The Building Blocks of Thought

A Rationalist Account of the Origins of Concepts

STEPHEN LAURENCE
ERIC MARGOLIS

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2024932365

ISBN 9780192898838 (hbk.)
ISBN 9780192898920 (pbk.)

DOI: 10.1093/9780191925375.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

For Kate and Shelly

Contents

PART III ALTERNATIVE EMPIRICIST PERSPECTIVES

PART IV FODORIAN CONCEPT NATIVISM

Preface and Guide to the Book

This book argues for a rationalist account of the origins of human concepts—that is, for a version of concept nativism. While this type of account comes in many varieties, they all take the mind to possess a rich innate structure that plays a central role in explaining the origins of concepts. Our own version of concept nativism holds that *many concepts across many conceptual domains are either innate or acquired via learning mechanisms that involve innate representations or other innate special-purpose elements*. Drawing on a broad range of evidence from many different disciplines, we argue that there is a powerful case to be made in favour of this view. However, we are also keenly aware of the fact that the rationalism-empiricism debate is widely seen as being irrelevant to contemporary theorizing about the mind—or worse, as being fundamentally confused—and that this has led many philosophers and cognitive scientists to dismiss it altogether. If this scepticism regarding the value and coherence of the rationalism-empiricism debate were warranted, our project would be doomed from the start. So in addition to making a case for our rationalist view over competing alternatives (both rationalist and empiricist), it is essential that we also address the fundamental challenges that call the debate itself into question.

Part I of the book provides a comprehensive rethinking of the theoretical foundations of the rationalism-empiricism debate which clarifies what exactly the debate is about—as well as what it is *not* about—and at the same time makes clear why it remains central to the study of the mind. In our view, the rationalism-empiricism debate should be understood to be about the differing views that rationalists and empiricists hold regarding the collection of innate psychological structures which constitutes the ultimate psychological basis for the acquisition of all further psychological traits. Likewise, the more specific debate about the origins of concepts should be understood to be about the differing views that rationalists and empiricists hold regarding the collection of innate psychological structures which constitutes the ultimate psychological basis for concept learning. This way of understanding the rationalism-empiricism debate is not new. But it has never been fully articulated and is frequently conflated with (or rejected in favour of) a number of prominent alternative ways of understanding the debate that turn out to be intellectual dead ends, especially the view that it is about nature versus nurture (or the relative contributions of genes versus the environment). Both critics of the rationalism-empiricism debate and its proponents and participants frequently conceptualize it in these mistaken and unproductive ways, often conflating several incompatible interpretations of the debate without

realizing it. We see the widespread scepticism regarding the value and coherence of the debate as stemming directly from such misunderstandings. While critics have rightly regarded these ways of understanding the debate as unworkable, they have been wrong to conclude that the debate itself should be abandoned as a result. Instead, what's needed is a better understanding of the debate. This understanding should be built around the idea that we started with—that rationalists and empiricists differ in terms of the ultimate psychological basis that they posit for acquiring all further psychological traits. By systematically developing this interpretation of the debate—and sharply distinguishing it from unproductive alternatives—Part I establishes a sound theoretical foundation for the debate, providing a detailed framework for understanding the diverse range of possible rationalist and empiricist theories and how they relate to one another.

In Part II, we turn to our positive case for concept nativism. As we see it, there is an overwhelming case to be made in favour of our view that many concepts across many conceptual domains are either innate or acquired via learning mechanisms that involve innate representations or other innate special-purpose elements. In making this case, we distinguish and clarify seven distinct types of argument supporting concept nativism, many of which have been poorly understood or insufficiently appreciated. Since our view is that a rationalist view about the origins of concepts is the right view to hold for many concepts across many conceptual domains, our discussion needs to cover a broad range of concepts from different conceptual domains. An exhaustive treatment of each of our seven arguments for concept nativism as it applies to every candidate concept and conceptual domain is out of the question. Instead, to make the discussion manageable, we have chosen to illustrate the breadth of our account—the range of concepts and conceptual domains that it covers—by bringing in new examples as we introduce each new argument. To illustrate the depth of the case for concept nativism—the fact that often many of these arguments apply to a given type of concept or conceptual domain—we examine a selection of conceptual domains from the vantage point of a number of these different arguments. While each of these seven arguments individually supports a rationalist perspective, the full force of the case for concept nativism comes from their collective impact and the recognition that they comprise what amounts to a single multifaceted inference to the best explanation argument for concept nativism. This argument not only demonstrates that a rationalist account of the origins of concepts should be adopted over competing empiricist accounts, it also shows why our version of concept nativism should be adopted over competing rationalist accounts (e.g., what are known as *core knowledge* accounts) which take there to be considerably less rich innate structure underlying concept learning.

Part III critically examines the empiricist opposition to concept nativism. Our critique of this opposition is organized around a representative selection of some of the most important and influential empiricist accounts of concept acquisition.

One common theme of Part III is that these empiricist proposals fail to do justice to the theoretical and empirical considerations that drive the rationalist accounts they are meant to be alternatives to. At the same time, however, we argue that work in the empiricist tradition contains valuable insights about conceptual development. We argue that not only are these insights consistent with concept nativism but that they can make a more significant contribution to explaining conceptual development when incorporated into a rationalist framework. Our discussion in Part III extends both the breadth and depth of conceptual domains covered in relation to the arguments for concept nativism in Part II by illustrating ways in which many of these arguments apply to new conceptual domains. We conclude that an examination of empiricist alternatives to concept nativism only serves to strengthen our case for rationalist accounts of the origins of concepts in general, and for our own version of concept nativism in particular.

Finally, Part IV addresses what is perhaps the most famous contemporary position in the rationalism-empiricism debate regarding the origins of concepts, namely Jerry Fodor's influential view that semantically primitive concepts (concepts that aren't composed of more basic representations) can't be learned and the corollary Fodor argued for that virtually all lexical concepts are innate (a view known as *radical concept nativism*). One of the reasons that Fodor's arguments against concept learning have figured so prominently in this debate—despite the wildly counterintuitive conclusions they are associated with—is that it has proven to be remarkably difficult to say exactly where they go wrong. But perhaps even more importantly, many theorists see Fodor's arguments as containing a deep insight about learning that imposes a fundamental constraint on any theory of concept learning; they just see Fodor as having drawn the wrong moral from this insight. While rejecting Fodor's radical concept nativism, these theorists agree with Fodor's claim that semantically primitive concepts cannot be learned and so must be innate. In fact, this view about conceptual structure and the limits on what can be learned lies behind a nearly universally accepted model of concept acquisition—endorsed in different ways by rationalists and empiricists alike—which we call the *Acquisition by Composition model* (or *ABC model*) of conceptual development. According to this model, concept learning requires that the learned concept be a complex concept which is formed from a compositional process that builds the new concept out of its semantic constituents. The heart of Part IV of the book is directed at showing why this model is mistaken. Our discussion encompasses an overview of the history of Fodor's views on these issues, which changed substantially over a period of more than thirty years. By carefully analysing Fodor's arguments, we show precisely how they go wrong, which in turn shows why the ABC model of conceptual development should also be rejected. The rejection of this model opens up a range of new possibilities for explaining how concepts can be learned which we explore in relation to a variety of different types of concepts and different theories of meaning for mental representations. This

discussion further underscores a major theme of the book—that rationalist accounts of the origins of concepts not only are consistent with concept learning but also offer the best overall account of how concept learning works. We end Part IV on this theme by highlighting the depth of the connection between our own rationalist account of the origins of concepts and cultural learning.

Since this is a long book, we have tried to arrange it in such a way that the four main parts of the book can be read on their own or out of order (though readers who do this may need to consult Chapter 2 and Chapter 6 for key terminology we use later on). Likewise, most chapters are sufficiently self-contained that readers who are interested in particular topics can jump ahead to the relevant chapter. However, it should be kept in mind that the theoretical framework in Part I and the many arguments, examples, and empirical findings that are discussed in different chapters in Parts II–IV are meant to interact with and support one another as part of a single integrated argument for concept nativism that runs through the entire book.

Acknowledgements

We would like to thank two anonymous referees for valuable comments and our colleagues, friends, and families for all of their help and support. We are also grateful for financial and institutional support from the University of Sheffield, the University of British Columbia, the Hang Seng Centre for Cognitive Studies, the Peter Wall Institute for Advanced Studies, the Arts and Humanities Research Council, and Canada's Social Sciences and Humanities Research Council. In addition, we would like to acknowledge that parts of Chapters 2, 4, 5, 13, 25, and 26 draw upon material from the following previously published articles and book chapters:

Margolis, E., & Laurence, S., "Making Sense of Domain Specificity," *Cognition*, 240, 105583 (2023); published by Elsevier.

Laurence, S., & Margolis, E. "Concept Nativism and Neural Plasticity," in E. Margolis & S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts*, MIT Press, 2015.

Margolis, E., & Laurence, S., "In Defense of Nativism," *Philosophical Studies*, 165:2 (2013), 693–718; published by Springer.

Laurence, S., & Margolis, E. "Abstraction and the Origin of General Ideas," *Philosophers' Imprint*, 12:19 (2012), 1–22.

Margolis, E., & Laurence, S. (2011), "Learning Matters: The Role of Learning in Concept Acquisition," *Mind & Language*, 26:5 (2011), 507–639; published by Wiley-Blackwell.

We are grateful to all of these journals and presses for granting us permission to reproduce this material. We are also grateful to MIT Press for permission to reprint as the epigraph a passage from Jerry A. Fodor's *Representations: Philosophical Essays on the Foundations of Cognitive Science*.

List of Figures

Note on Authorship

This book was a fully collaborative project; the order of the authors' names is arbitrary.

It seems to me that Anglo-American theorizing about concept attainment has, for several hundred years now, restricted itself to the consideration of a very small range of theoretical options. It also seems to me that the results have not been extraordinarily encouraging. Perhaps it is time to throw open our windows, kick over our traces, upset our applecarts and otherwise wantonly mix our metaphors. If we are going to have a cognitive science, we are going to have to learn from our mistakes. When you keep putting questions to Nature and Nature keeps saying "no", it is not unreasonable to suppose that somewhere among the things you believe there is something that isn't true.

Jerry Fodor

1
Introduction
Whatever Happened to the Debate over Innate Ideas?

One of the most remarkable features of the human mind is the breadth and richness of what it can represent. We aren't limited to thinking about current sensations or even to the objects in our immediate environment. Our thoughts can also turn to abstract matters (truth, beauty, justice), to things that are far away in space and time (the rings of Saturn, the Crimean War), to things that haven't happened (a world in which the dollar remained on the gold standard), and even to things that don't exist (Santa Claus, vampires, phlogiston).

This fact about the mind's representational powers leads to a question that has been at the centre of an enduring and, we think, highly productive debate that traces back to antiquity—a question that is integral to nearly all philosophical theorizing about human nature and that has motivated an enormous amount of work in cognitive science. There are a number of ways of putting this question, but perhaps the most recognizable and eloquent formulation is owing to John Locke:

> How comes it [the mind] to be furnished? Whence comes it by that vast store, which the busy and boundless Fancy of Man has painted on it, with an almost endless variety? Whence has it all the materials of Reason and Knowledge? (1690/1975, II.I.2, p. 104)

In short, where do our concepts or ideas come from?[1]

[1] Although there are some differences between how ideas were understood in the history of philosophy and how concepts are understood today, there is enough of an overlap that we will use the terms "concept" and "idea" interchangeably. What exactly concepts are has been a matter of significant controversy, both in philosophy and in cognitive science (Margolis and Laurence 1999, 2015). We will say more about some of these controversies, and about what concepts are and how they relate to the rationalism-empiricism debate in Chapter 6. For the time being, what matters most is that we take the representational components that make up thoughts to be concepts, where thoughts are the representations that are involved in such high-level cognitive processes as categorization, decision making, recalling facts, analogical reasoning, interpreting discourse, forming explanations, planning a course of action, and problem solving. For example, when you think *blue whales are the largest animals to have ever existed*, your having this thought involves the activation of a mental representation that is composed of simpler representations—concepts—including ones for blue whales, animals, and existence (among others). We will follow the convention in which mentioned concepts and ideas appear in small caps—for example, BLUE WHALE for the concept of blue whales and ANIMAL for the concept of animals.

This book offers an answer to Locke's question that is inspired by the speculations of Plato, Descartes, Leibniz, and other rationalist thinkers in the history of philosophy, and that owes a great deal to the rationalist theorizing in cognitive science that began with Noam Chomsky's pioneering work in linguistics. In contemporary philosophy and in much of cognitive science, scepticism about rationalist views of the mind is common. Nonetheless, we think that a strong case can be made for a rationalist view of the origins of concepts—a view we refer to as *concept nativism*—and in the course of this book, we present this case in detail.

1.1 The Rationalism-Empiricism Debate about the Origins of Concepts

We will begin by providing an initial overview of what we take to be at stake in the debate between rationalist and empiricist accounts of the origins of concepts. Our aim in this initial overview is simply to sketch the basic outlines of the debate. In Chapter 2, we will revisit these issues and provide a more detailed official statement of our view of rationalism, empiricism, and the rationalism-empiricism debate. This will cover both the full scope of the rationalism-empiricism debate—which is about the origins of many different types of psychological traits—and the intricacies of how the debate should be understood when the focus is on the origins of concepts. But for now we will leave most of those details out and just sketch the general contours of how rationalist and empiricist views of concepts differ from one another.[2]

The debate is sometimes characterized so that rationalism is the view that there *are* innate ideas or concepts and that empiricism is the view that the mind is initially a blank slate in that it has no innate structure whatsoever. However, this way of distinguishing empiricism from rationalism is problematic. First, to characterize empiricism as the view that the mind begins with no innate structure would have the unfortunate consequence that there aren't really any empiricists. It has long been recognized by all parties to the rationalism-empiricism debate that a mind without any innate structure—a truly blank slate—wouldn't be

[2] Rationalists and empiricists in the history of philosophy took contrasting stands on a range of issues in epistemology and the philosophy of mind that are independent of the questions about the origins of concepts that are our concern in this book. Unless otherwise indicated, when we refer to *rationalism* and *empiricism* (and to the *rationalism-empiricism debate*), we are referring only to these views insofar as they bear on questions regarding the innate structure of the mind and the psychological basis of cognitive and conceptual development. There are other terms that have been used for rationalism, including *innatism*, *innativism*, and *nativism*. Moreover, it is not uncommon to refer to the psychological debate between rationalists and empiricists as the *nativism-empiricism debate*, in which competing rationalist and empiricist views of the origins of concepts are known as *concept nativism* and *concept empiricism*. We will sometimes use this terminology too. In particular, we will often refer to our own view as a form of concept nativism, since this term is commonly used to refer to rationalist accounts of the origins of concepts in both philosophy and cognitive science.

capable of learning. There has to be something that accounts for why human beings come to know anything at all about the world around them while things like rocks and chairs don't.[3] Second, although it is true that rationalists are more likely than empiricists to embrace innate concepts in addition to other types of innate psychological structures, focusing exclusively on whether concepts are innate or not doesn't do justice to mainstream views of the origins of the conceptual system. Empiricists may accept some innate concepts, and rationalists may hold that what matters is not innate concepts per se but rather the existence of a rich innate basis for acquiring concepts.

For these reasons, we think it best to characterize the debate in other terms, which more accurately reflect the nature of the actual disagreement between empiricists and rationalists. For contemporary theorists in philosophy and cognitive science, this revolves around the character of the innate psychological structures that underlie concept acquisition. While both empiricists and rationalists posit innate psychological structures in order to explain how concepts are acquired, they diverge in terms of the *number* and *kinds* of innate psychological structures they accept.

According to empiricist approaches, there are few if any innate concepts and concept acquisition is, by and large, governed by a small number of innate general-purpose cognitive mechanisms being repeatedly engaged. Sometimes this point is put by saying that empiricists claim that concepts are largely acquired on the basis of experience and hence that the conceptual system is predominantly a product of learning. But the crucial fact here isn't that empiricists place a lot of weight on learning or experience. (As we'll see in a moment, rationalists do too.) Rather, what is unique to empiricism in the rationalism-empiricism debate is its *characteristically empiricist approach to concept learning*. The empiricist view is that concept learning ultimately traces back almost exclusively to general-purpose (domain-general) cognitive mechanisms and that these provide the psychological underpinning for the many varied concepts that humans come to possess. For example, on a typical empiricist view, concepts related to things that are agents (as opposed to inanimate objects) and concepts related to number are both the product of the same kind of psychological processes embedded in the same general-purpose concept learning mechanisms. The mechanisms produce agency

[3] Some contemporary theorists who undoubtedly fall on the empiricist side of the rationalism-empiricism divide have rejected the label "empiricism" because of its association with the view that the mind lacks innate structure. For example, in a discussion relating work in neuroscience to theories of conceptual development, Steven Quartz remarks, "I have avoided using the term empiricism, instead stating the strategy in terms of not being strongly innate. My reason for this lies in the common identification of empiricism with *Tabula Rasa* learning" (2003, p. 34). Similarly, Elman et al. (1996) reject the label "empiricism", identifying empiricism with the view that genes play no role in determining behaviour: "There can be no question about the major role played by our biological inheritance in determining our physical form and our behaviors. We are not empiricists" (p. 357). Since we take there to be a substantial issue at stake between theorists like Quartz or Elman et al. and concept nativists, a better characterization of empiricism is clearly needed.

representations in one case and number representations in another not because they have special-purpose elements that dispose them to do this, but simply as a product of processing input in these domains.

Rationalist approaches, in contrast, typically embrace some innate concepts, but more importantly, they suppose that concept acquisition isn't governed solely by a few innate general-purpose cognitive mechanisms. Rather, rationalists maintain that, in addition to innate general-purpose cognitive mechanisms and some innate concepts, there are a number of special-purpose learning mechanisms (with varying kinds and degrees of specialization) that play a vital role in conceptual development. Each of these rationalist special-purpose learning mechanisms governs the acquisition of a restricted range of concepts and is either itself an innate mechanism or constructed in part from innate special-purpose resources. So, what is unique to rationalism in the rationalism-empiricism debate is, first, that it typically posits a stock of innate concepts and, second, that it has a *characteristically rationalist approach to concept learning*. A rationalist view is perfectly at home with the claim that representations of agency might depend on psychological processes that reflect the operation of innate agency-specific concept learning mechanisms, while representations of number depend on separate, innate number-specific concept learning mechanisms. The reason why agency representations form in the one case and numerical representations in the other would then be due as much to the fact that they are governed by different innate special-purpose learning mechanisms as it is to the differing input to these mechanisms.

Rationalism and empiricism are not specific theories. Rather, they are each theoretical frameworks within which there are many different theoretical options. For example, within the empiricist framework, some empiricists claim that there are no innate concepts whatsoever and that concept acquisition depends exclusively on a small number of general-purpose cognitive mechanisms. Jesse Prinz defends a view along these lines, holding that concepts "are all learned, not innate" (Prinz 2005, p. 679). After arguing against what he takes to be the main proposals for special-purpose innate concept learning mechanisms, he summarizes his discussion by noting, "I do not believe that any of these domains is innate. That is to say, I do not think we have innate domain-specific knowledge that contributes to structuring our concepts" (Prinz 2005, p. 688). A different type of empiricist account accepts a limited number of innate special-purpose mechanisms that constrain how the conceptual system develops in certain isolated cases.[4] However, such cases are often seen as constituting minor exceptions to the general rule that concept acquisition is governed solely by general-purpose learning

[4] The innate special-purpose mechanisms that empiricists posit are typically relatively simple and geared towards low-level perceptual features, such as a bias to attend to movement or to high- or low-frequency visual stimuli.

mechanisms. For example, Rogers and McClelland (2004) defend a view of this sort. After arguing that a general-purpose learning model can explain the way that adult semantic memory is organized, they suggest that there may be a handful of instances where special-purpose cognitive mechanisms constrain conceptual development, including a tendency to withdraw from strong stimuli and to respond favourably to the taste of fat and sugar. Yet Rogers and McClelland state that they are "reluctant…to accept that, in general, human semantic cognition is prepared in this way", arguing instead that "domain-general mechanisms can discover the sorts of domain-specific principles that are evident in the behavior of young children" (Rogers and McClelland 2004, p. 369). In later chapters, we will be discussing these and other empiricist views.

Within the rationalist framework, there is also a broad range of different possibilities. What these different rationalist accounts have in common is that, in addition to positing the types of innate structures found in empiricist accounts (innate general-purpose learning mechanisms), they also posit further innate structures that are involved in cognitive and conceptual development. In particular, these include innate concepts, innate special-purpose learning mechanisms for acquiring concepts in a particular domain, and innate special-purpose resources that contribute to other learning mechanisms for acquiring concepts in a particular domain (i.e., for learning mechanisms that aren't wholly innate but that have critical special-purpose parts that are innate). But rationalists will differ over such things as how many and what kinds of concepts are either innate or acquired via such *rationalist learning mechanisms*, as well as how rich the innate endowment is in any given conceptual domain.[5]

One influential rationalist view, known as *core knowledge* and *core cognition*, has been championed by Susan Carey, Elizabeth Spelke, and others. This view holds that:

> Just as humans are [innately] endowed with multiple, specialized perceptual systems, so we are [innately] endowed with multiple systems for representing and reasoning about entities of different kinds…studies suggest that there are at least four core conceptual systems encompassing [innate] knowledge of objects, agents, numbers, and space. (Carey and Spelke 1996, p. 517)

This view is sometimes understood to posit innate concepts in these core domains—for example, the concept of an object, the concept of an agent, the concept of belief, and so on—alongside special-purpose learning mechanisms involved in the acquisition of further concepts in these domains. But it is also

[5] We will use the term *rationalist learning mechanism* to refer to psychological mechanisms that are either innate special-purpose learning mechanisms or learning mechanisms involving innate special-purpose resources. See Chapter 2 for further discussion and qualifications.

possible to understand core cognition as not positing any specific innate concepts, and instead taking concepts like these to be the product of rationalist learning mechanisms. On this understanding, although the concept AGENT wouldn't be innate, a special-purpose learning mechanism that is responsible for its acquisition might be—one that underlies the acquisition of perhaps a range of concepts related to agency but not other types of concepts. On either understanding, core cognition is committed to innate resources that are particular to concepts in at least four domains. Other rationalist views might differ regarding the innate basis for acquiring concepts in these and other domains. Some would posit fewer innate concepts or fewer rationalist learning mechanisms, or would posit a less rich innate endowment in the domains singled out by core knowledge. Others would posit more innate concepts or rationalist learning mechanisms than core knowledge views, or posit a richer innate endowment in domains singled out by core knowledge. To give just one example, Lance Rips argues that the mechanisms posited by the core knowledge view "don't provide mental components that are sufficient to explain adult concepts" and concludes that they need to be supplemented by further innate concepts (Rips 2017, p. 159). Just as there are many different types of empiricist views, there are many types of rationalist views, and we will encounter a number of them later on too.

The central thesis of this book is that the right framework for understanding the origins of concepts is a rationalist framework. In defending this thesis, we will develop an extensive series of arguments in favour of rationalist accounts and will respond to empiricist criticisms. These arguments can be seen, in the first instance, as arguments in favour of rationalist accounts in general (the rationalist framework). When considered in isolation, each is consistent with a range of different rationalist views—that is, with many different forms of concept nativism. At the same time, however, we will argue that, taken together, they show not only that empiricist views substantially underestimate the innate endowment underpinning human conceptual development but that many rationalist views do as well. Ultimately the view that we favour is a form of concept nativism that holds that *many* concepts across *many* different conceptual domains are either innate or acquired via rationalist learning mechanisms.

At this stage of inquiry, it is not possible to say precisely how many (or precisely which) concepts are innate or acquired via rationalist learning mechanisms. There remains considerable room for reasonable disagreement. Accordingly, our aim is not to offer a complete or final catalogue of such concepts. Any attempt to do so would be premature if only because many conceptual domains remain largely unexplored. That said, we will argue that there is overwhelming support for the claim that the ultimate catalogue of these concepts will be extensive. As part of our case for concept nativism, we will argue for a rationalist treatment of a wide range of concepts across diverse conceptual domains. Likely candidates, in our view, include concepts associated with the representation of objects, space,

time, geometry, number, agency, individuals, mental states (e.g., perception, belief, emotions), communication, causation, animals, plants, food, danger, disease, goals, paths, movement, events, feature/property, stuffs/substances, logic, modality (e.g., possibility, necessity), sameness/difference, sex, life stages, kinship, social groups, social status, tools, function/purpose, norms, cooperation, and morality (e.g., fairness, harm, obligation).

While this amounts to a considerable number of concepts across a wide range of conceptual domains, we want to make it absolutely clear that we nonetheless think that *most* concepts are *not* innate and that relatively general-purpose learning mechanisms play an important role in the acquisition of many concepts. In fact, we don't see how any tenable form of rationalism could deny these things. We emphasize these points because one of the most famous—many would say infamous—rationalist accounts of the origins of concepts, Jerry Fodor's *radical concept nativism* (Fodor 1975, 1981), denies them. And Fodor's extreme view is often mistakenly taken to be representative of rationalist accounts of the origins of concepts in general.

According to Fodor's radical concept nativism nearly all lexical concepts are innate, including the likes of LINGUINI, CARBURETTOR, BEATNIK, and QUARK.[6] Indeed, Fodor argues that it is *impossible* for such concepts to be learned. Notice that it isn't just the sheer volume of innate concepts that makes this view so outrageous—the thousands and thousands of concepts corresponding to actual and potential natural language words—but also the fact that most of these concepts are clearly newcomers in human history, dependent upon specific historical, cultural, and technological conditions for their appearance. As implausible as Fodor's view is, we think that there is nonetheless much to be gained from a proper analysis of the ingenious arguments he has put forward for this view (see Part IV).

Setting aside Fodor's extreme and highly unrepresentative radical concept nativism, it is clear that rationalism isn't confined to postulating innate concepts.[7] Rather, a big part of concept nativism is the use it makes of rationalist learning mechanisms in explaining the origins of concepts, and, as their name suggests, these mechanisms are often best understood as *learning* mechanisms. For instance, a special-purpose mechanism for food might support the learning of which items in the environment can be eaten and which are potentially toxic and

[6] Lexical concepts are ones that are expressed by individual words in natural language.

[7] Fodor's extreme account stands in an analogous relation to rationalism as the view that the mind is a blank slate stands in relation to empiricism. Just as there are few if any real advocates of the view that the mind is a blank slate, one is hard pressed to find any real advocates of Fodor's radical concept nativism. And while Fodor's view may not quite be incoherent, it is very nearly as implausible as the blank slate view. Finally, despite being widely rejected by rationalists, Fodor's extreme account has often been mistakenly taken to represent the rationalist view, just as the blank slate view has been mistakenly taken to be representative of the empiricist view of cognitive and conceptual development despite being widely rejected by empiricists.

to be avoided, guiding food preferences and food-seeking behaviour. Or a special-purpose mechanism for faces might support the learning of concepts of individuals. These hypothesized mechanisms are very much in the business of learning about the world, according to the rationalist. They are just specialized for learning particular information in a way that is highly constrained by the nature of the learning mechanism.

To a large extent, then, the difference between rationalism and empiricism isn't *whether* learning is central to human concept acquisition but rather their differing views of *how* learning works. While empiricists take learning to be almost exclusively mediated by innate general-purpose learning mechanisms, rationalists maintain that general-purpose learning mechanisms, though real and important, are not sufficient, and so rationalists also take there to be innate concepts and numerous innate special-purpose psychological structures involved in learning.

1.2 Philosophy, Psychology, and the Naturalistic Study of the Mind

As we mentioned in the previous section, concept nativism has deep roots in the history of philosophy. But over the years, the intellectual landscape has changed in a number of important ways, making for a complicated relationship between the historical debate over innate ideas and the debate between contemporary empiricists and rationalists. Perhaps the most important difference has to do with the broader set of philosophical issues that were wrapped up with the status of innate ideas in historical discussions (Cowie 1999; Samet 2008). In the historical debate, the issue of innate ideas was taken to have far-reaching metaphysical and epistemological consequences, including implications for the existence of God, the relation between soul and body, and the nature of morality.

Plato, for example, argued in the *Phaedo* that the idea of equality is innate on the grounds that sensory experience cannot give us this idea, since things that appear to be equal in length are never really exactly equal. From this he concluded that the idea must be one that we have prior to perceiving the world—that coming to understand equality is, in effect, a matter of recollecting something we knew but have since forgotten. It was then a short step to the view that people have a soul that was once situated in a non-physical realm, the only realm in which true equality itself exists. Or consider Descartes' musings about God in the *Meditations*. Part of his rationale for believing in God turned on an argument that the idea of God couldn't be acquired by ordinary experience. For Descartes, the cause of this idea wouldn't have enough "formal reality" if it weren't a perfect being—God himself—that was its cause.

However, by far the most prevalent extra-psychological issue in the historical debate involved questions about the justification of human knowledge. Rationalists in the history of philosophy gave a priori knowledge (roughly, knowledge not justified through experience) a central role in epistemology, which generated the vexed problem of explaining how such knowledge is possible. The answer for many rationalists was to postulate innate ideas, principles, and faculties of the mind. The thinking was that the knowledge these innate structures lead to is justified by virtue of its psychological origins, often backed by God's goodness for giving us the innate endowment in the first place. However, it can't be taken for granted that a higher being ensures the truth or validity of any innate beliefs we might happen to have. And in hindsight it isn't hard to see that, in principle, a belief that requires empirical justification could be innate (e.g., the belief that humans have hands), while a belief that requires a priori justification might not be (e.g., the belief that arithmetic is incomplete). Justification is one thing, psychology another.

Does this mean that the earlier debate about innate ideas is merely a historical curiosity? Not at all. Although its participants held some questionable subsidiary views and injudiciously mixed up their epistemology and their psychology, they were nonetheless interested in the workings of the mind.

What's more, while their approach to psychological matters wasn't scientific by today's standards, it wasn't entirely devoid of empirically grounded argumentation either. For example, Descartes' views about innate ideas were informed by the observation that people arrive at ideas that are not exemplified by the perceptible objects that cause them. Seeing a triangular shape on a piece of paper may lead to the idea of a perfect Euclidean triangle even though the form on the paper invariably falls short of being perfectly triangular in various ways (e.g., its lines can't help but have a certain amount of breadth) (see Descartes 1641/1984, p. 262). Descartes' rejection of the theory that ideas come from the senses was also based on considerations deriving from his study of the physiology of vision: "for the sense-organs do not bring us anything which is like the idea which arises in us on the occasion of their stimulus, and so this idea must have been in us before" (Descartes 1641/1991, p. 187).

Similar sorts of forays into empirical argumentation can be found in Locke, the most famous critic of innate ideas. Among other things, Locke pointed to what he took to be unassailable facts about the minds of children and people from remote parts of the globe. For him, these were of the utmost importance because of the presumed link between a principle being innate and its being universal. "I agree with these Defenders of innate Principles, That if they are *innate*, they must needs *have universal assent*" (1690/1975, I.ii.24, p. 61). Locke's tactic was to argue against his rationalist opponents by providing examples of people who don't endorse or appreciate philosophical principles that had been upheld as being innate:

> But he that from a Child untaught, or a wild Inhabitant of the Woods, will expect these abstract Maxims, and reputed Principles of Sciences, will I fear, find himself mistaken. Such kind of general Propositions, are seldom mentioned in the Huts of *Indians*: much less are they to be found in the thoughts of *Children*… (Locke 1690/1975, I.ii.27, p. 64)

Still, the true potential of empirical argumentation to address the debate over innate ideas didn't really come out until the work of Chomsky and others at the forefront of the cognitive revolution in the 1960s. Chomsky particularly drew attention to a wealth of new empirical data concerning the innate basis of language, while explicitly linking his theories of language acquisition to neglected rationalist views of language and thought in the seventeenth, eighteenth, and early nineteenth centuries—a body of work he referred to as *Cartesian linguistics*. All of this new empirical data, Chomsky argued, served to vindicate the general view of the mind associated with philosophers in the rationalist tradition:

> It seems to me that the conclusions regarding the nature of language acquisition, discussed above, are fully in accord with the doctrine of innate ideas, so understood, and can be regarded as providing a kind of substantiation and further development of this doctrine. (Chomsky 1967, p. 131)

Chomsky's focus was on language, not concepts. But language was clearly meant as just one well-developed case study, and others were expected to follow.[8]

There are two important features to Chomsky's claim that ought to be emphasized and that are central to our own defence of concept nativism. The first is his insistence that empiricist and rationalist theories constitute empirical proposals regarding the mind and consequently ought to be evaluated in the same way as other empirical proposals. For Chomsky, there is no a priori method for discovering the structure of the mind any more than there is an a priori method for discovering the structure of the circulatory system:

> Particular empiricist and rationalist views can be made quite precise and can then be presented as explicit hypotheses about acquisition of knowledge, in particular, about the innate structure of a language-acquisition device…When such contrasting views are clearly formulated, we may ask, as an empirical question, which (if either) is correct. (Chomsky 1965, pp. 52–53)

[8] The relation between contemporary work in linguistics and the historical debate over innate ideas is a major theme in Chomsky's writings in the 1960s and 1970s. See also Chomsky (1965, 1966, 1971, 1972/2006, 1975).

Given that empirical methods of inquiry are more strongly associated with empiricism in the history of philosophy, the link that Chomsky draws between his own work and the rationalist philosophical tradition may seem surprising at first. But for Chomsky, these issues about methods of inquiry are entirely distinct from the sorts of psychological views that rationalists and empiricists considered to be viable. To the extent that empirical methods are thought to be inherently tied to empiricism, we might say that Chomsky's *epistemology* is empiricist even if his *psychology*—and, in particular, his view on the innate structure of the mind—is rationalist. In any case, Chomsky's view illustrates how there is no incompatibility between using empirical evidence to support rationalist accounts of the mind (or to argue against empiricist accounts); empiricism shouldn't be seen as having a monopoly on the use of empirical evidence and argument.

The second important feature of Chomsky's revival of rationalism was his rejection of the widely held view that empiricism is superior to rationalism for being more parsimonious:

> Where empiricist and rationalist views have been presented with sufficient care so that the question of correctness can be seriously raised, it cannot…be maintained that in any clear sense one is "simpler" than the other in terms of its potential physical realization, and even if this could be shown, one way or the other, it would have no bearing on what is completely a factual issue. (Chomsky 1965, p. 53)

Chomsky's point is that the way to choose among competing proposals about the mind is by assessing the depth and cogency of the way they handle empirical data, not how simple they are according to some preconceived understanding of simplicity. To suppose otherwise is to adopt a dogmatic approach to the study of the mind. Thus the burden on all theorists (empiricists and rationalists) is to formulate sufficiently articulated theories that can be evaluated for their explanatory power.[9]

Chomsky's own proposals were exciting precisely because they were so well developed and because they were able to accommodate an abundance of linguistic data that previous theorists hadn't taken sufficiently seriously or had completely overlooked. Chomsky stressed that natural language speakers regularly

[9] We should note that Chomsky isn't claiming that explanatory power is the only measure of theoretical goodness, or that all forms of simplicity are theoretically irrelevant; rather, he is arguing that simplicity considerations that aren't tied to explanatory power should carry little weight in theorizing about the mind. Also, the reason we have added the qualification *some preconceived understanding of simplicity*, as opposed to speaking of simplicity in an absolute sense, is that there are different ways to measure the simplicity of a psychological theory. This fact makes it even harder to maintain that empiricist models are to be preferred for being simpler, since on a number of pertinent standards of simplicity, empiricist models turn out to be *less* simple than rationalist models. (See Chapters 4 and 17 for further discussion of these issues.)

use and understand novel sentences and consequently that a person's knowledge of language can't reside in a memorized list of sentences that are reinforced as a response to specific stimuli. He also noted that parents don't generally correct children's syntactic errors and that this places significant constraints on the way that children come to learn the language of their community. But most importantly, Chomsky called attention to a striking range of facts regarding people's linguistic knowledge—patterns in their intuitions about acceptable and unacceptable sentences and their potential interpretations.[10] This data played a pivotal role in the emergence of modern linguistics and in the development of sophisticated competing rationalist and empiricist models of language acquisition.

Following the cognitive revolution, an enormous amount of empirical work has been done that bears on the rationalism-empiricism debate. As a result, contemporary theorists interested in the origin of human concepts have a truly unprecedented body of empirical data at their disposal. Where Locke and Descartes were confined to rudimentary observations and conjectures, contemporary theorists have access to a huge range of important discoveries about the mind that could hardly have been imagined even a few decades ago. The list of disciplines whose findings promise to shed light on Locke's question is long and impressive. It includes anthropology, archaeology, behavioural ecology, behavioural economics, clinical psychology, cognitive psychology, comparative

[10] While it isn't possible to do full justice to this data here, we can give a few illustrative examples. (Following the standard convention in linguistics, we mark unacceptable sentences with an asterisk, "*".) To a native English speaker, (1) and (2) are perfectly acceptable sentences. But although (3) and (4) are closely modelled on (1) and (2), and although (3) is acceptable, (4) is not (Chomsky 1957):

(1) The book is interesting. (2) The book seems interesting.
(3) The child is sleeping. (4) * The child seems sleeping.

Likewise, even though (7) and (8) seem to follow the same pattern as (5) and (6), native English speakers reject (8) (Chomsky and Lasnik 1977).

(5) Who do you want to see? (6) Who do you wanna see?
(7) Who do you want to see Bill? (8) * Who do you wanna see Bill?

Speakers of English also have intuitions about possible interpretations of English sentences, including some interpretations that are not immediately obvious. Notice that (9) can mean the same as either (10) or (11)—or even (12) (Chomsky 1965):

(9) I had that book stolen.
(10) Someone stole the book in question from me.
(11) I hired someone to steal the book in question from someone else for me.
(12) I was about to succeed in stealing the book from someone, but was caught at the last minute.

(This last reading can take a good amount of work to hear, but compare: "I had the race in the bag. But then I tripped, and injured myself just before the finish line.") English speakers also have subtle intuitions about such things as when "he" and "him" can be co-referential with a proper name in the same sentence—e.g., "John" in (13)–(16)—despite the fact that the principles governing this are not at all intuitively obvious (Chomsky 1995):

(13) John criticized him. (co-reference not possible)
(14) John said Mary criticized him. (co-reference possible)
(15) He said Mary criticized John. (co-reference not possible)
(16) After he left the room, Mary criticized John. (co-reference possible)

(animal) psychology, computational modelling, developmental psychology, endocrinology, ethology, evolutionary psychology, genetics, linguistics, neuroscience, robotics, and social psychology, among others. In each case, however, it must be recognized that the implications for the debate over innate ideas remain tremendously controversial. Even researchers from the same discipline, but with differing perspectives, have questioned each other's methods and have found themselves drawing very different conclusions from the same data. Given the diversity of this research, the sheer scope of its output, and the many perplexing philosophical and theoretical issues that invariably come up, an empirically informed assessment of the rationalism-empiricism debate about the origins of concepts is a challenging task, to say the least. We will end this initial chapter by briefly illustrating both the promise and some of the challenges of evaluating this broad range of empirical data by discussing a kind of concept—geometrical concepts—which were central to the historical debate over innate ideas.

1.3 An Example: Geometrical Concepts

In Plato's dialogue the *Meno*, Socrates is depicted as demonstrating the existence of innate geometrical knowledge by helping an uneducated boy complete a subtle geometrical proof with only minimal prompting. The dialogue involves a fictional interaction in a literary work. But suppose Plato's dialogue was a faithful account of an actual historical event. Would the event it recounts be enough to show that geometrical knowledge is innate or that there are any innate geometrical concepts? One obvious difficulty is that we can't take the boy's purported ignorance at face value. Even though the boy wasn't formally educated, he did live in a literate community filled with manufactured shapes, symbols, maps, and other technologies for conveying geometrical knowledge. Maybe this feature of the boy's environment was doing more work than Plato realized.

How might we obtain stronger evidence for innate geometrical concepts? A good start would be to design experiments that can be performed with young children—children who have had limited experience with geometry. One type of experiment that has been especially fruitful uses a reorientation paradigm in which experimental participants witness an object being hidden in a given location in a room and are subsequently gently spun around until they become disoriented. Then their task is to recover the object by exploiting what they remember of the layout of the room and any available potential cues that could help them to reorient themselves.

Using this procedure, Hermer and Spelke (1996) examined the ability of 18- to 24-month-old children (and adults) to locate an object hidden in one of four identical containers, one in each corner of a rectangular room with white walls

and no further cues (see Figure 1.1). In this situation, the best one can do is use the geometrical properties of the room, for example, by noting the fact that the object was hidden in a corner with a long wall on the left and a short wall on the right. But of course, if one does use this strategy, then there is no way to reliably pick the correct corner. All that can be hoped for is to search equally in the correct corner and the geometrically equivalent opposite corner, since both of these corners will have a long wall on the left and a short wall on the right. This is exactly what Hermer and Spelke found that the participants did, not just the adults but also the children. Next, Hermer and Spelke continued their investigation using essentially the same task but this time one of the short walls was covered by a blue cloth. In this case, it is possible to narrow things down to the correct corner by using geometrical and landmark information, for example, by representing the hidden object as being in the corner with a long wall to the left and the short blue wall to the right. As you would expect, adults had no trouble with the task; the overwhelming majority chose the correct corner on the first try. But evidentially the blue wall didn't help the children. They continued to look in the correct corner and the geometrically equivalent opposite corner equally, just as before. This suggests that not only can young children represent geometrical properties and use them to reorient themselves but that they may exclusively rely on geometrically properties, ignoring even highly salient landmarks.

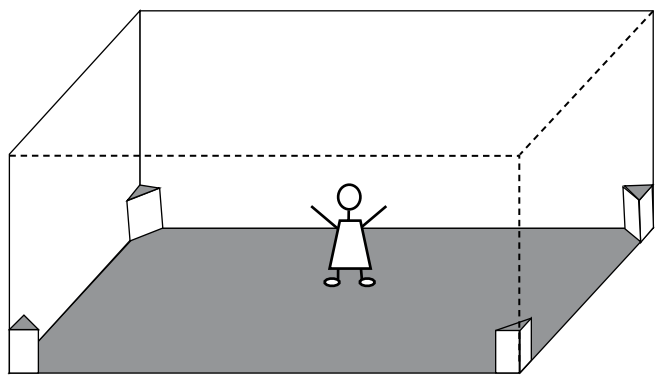

Figure 1.1 The basic reorientation experiment. Individual experimental participants watch an item being hidden in one of the four containers in the corners of a rectangular room. They are then disoriented and asked to recover the hidden item. See Hermer and Spelke (1996) for further details.

What are the implications of this experiment for the debate about innate ideas? Obviously this is only one experiment, and no single experiment can be expected to settle the question of whether geometrical concepts are innate. But it does offer some hope that Plato was on to something. Notice that the children in question

are very young—on average, they are less than 2 years old—and yet apparently make use of geometrical properties. At the same time, however, the experiment illustrates some of the difficulties of moving from empirical data towards a defensible position regarding the innate structure of the mind. To the extent that children's success in exploiting the geometrical layout of the room speaks to an innate mechanism for representing geometrical properties, this is because of certain assumptions about the children being too young to have learned the relevant geometrical concepts via general-purpose learning. But are they really too young in this instance? They did have eighteen to twenty-four months of experience. Maybe for a powerful enough general-purpose learning mechanism, that is enough time. How to determine what counts as enough or too little time is itself a complex question and one that is very much in dispute (see Chapters 8 and 9).

One solution might be to insist that we should focus on studies with newborns, children who have only had minutes of postnatal experience, not months. However, children who are this young can't move around a room on their own, so we can say with 100% confidence that geometrical concepts aren't going to show up in their behaviour in a reorientation task of this sort.

A popular rationalist move to make at this point has been to note that even if infants don't manifest a concept in their behaviour, they might still be said to have the concept innately in virtue of having a disposition to use it. But this raises a further worry. How can we cash out the relevant dispositions in a way that is congenial to the rationalist position without trivializing this claim? After all, empiricists don't deny that children can acquire geometrical concepts and that children are therefore in some sense disposed to have an understanding of geometry. The danger, in other words, is that a reliance on a dispositional account may render the debate about innate ideas vacuous (Stich 1975). We ourselves think that this impasse can be resolved (see Chapter 2), but as the example illustrates, there are theoretical difficulties ahead. The debate about innate ideas may be an empirical issue, but that doesn't mean its solution can be read directly off of a given body of data.

Nevertheless, more empirical data can only help. Another potential source to consider is cross-cultural evidence. We saw earlier that Locke developed arguments bearing on the rationalism-empiricism debate based on what he took to be facts pertaining to the existence (or non-existence) of cross-cultural universals. Here too there is recent experimental evidence that is germane to the question of whether geometrical concepts are innate. Geometrical concepts and geometrical knowledge are common in large-scale industrial societies around the world. But what about remote small-scale societies in which there is no formal education, relatively little exposure to Western culture, and no experience with things like rulers, compasses, and maps? Izard et al. (2011) explored this question by studying the Mundurucú, an indigenous Amazonian people of just this type. Izard et al. queried Mundurucú children and adults to elicit intuitions associated with

Euclidean geometry, asking such questions as: Can a straight line be drawn through three non-aligned points? Can more than one straight line be drawn through the same two points? Given a line and a point not on the line, can a second line be drawn that passes through the point but never intersects the first line? Both Mundurucú children and adults did well even for questions that required them to form judgements about matters that go well beyond sensorimotor experience (e.g., conditions pertaining to infinite parallel lines). Izard et al. also asked Mundurucú children and adults to estimate the size of the third angle of a triangle (based on a depiction of the triangle's other two angles) using their hands or a custom-made device for indicating the sizes of angles. Their estimates were the same as Westerners, with averages for the sum of the resultant triangle's three angles nearly identical to 180°.

In another study, Dehaene et al. (2006) presented Mundurucú children and adults with a series of tasks with six images in which they were required to choose the one image that differed from the others. The test materials were carefully designed so that the exceptional image differed from the others in terms of a geometrical property (e.g., a diamond shaped parallelogram among five rectangles, or an open figure among five closed figures). Dehaene et al. also tested Mundurucú children and adults on a task that required using a map to locate a hidden item among three containers. Importantly, the maps that Dehaene et al. employed didn't indicate the geocentric orientation of the layout (north, south, etc.) and didn't include any features of the terrain. They only represented the containers and their geometrical relations to one another, with one container marked as the target location. In both of these experiments, Mundurucú participants performed significantly above chance, with both groups on a par with Western children.[11]

Undoubtedly, Locke would have found this work to be of much interest. It is also a major improvement on Plato's thought experiment. But it raises a number of theoretical questions regarding the import of cross-cultural data in the evaluation of concept nativism. The Mundurucú are an important test case given their lack of formal education and their unfamiliarity with compasses, maps, and so on. But this case doesn't demonstrate that geometrical concepts are *universal*. Indeed, it could be objected that it is impossible to show that a concept or system of representation is a human universal because it is impossible to examine people from every actual human culture (many are long gone), much less every possible human culture.

One might also wonder what conclusions could be drawn if the Mundurucú hadn't succeeded on any of these geometrical experiments. Would a single

[11] Interestingly, Western adults did significantly better on these tasks than the other groups (i.e., better than Western children, Mundurucú children, and Mundurucú adults). This suggests that formal education or greater experience with maps has an impact on the development or use of geometrical concepts beyond a shared baseline.

counterexample—a culture that apparently lacks geometrical concepts—disprove rationalism, as Locke seems to think? In our view, arguments from universality make an important contribution to the case for concept nativism, but these arguments have understandably been the focal point of much disagreement and it will take some time to sort out their import (see Chapter 11).

Another source of evidence that has been invoked in the debate between rationalists and empiricists stems from cases where a person has been deprived of normal experience because of a sensory deficit or because of lack of access to a normal environment. Molyneux famously called attention to this matter when he asked Locke whether a person who was blind from birth and who suddenly regained his sight as an adult would be able to distinguish a cube from a sphere just using vision, having previously only encountered cubes and spheres through tactile perception. Locke's answer was that they wouldn't be able to because the connections between tactile perception and vision need to be learned (Locke 1690/1975, II.ix.8). Whether Locke is right about the need for these connections to be learned or not,[12] research based on the reasoning embodied in Molyneux's question is well worth exploring. Might a deprivation experiment help to resolve the question of whether or not geometrical concepts are innate? For ethical reasons, there are of course limits to how much can be done with humans. However, there are natural experiments (cases of unplanned deprivation), and we can also turn to studies with animals.

As it happens, reorientation tasks like the ones performed by Hermer and Spelke have been successfully implemented with a diverse range of species, including instances in which the animals were subject to very strict controls regarding their experience of geometrical properties prior to testing. For example, Brown et al. (2007) reared fish (*Archocentrus nigrofasciatus*) from birth in either rectangular or circular tanks which had curtains around them so that the fish couldn't see beyond their tank and pick up on the geometry of the room or any landmarks it might contain. Once they had reached maturity at approximately 4 months of age, the fish received a brief amount of training in a rectangular enclosure analogous to the rectangular room used by Hermer and Spelke. The rectangular enclosure was inside a tank with other fish in the tank outside the enclosure. Each corner of the enclosure had a "door" but only one was open and this was the only one that permitted access to the area outside the enclosure. The purpose of the training was to get the fish to learn the location of this special

[12] It turns out that Locke was wrong about these connections needing to be learned. Although newly sighted adults do have difficulties relating visual perception to tactile perception, this isn't because the connections between the senses need to be learned. Evidence from newborn infants shows that there is an innate link between vision and touch (Sann and Streri 2007). On the other hand, he was right that newly sighted adults would struggle with this task, just for a different reason. The difficulties facing newly sighted adults stem from the fact that the neural tissue that would normally be devoted to visual perception doesn't lie idle in the blind, but rather takes on new functionality (Pascual-Leone and Hamilton 2001).

door—something that the fish were well motivated to do, as it allowed them to join a group of fish outside the enclosure. The fish were then tested using the same enclosure but with all four doors closed. Interestingly, when this enclosure consisted of four white walls—so no visible landmarks at all—the fish chose the correct door and its geometrically equivalent opposite door equally. What's more, this was true for both groups of fish—ones that had been reared in a rectangular tank *and* ones that had been reared in a circular tank. It would appear, then, that the fish were encoding the geometrical properties of the testing chamber regardless of whether they had the need or opportunity to navigate on the basis of this type of information prior to the experiment.

Studies in this vein might naturally be taken to suggest that non-human animals have innate mechanisms for representing the geometry of their environment.[13] Nonetheless, concerns might be raised that leave room to wonder exactly how much may be concluded from this type of research. One important point is that the critical test usually doesn't disclose a spontaneous response—it requires some training.[14] Also, the data are from a different species, not from humans. Even if a mechanism for geometrical representation is innate in fish or other animals, things might be different for humans. There is also a deeper methodological issue regarding the logic of deprivation experiments. Some have argued that it is impossible to impose perfect conditions of deprivation and thereby rule out the potential influence of an animal's environment, since there is always some kind of organism-environment interaction (e.g., Griffiths and Machery 2008). If this is right, then there may be a principled reason to think that deprivation experiments shouldn't be given much weight. Our own view is that deprivation experiments and data from comparative psychology can and do support concept nativism, but once again, there is work to be done in making clear how these sorts of considerations discriminate between the empiricist and rationalist viewpoints (see Chapters 4 and 10).

Yet another promising potential source of evidence concerns cases in which there is an impairment involving geometrical representation that has a genetic component. An important line of investigation of this sort focuses on Williams syndrome, a condition affecting individuals lacking a small set of genes (Landau and Hoffman 2012). It turns out that many people with Williams syndrome have inordinate difficulty with versions of the reorientation task that require attending to the geometrical properties of the room, yet they are perfectly able to use landmark information to locate a hidden object. While the exact nature of their spatial

[13] Similar results have been obtained for many species, including rhesus monkeys (Gouteux et al. 2001), rats (Cheng 1986), newborn chickens (Chiandetti and Vallortigara 2010; Chiandetti et al. 2015), and even bumblebees (Sovrano et al. 2012).

[14] Although, see Chiandetti et al. (2015) (experiment 2) for a variation on the reorientation task in which newborn chicks have no training and their experience of relevant geometrical properties is confined to the test conditions.

and navigational difficulties is a matter of ongoing investigation, there are indications that it is fairly specific (Lakusta et al. 2010; Landau and Hoffman 2012). For instance, when children and adults with Williams syndrome remain oriented in space, they can locate objects in a fixed location after they themselves have moved, performing at the same level as 3- to 4-year-olds who don't have Williams syndrome. In contrast, when they are disoriented and have to regain their orientation using just geometrical information, their performance can be so impaired that their choices are effectively random. Moreover, they show an idiosyncratic cognitive profile of relatively spared representational abilities accompanied by severe representational impairments. For example, Williams syndrome is associated with intact biological motion representation alongside deficits for other types of motion representation (Jordan et al. 2002; Reiss et al. 2005).

This work may be taken to suggest that geometrical abilities are innate because they can be selectively impaired in individuals with genetic anomalies. But as we will see in Chapter 3, the relationship between genes and the rationalism-empiricism debate is not at all simple or straightforward. Many researchers have argued that the fact that all traits depend on both genes and the environment shows that rationalist views about the origins of concepts are untenable. Some researchers have also argued that the developmental pattern associated with a genetic disorder like Williams syndrome can never provide good grounds for postulating innate concepts or special-purpose learning mechanisms because the whole idea of drawing conclusions about cognitive development in general from atypical cases is itself unsound (Karmiloff-Smith 2009a). Others may worry that while cognitive deficits associated with genetic anomalies might in principle be used to support these types of inferences, the details in this particular case can't support any rationalist conclusions. We disagree with these views, but they raise important objections that can't be ignored (see Chapter 20).

Finally, we will mention one further potential source of evidence bearing on the debate about innate ideas, involving research that is motivated by evolutionary theorizing or that is informed by considerations having to do with the evolutionary pressures that are likely to have shaped the human mind. We have seen that geometrical representation is involved in aspects of navigation. Spelke and Lee (2012) have suggested that a well-designed system of representation for navigation will focus on what they call an environment's *extended surface layout*—the three-dimensional contours of the environment's surfaces. Imagine standing in a natural environment that is shaped a little like a baseball field with a long ridge making an arc around the "outfield" and small groups of flowers or dry patches in the grass roughly where first, second, and third base would be. The ridge establishes part of the environment's three-dimensional contours, whereas the flowers and dry patches would be considered objects or surface markings that happen to occur inside the area, not features that define its contours. Spelke and Lee's thinking is that, in natural environments, such objects and surface markings may not

be reliable navigational cues. They can change relatively easily (an animal could eat the flowers) and can be difficult to distinguish from similar cues (different dry patches may all be pretty similar), and taking into account many of the details of such things could impose a high processing cost. By contrast, an environment's extended surface layout is likely to be more stable and reliable and can be represented more economically. It would therefore make good evolutionary sense for organisms to have a cognitive system capable of exploiting the geometrical structure of an environment.

By placing our navigational abilities in this evolutionary context, Spelke and Lee were able to make, and subsequently confirm, a series of otherwise highly surprising and subtle behavioural predictions. For example, 4-year-olds have been found to use certain geometrical properties of a rectangular arrangement of walls even when they are short enough to see over (30 centimetres high)—in fact, they will use the geometry of the walls even if the "walls" are short enough to be easily stepped over (2 centimetres high). However, they fail to exploit the same geometrical properties when they are instead properties of a coloured rectangular patch on the floor (a surface marking) or of four freestanding pillars that implicitly define a rectangle (landmark cues) (Lee and Spelke 2008, 2011) (see Figure 1.2).

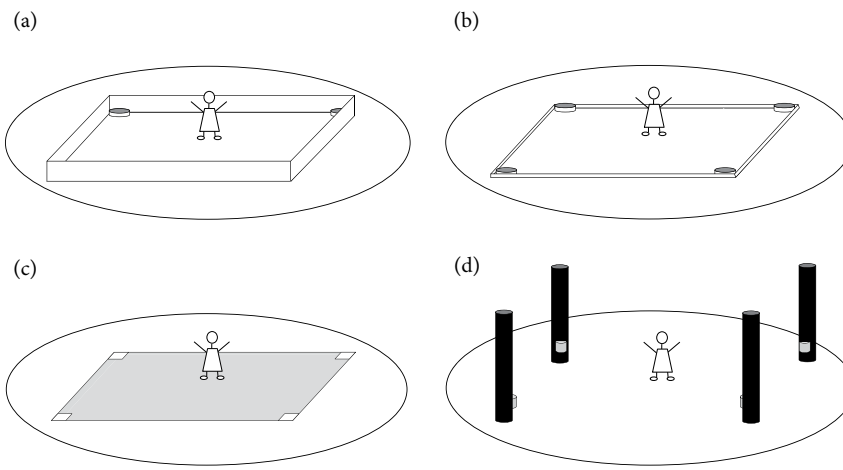

Figure 1.2 Reorientation experiment variations. In these variants on the basic reorientation experiment, (a) the walls of the rectangular enclosure are 30 cm high, (b) the walls of the rectangular enclosure are 2 cm high, (c) there are no walls but just a floor marking that covers same rectangular space, and (d) the rectangular space is framed by four freestanding columns. Children succeed in using geometrical cues in (a) and (b), but fail to use them in (c) and (d). Containers in the corners of the rectangular space varied across conditions; see Lee and Spelke (2008, 2011) for details. (Figure based on figure 1 in Lee and Spelke 2008 and figure 1 in Lee and Spelke 2011.)

This pattern of results is utterly bizarre if one doesn't take into account the navigational environments that our ancestors had to contend with. However, it starts to make sense if we suppose that selectional pressures stemming from navigational needs in our evolutionary history led to an adaptation for representing the large-scale three-dimensional arrangement of an area's extended surfaces. And if there is an adaptation for representing geometrical properties for purposes of navigation, this might naturally be taken to suggest that humans have an *innate* mechanism that represents geometrical properties.

That said, evolutionary arguments regarding the mind are immensely controversial, as is the field of evolutionary psychology in general. Many theorists question the viability of this approach to the study of mind on methodological grounds. Fodor has bemoaned "the outpouring of just-so stories by which the mainstream of evolutionary cognitive psychology is very largely constituted" (Fodor 2001, p. 627). And Stephen J. Gould (1997, p. 51) has charged that:

> Much of evolutionary psychology…devolves into a search for the so-called EEA, or "environment of evolutionary adaptation" that allegedly prevailed in prehistoric times. Evolutionary psychologists have gained some sophistication in recognizing that they need not postulate current utility to advance a Darwinian argument; but they have made their enterprise even more fatuous by placing their central postulate outside the primary definition of science—for claims about an EEA usually cannot be tested in principle but only subjected to speculation…the chief strategy proposed by evolutionary psychologists for identifying adaptation is untestable, and therefore unscientific.

We ourselves don't think these objections hold up. Good evolutionary theorizing about the mind isn't a matter of making up stories and its claims aren't untestable. Seeing why will require some reflection on a number of general methodological issues (Chapter 4) but also, and perhaps more importantly, working through some examples in which evolutionary arguments inform rationalist theories of the human conceptual system (see Chapters 14 and 15).

As we hope the example of geometrical representation illustrates, we have come a long way since Plato first speculated about innate geometrical knowledge. What's more, the experimental findings briefly mentioned in this section represent only a tiny fraction of a huge enterprise in which researchers in cognitive science have, in effect, taken up Chomsky's call to approach the dispute over innate ideas as an empirical question of the first importance. But as the example of geometrical representation also illustrates, the fact that we are now dealing with an empirical issue about the mind doesn't mean that it is going to be a straightforward matter to settle. Over the course of the book, we will see that there is an enormous wealth of similarly exciting findings that can feed into a case for concept nativism beyond anything historical philosophers like Plato and

Descartes could ever have imagined. But we will also see that building this case involves working through a landscape full of theoretical and philosophical difficulties (like the ones we have briefly mentioned here) that are associated with all of this data and that bear on how best to interpret it and what it can truly tell us about the origins of concepts.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0001

PART I

THE RATIONALISM-EMPIRICISM DEBATE

2

What the Rationalism-Empiricism Debate Is Really About

Our principal objective in this book is to defend concept nativism. Our positive case for concept nativism (in Part II) is built around a series of independent but complementary arguments that combine a wealth of empirical and theorical considerations. Taken together, they form a powerful inference to the best explanation argument for a rationalist account of the origins of many concepts across many different conceptual domains. But while this is our principal objective—to argue for our version of concept nativism—we have a second equally important aim that will be the main focus in Part I of the book. This second aim is to comprehensively rethink the rationalism-empiricism debate about the origins of psychological traits and to clarify the key theoretical notions that are critical to the debate. This undertaking is of fundamental interest in its own right. But it is especially pressing in light of the challenges and objections that have been raised to pursuing a project like ours. Many contemporary theorists see the rationalism-empiricism debate as resting on an irreparably confused theoretical foundation, leading them to suppose that it is a completely worthless debate that should just be abandoned. Others take rationalism in particular (regarding any type of psychological trait) to be so deeply flawed as to warrant dismissing the entire rationalist framework as not being worthy of consideration. We completely disagree with these assessments, but they underscore the urgent need to get clear about and what the rationalism-empiricism debate is really all about and what rationalism is truly committed to. If these views regarding the value and coherence of the rationalism-empiricism debate or the viability of rationalism in particular were warranted, our project would be doomed from the start. Moreover, among those who maintain that the rationalism-empiricism debate remains valuable, or are active participants in the debate, there nonetheless is much confusion about precisely what the debate is about; it is not uncommon for both rationalists and empiricists to conflate several incompatible understandings of the debate without realizing it. What all of this means, in our view, is that a thorough rethinking of the foundations of the debate is long overdue. So, while the book as a whole argues for a rationalist account of the origins of concepts, Part I is meant to establish the integrity of the foundations of the rationalism-empiricism debate and the viability of rationalist approaches to the origins of psychological traits, laying the theoretical groundwork for the rest of the book.

We will begin, in this chapter, by spelling out in considerably more detail what we take the rationalism-empiricism debate to be about. The account that we develop expands upon the sketch given in the previous chapter by introducing a number of important new distinctions, clarifying the key theoretical notions that are critical to the debate, and elucidating the dimensions of variation that help to differentiate one view from another within the large spectrum of possibilities. Once this account is in place, we will be in position to address the charge that our project is doomed from the outset. This charge stems from two types of criticisms—ones that are directed at the entire rationalism-empiricism debate and ones that are directed at rationalism in particular. In Chapter 3, we address the first of these two sets of criticisms, a collection of interrelated challenges that call into question the value and coherence of the rationalism-empiricism debate as a framework for theorizing about the origins of any type of psychological trait. Many of the critics who raise these challenges equate the rationalism-empiricism debate with the so-called *nature-nurture debate*, understood as a dispute about the relative contributions of genes and the environment. And so addressing these challenges requires evaluating the nature-nurture debate and its relation to the rationalism-empiricism debate. In Chapter 4, we turn to the second set of criticisms, in which it isn't the rationalism-empiricism debate that is called into question but rather the value and coherence of just one side of this debate—the rationalist side. The charge here is that rationalism is so fundamentally flawed that it can essentially be dismissed in advance of any detailed consideration of the arguments and evidence that might be given in its favour. In addition to addressing a number of specific charges along these general lines, Chapter 4 briefly looks at two of the most important types of argument for adopting a rationalist approach to the origins of at least some psychological traits.

Our discussion in Chapters 2, 3, and 4 is largely focused on the rationalism-empiricism debate with respect to psychological traits in general—that is, as it might pertain to a psychological trait of any type (not just concepts). Collectively these chapters show how the rationalism-empiricism debate is not only perfectly coherent but integral to understanding the human mind, and that far from being fatally flawed or riddled with confusions, rationalism can in principle provide a powerful theoretical framework for explaining the origins of a variety of psychological traits. In Chapters 5 and 6, we shift our attention from the rationalism-empiricism debate in general, which concerns the origins of many different types of psychological traits, to the rationalism-empiricism debate regarding the origins of concepts—the debate between concept nativism and concept empiricism. In Chapter 5, we draw attention to how superficial and illusory explanations of development can be difficult to see for what they are and how this can illicitly lead to the neglect of rationalist accounts of cognitive and conceptual development. We illustrate the detrimental appeal of such illusory explanations with the enormously influential empiricist idea that concepts are acquired via a process of abstraction. We also show that

abstraction can be rehabilitated and turned into a substantive account of conceptual development, but that doing so involves abandoning much of what has made it attractive to its empiricist advocates. Finally, in Chapter 6, we discuss theories of the nature of concepts and their bearing on the debate between concept nativism and concept empiricism. We also say more about how we understand innateness, and why concept nativism is about more than just innate concepts.

2.1 Philosophical Hostility to the Rationalism-Empiricism Debate

Following his groundbreaking work in linguistics in the late 1950s and early 1960s which essentially founded the contemporary paradigm for the scientific study of language, Chomsky began to explore the broader theoretical context for this work. As we noted in Chapter 1, this gave rise to the contemporary incarnation of the rationalism-empiricism debate, reviving a debate that had been relatively dormant since its heyday in the sixteenth and seventeenth centuries. This revival of the rationalism-empiricism debate was quickly greeted by a substantial amount of philosophical hostility. Writing in 1966, Nicholas Rescher noted that:

> The mention of innate ideas evokes little responsive sympathy in modern philosophers. When, for example, I inquired of Nelson Goodman regarding publication of a recent symposium on the topic on which he had participated, he wrote in reply that: "Personally, I am rather of two minds about seeing this published since I feel in some ways that the less attention given the matter of innate ideas the better" (Letter of 22 July I965). (Rescher 1966, p. 205)

Incredibly, Goodman's remark to Rescher understates his animosity towards rationalist accounts of cognitive and conceptual development. In what we assume to be the paper from this symposium, Goodman presents his views on innate ideas in the form of a dialogue. He explains:

> This recasting and expansion of the material in my symposium talk 'On Some Inimical Ideas' reflects no literary ambitions. The dialogue form offered advantages both in organization and in giving an appropriate tone to discussion of a theory that only my respect for its advocates [i.e., Chomsky] enables me to take at all seriously. (Goodman 1967, p. 23)

In the dialogue, Goodman presents the rationalist proponent as being forced into the position of maintaining that their position amounts to "the trivial truth that the mind has certain capacities, tendencies, limitations" (p. 28). Goodman then goes on to summarize his assessment of Chomsky's proposal that innate ideas play a role in language acquisition by accusing Chomsky of making

"unsubstantiated conjectures that cry for explanation by implausible and untestable hypotheses that hypostatize ideas that are innate in the mind as non-ideas" (p. 28).

While Goodman's way of expressing his hostility to the rationalism-empiricism debate is particularly colourful, his exasperation at the thought of revisiting the rationalism-empiricism debate was by no means unique. Writing at much the same time, P. F. Strawson, another major figure in twentieth century philosophy, also saw little of value in the historical debate between rationalists and empiricists, claiming that:

> whatever genuine questions were at issue in these debates, they tended to be hopelessly obscured by the terms in which the debates were conducted. Those terms are rich enough, over-rich, in metaphorical suggestion: ideas as characters written on the tablets of the mind (copied from experienced originals or inscribed by the hand of God); or ideas as furnishings of the mind's house (picked up at that general store, experience, or built-in structural features). Even if a determined effort is made to escape from such pictures, the debate about origins is apt to remain a sterile exchange of points: on the one hand, that all capacities to think, recognize, classify, etc., have to be acquired (for an infant does not think at all), on the other, that the acquisition of such capacities presupposes the capacity to acquire them. (Strawson 1966, pp. 68–69)

In many areas of philosophy and cognitive science, this sort of hostility to the rationalism-empiricism debate has never really abated. In fact, in many ways it has only intensified. For this reason, it is important to show that it is unfounded. Since Goodman's and Strawson's views are not at all unrepresentative and provide an instructive contrast with our own views, it pays to see how such views go wrong.[1]

On Goodman's and Strawson's reconstructions of the rationalism-empiricism debate, there is no substantive issue at stake since they take both sides in the debate to endorse the truism that infants possess the capacity to acquire cognitive capacities. Why do they see it this way? The key factor, we'd suggest, is their assumption that rationalists cannot actually hold that infants are truly in a position to think and reason or that they possess genuine representations, much less abstract representations that, in Goodman's words, are "implanted in the mind as original equipment" (Goodman 1967, p. 27). In that case, rationalists must have some other understanding of what it is for ideas and cognitive abilities to be innate, one that isn't a complete non-starter. The alternative, for Goodman and Strawson, is that infants are merely predisposed to acquire these things. This in

[1] Later, in Chapter 3, we will examine how this hostility to the rationalism-empiricism debate has developed in more recent discussions of the debate, both in philosophy and cognitive science, and why such concerns continue to be misplaced.

turn amounts to nothing more than the claim that infants, in some sense, have the capacity to acquire whatever capacities or ideas they eventually acquire—a claim that is unobjectionable yet so trifling that it extinguishes any interest there could be in the debate between rationalists and empiricists.[2]

Strawson gives no justification for his claim that "an infant does not think at all". He just asserts it. Still, it is not hard to discern the rationale behind this assertion. The reason so many philosophers have held that there isn't much happening in infants' minds is the fact that it isn't apparent from their behaviour that there is. The problem with this line of thinking, however, is that it ignores the *competence-performance distinction*—the distinction between the capacities of an organism or system (competence) and its actual behaviour under a given set of circumstances (performance).

Complex systems typically have many capacities that are not evident in their behaviour depending on the circumstances. Standard laptop computers, for example, can do word processing, arithmetical calculations, video editing, and so on, but of course they don't typically do *all* of these things *all* of the time. A laptop that happens to be engaged in a word processing task, such as inserting a footnote into a document, may not, at the same time, be exercising its ability to crop photos. Yet it still has the ability (competence) to crop photos, and not merely in the sense that the laptop is the type of thing that can crop photos in principle (e.g., by acquiring this ability later on). Rather, the point is that it currently possesses a mechanism for cropping photos, a mechanism that is in place and ready to be exercised. The laptop's photo-cropping competence simply isn't evident while the machine's operations are directed elsewhere.

Various circumstances can also impede a system's performance despite its competence remaining intact. If there is a power outage or if the laptop's battery runs out of charge, it still has the ability to do word processing. Just recharge the battery or plug the computer into another power source, and once again it will take on a word processing task just as before. (By contrast, if its word processing program were deleted, then the laptop would no longer possess the competence, though it would of course still be the kind of system that could acquire—or reacquire—this competence.)

Likewise, the laptop may have abilities that it isn't able to exhibit in its behaviour owing to the interaction of the mechanisms that implement these abilities with other mechanisms. Its word processing function could be hampered by other programs monopolizing its random-access memory. Or, to take another example, its ability to catalogue consumer products by their barcodes might

[2] Goodman's and Strawson's views of the rationalism-empiricism debate trace back to Locke: "if the Capacity of knowing be the natural Impression contended for, all the Truths a Man ever comes to know, will, by this Account, be, every one of them, innate; and this great Point will amount to no more, but only to a very improper way of speaking; which whilst it pretends to assert the contrary, says nothing different from those, who deny innate Principles" (Locke 1690/1975, I.ii.5, p. 50).

require higher resolution input than its camera is able to deliver. In this case, the laptop could nonetheless possess the ability for identifying different barcodes given appropriate input. It is just that in order to tap this competence, a higher resolution camera would be needed to provide the appropriate input. In short, there are many reasons why a system's performance (its readily observable behaviour) might fail to reflect the full range of competences it possesses (these being its possibly hidden or inactive abilities, which may require other conditions to be satisfied in order to be manifested).

Needless to say, infants are considerably more complex than laptops. (Infants' brains contain billions of neurons with the same general organization as the adult brain, including its six horizontal layers of neocortex; see Marcus 2004; Larsen et al. 2006.) So we should expect there to be any number of factors that can interfere with their ability to display the cognitive and representational abilities they may possess. Here are a few examples. Infants may have insufficient muscle strength to control their bodies in appropriate and telling ways. They may have insufficient coordination for expressing their intentions in action. And, given their limited memory and attention capacity, they may have insufficient memory and attention for a given type of task and become distracted or to lose track of what they are doing. This is not to say that every instance in which infants appear to lack a given ability must be due to some performance factor obscuring a hidden competence. However, in many specific instances, this possibility ought to be entertained—something that Strawson and Goodman clearly fail to do. So, perhaps it is not true that, as Strawson puts it, "an infant does not think at all", but only that an infant's thinking—its ability to recognize, classify, draw inferences, and so on—isn't obviously manifested in its readily observable behaviour.

In considering Goodman's anti-rationalist sentiments, it is also important to bear in mind that it isn't just *rationalists* who hold that there are abilities or ideas that are "implanted in the mind as original equipment". This is something that is common to rationalists *and* empiricists. As W. V. O. Quine (who was himself an empiricist and a behaviourist) once noted, even "the behaviorist is knowingly and cheerfully up to his neck in innate mechanisms" (Quine 1969a, p. 57). Moreover, contrary to what Goodman suggests, the posited innate mechanisms are necessarily a matter of speculation for both empiricists and rationalists alike; the nature and workings of such mechanisms will only be discoverable through an enormous amount of painstaking theoretical and empirical work.

Finally, the most crucial point that both Goodman and Strawson seem to miss is that, while rationalists and empiricists do both posit innate psychological structures, they don't posit the *same* psychological structures. This fact alone shows that the debate between rationalists and empiricists doesn't boil down to the truism that all infants possess the capacity to acquire the various cognitive capacities that they will acquire. To be sure, both rationalists and empiricists do accept this truism—it is a truism after all. But there is still plenty of room for

them to disagree about the nature of the psychological underpinnings of cognitive development and, therefore, to disagree about what infants' capacity for acquiring various cognitive capacities involves.[3]

We can see this point by returning to what the rationalism-empiricism debate looks like on an account like the one that we sketched in Chapter 1. As was noted there, empiricists posit a highly limited set of distinct types of psychological structures or traits as the innate basis for acquiring other psychological traits, and they suppose that the very same learning mechanisms underlie disparate types of acquired psychological traits.[4] Since the same learning mechanisms are responsible for the acquisition of psychological traits of many different kinds and aren't restricted to any particular domain, these are often said to be *domain general*. When empiricists do allow that there are *domain-specific* learning mechanisms—special-purpose systems that are tied to a narrower range of psychological traits and that are restricted to particular domains—they typically claim that these are learned, and that the learning is achieved on the basis of more fundamental domain-general learning mechanisms. Empiricists are also extremely frugal when it comes to innate concepts. Empiricists maintain that most concepts, maybe even all, are acquired almost exclusively on the basis of these domain-general learning mechanisms.

Chapter 1 noted that rationalists, by contrast, posit many distinct types of psychological structures as the innate basis for acquiring psychological traits. In addition to domain-general learning mechanisms, rationalists posit a large number of specialized components of the mind as part of the innate foundation for cognitive and conceptual development. These innate specialized components (which rationalists take to not be acquired on the basis of more fundamental domain-general psychological mechanisms), figure in specialized learning mechanisms for acquiring further psychological traits. Rationalists also often embrace a significant number of innate representations of different types, including a variety of innate concepts. Thus, rationalists will typically hold that many concepts stand alongside the innate domain-general and domain-specific components which they maintain are the psychological starting point for subsequent learning.

This brief characterization highlights what we take to be the general nature of the contrasting commitments of rationalism and empiricism. The key point to notice in relation to Goodman and Strawson is that, on this understanding of the rationalism-empiricism debate, the debate doesn't simply reduce to two

[3] This point applies as much to historical rationalist views as it does to contemporary ones. Locke was certainly right that the rationalists in his time explicated their rationalist views of the mind in terms of dispositional properties (see, e.g., Descartes 1648/1985, p. 304; Leibniz 1705/1996, p. 52). However, this doesn't mean that there would be no substantive distinction between the types of dispositional properties that figured in historical rationalist and empiricist theories.

[4] We will use "psychological structure", "psychological trait", and "cognitive trait" interchangeably as generic terms for any type of psychological entity, including mental representations, knowledge structures, processing mechanisms, processing links, and so on.

contrasting ways of asserting the same truism—that all infants possess the capacity to acquire the various cognitive capacities they will acquire—with no substantive differences between the two sides in the debate. On our view, there are very clear differences between rationalism and empiricism. We take this to be a major selling point of our approach, since there really is a substantial theoretical disagreement between rationalists and empiricists. Eliminating the debate by interpreting it in such a way that nothing could be at stake in the debate doesn't make these very real differences go away.

While this brief sketch of how we understand the rationalism-empiricism debate succeeds in conveying much of its spirit, it leaves out many important details that will matter later on. To properly make sense of the debate—and to avoid the many confusions surrounding it—this basic account needs further elaboration. That will be the job of the remainder of this chapter. The account that we will present is not entirely new; in broad outlines, it shares key features with accounts that others have endorsed.[5] However, there is much that is new. We will need to clarify theoretical notions that have been disputed or considered problematic and introduce a number of new terms to highlight important distinctions that have been neglected or overlooked before arriving at a comprehensive statement of how we understand rationalism, empiricism, and the rationalism-empiricism debate. All of this will all take some time to spell out, and we will approach it in stages, with each stage adding further details to the account.

2.2 The Acquisition Base

The first idea that we need to introduce is that of the acquisition base. The *acquisition base* posited by a theory is the collection of psychological structures that the theory takes to be *psychologically primitive* (in the sense that they are not acquired via any psychological process of acquisition) and to provide the basis for acquiring further psychological traits (Margolis and Laurence 2013).[6] The structures in the acquisition base provide the ultimate psychological basis for explaining the origins of all psychological traits. Learning has to start somewhere, and the acquisition base is where it starts.

[5] As we will see in Chapter 3, both rationalists and empiricists often state their disagreement in broadly the same way that we have characterized the debate. On the other hand, we will also see that theorists on both sides of the debate can fall prey to errors and confusions in formulating the terms of the debate and that both participants in the debate and critics of it frequently conflate the core idea of our account with incompatible understandings of the debate that turn out to be intellectual dead ends—especially the view that it is about nature versus nurture (or the relative contributions of genes versus the environment).

[6] The idea of a psychological primitive is introduced in Samuels (2002) in the context of developing an account of what makes a psychological trait innate (see also Cowie 1999). We discuss Samuels' view of innateness in relation to our own in Chapter 6.

Box 1

Acquisition Base—the acquisition base posited by a theory is the collection of psychologically primitive psychological structures that the theory takes to be the ultimate psychological basis for explaining the developmental origins of all psychological traits.

A fundamental point regarding the acquisition base is that *all* theorists who hold that there are any psychological traits at all are committed to the existence of an acquisition base of one kind or another—whether they are aware of it or not and regardless of where they stand in the rationalism-empiricism debate. This is because, for any given psychological trait, either that trait will be acquired on the basis of other psychological traits (learning mechanisms, for example) or it won't be. If the trait isn't acquired on the basis of other psychological traits, then it is in the acquisition base simply in virtue of that fact. And if the trait is acquired on the basis of other psychological traits, the same point will apply to whichever psychological traits mediate its acquisition—those psychological traits will either be acquired on the basis of still further psychological traits (other learning mechanisms, for example) or not. Since this process can't continue indefinitely, all theorists who posit psychological structures of any kind at all are committed to the existence of an acquisition base. None of this is to say that everyone is committed to *the very same* acquisition base or even to acquisition bases that are particularly similar to one another, just that everyone is committed to some acquisition base or other.

Now that we have the idea of an acquisition base to work with, we can use it to reformulate our initial account of the differences between rationalist and empiricist accounts, which was based on the psychological structures which each of these approaches takes to be innate.[7] To a first approximation, on this reformulation of the debate, what rationalists and empiricists disagree about is the character of the acquisition base. In particular, rationalist and empiricist views differ regarding the number and types of psychological structures that they take to be in the acquisition base.[8] Empiricist views take the acquisition base to be very

[7] Recall from Chapter 1 that rationalism and empiricism are not specific theories. Each is a theoretical framework within which many specific theories can be developed, including competing theories within each framework for explaining the very same psychological traits. This means that there is a wide spectrum of theories at issue—with both a range of rationalist theories and a range of empiricist theories—and that we need to be sensitive to how they differ regarding the contents of the acquisition base (which on all views will provide the ultimate psychological basis for cognitive and conceptual development).

[8] This can be put as a disagreement about what is in the acquisition base or as a debate about which type of acquisition base should be accepted as correct. We will also use the terms *empiricist acquisition base* (an acquisition base associated with an empiricist account) and *rationalist acquisition base* (an acquisition base associated with a rationalist account).

sparse relative to rationalist views. They see the acquisition base as containing a highly limited set of distinct types of psychological structures and, for the most part, suppose that there is a limited number of instances of these types as well. These psychological structures provide the ultimate basis for acquiring all other psychological traits. Since the same highly limited number of psychological structures is responsible for the acquisition of psychological traits of many different kinds, they must comprise learning mechanisms that are generally operative across many different domains, that is, they are domain-general learning mechanisms. Such domain-general learning mechanisms form the core of any empiricist acquisition base.[9] When empiricists do allow that there are domain-specific learning mechanisms, they typically claim that these are learned and so are not part of the acquisition base. Empiricists are also very frugal in terms of the types of representations they admit into the acquisition base. Empiricists will typically accept that low-level sensorimotor representations are in the acquisition base, but hold that there are few, if any, concepts or abstract representations in the acquisition base. Empiricists maintain that most concepts, maybe even all, are acquired almost exclusively on the basis of domain-general learning mechanisms.

What about rationalism? Rationalist views hold that the acquisition base isn't restricted to domain-general learning mechanisms and low-level sensorimotor representations. It contains a substantial number of more specialized kinds of psychological structures as well. In particular, rationalists typically hold that the acquisition base contains a number of domain-specific learning mechanisms or psychological structures that contribute to specialized learning mechanisms. Rationalists also often hold that there are abstract representations of different types in the acquisition base, including, on many rationalist accounts, a significant number of concepts.

This sketch of how rationalist and empiricist accounts of the acquisition base differ is a good first approximation and will do for the moment. Later, we will have more to say about the types of psychological structures in rationalist and empiricist accounts of the acquisition base. For now, we want to turn to a discussion of several general theoretical points connected with this way of characterizing the debate.

The first is that in saying that a psychological structure is part of the acquisition base, we are saying that there is no explanation at the psychological level—for example, no learning-based account—of the acquisition of this structure. This doesn't mean that there is no way to account for the origins of the structure at all, or that its acquisition is in any way mysterious. Its origins will still be explainable. We assume that there is a broadly physical story in terms of biological processes

[9] This makes sense since positing domain-general learning mechanisms that apply across many different content domains allows empiricists to posit a sparser acquisition base, as such mechanisms can be used across many domains rather than having different mechanisms for different domains.

(neural, genetic, and so on) regarding the origins of all psychological structures. For many psychological structures, there is also a psychological story of the origins of the structure which is grounded in these lower-level physical processes. What makes it the case that a psychological structure is in the acquisition base is that there is no psychological-level acquisition story grounded in the lower-level physical processes; there is *only* the lower-level story of its origins. So, in saying that a psychological structure is primitive, all we are saying is that it is not acquired via psychological-level processes, not that there is no explanation of its origins at all.[10]

The next point that we want to highlight concerns the relation between this formulation of our account of the rationalism-empiricism debate—in terms of the notion of an acquisition base—and our initial way of framing the debate in terms of what is innate. One way of looking at the characterization of the debate in terms of the notion of an acquisition base is that this characterization essentially bypasses issues about what innateness is and the various controversies that are tied to how best to understand that notion. That is, it can be seen as providing an alternative way of formulating the debate, which doesn't require the notion of innateness and perhaps even paves the way for its elimination. A different way of looking at this characterization is to see it as providing a way of explicating or spelling out what it is for a psychological trait to be innate: A psychological trait's being innate amounts to its being part of the acquisition base, that is, to its being a psychological trait that isn't learned or acquired via any psychological-level process of acquisition.

For present purposes, there is no need for us to decide between these two options. While we ourselves think that it is appropriate and useful to understand the rationalism-empiricism debate in terms of the notion of innateness, and so are happy to see the characterization in terms of the acquisition base as an explication of the notion of innateness, we are aware that many theorists find the notion of innateness to be problematic.[11] Theorists who are sceptical of the notion of innateness are free to treat the characterization in terms of the acquisition base as a way of avoiding any reference to innateness. In any case, going forward our discussions will primarily be framed in terms of the acquisition base, since this is what underpins our account, regardless of whether this understanding is taken to provide an explication of the notion of innateness or not. And for readers who

[10] The non-psychological processes involved in the acquisition of such structures are often conceptualized as biological maturation processes involving the unfolding of biological processes of development that eventuate in the development of such psychological structures in much the way that these kinds of processes eventuate in the development of structures in bones or the liver, for example. One's liver isn't the product of learning, but this doesn't mean that its acquisition is intrinsically mysterious, even if the details of its development are complex and not yet fully understood.

[11] In our view, using the notion of innateness to characterize the rationalism-empiricism debate nicely links contemporary theorizing about this debate to the historical debate about innate ideas and to the resurgence in interest in these historical views in the early days of cognitive science (see Chapter 6 for more discussion of this issue).

prefer to avoid the notion of innateness altogether, any references to innate traits in what follows can be read instead in terms of these traits being part of the acquisition base.

A related consequence of our account of the rationalism-empiricism debate in terms of the notion of the acquisition base is that it makes explicit and so highlights the fact that this debate is not a debate about the notion of innateness. Some philosophers, however, have seen the rationalism-empiricism debate in just these terms—as a debate about whether there is a viable notion of innateness and, if there is, what this might be (see, e.g., Stich 1975). On this way of looking at things, rationalists are seen as embracing the notion of innateness, while empiricists are seen as being sceptical about whether the notion withstands scrutiny and so as rejecting the notion. It should be clear that, from our perspective, construing the debate in this way is a mistake. This is because both rationalists and empiricists are equally dependent on there being a viable notion of innateness. On our initial characterization of the debate, it is clear that both rationalists and empiricists posit innate psychological traits—what they disagree about is which kinds of psychological traits are innate, not whether any psychological traits are innate. And on our reformulation of the debate in terms of the acquisition base, rationalists and empiricists are still equally dependent on there being a viable notion of innateness. If the idea of the acquisition base is seen as explicating the notion of innateness, then they are equally committed to there being innate psychological traits; if it is seen as eliminating the notion of innateness, then neither has this commitment. In both cases, rationalists and empiricists are either both committed to a notion of innateness or neither is.

A related mistaken idea about how rationalists and empiricists differ from one another is that rationalists hold that something is innate and that empiricists deny this. Our account of the debate also makes clear that this isn't right. As we argued earlier, any theorist who acknowledges the existence of any psychological traits at all is committed to there being an acquisition base, so empiricists, just like rationalists, have to accept that there are psychological traits that are psychologically primitive. If innateness is understood as being explicated in terms of the notion of being psychologically primitive, then clearly everyone must accept that there are at least some innate psychological traits (though of course different theorists would take different psychological traits to be innate). On the other hand, if the understanding of the debate in terms of the acquisition base is taken as a way of avoiding or eliminating the notion of innateness, then neither rationalists nor empiricists would be committed to there being anything that is innate. Either way, the debate wouldn't be about whether anything is innate.[12]

[12] Samuels' (2002) discussion of innateness and the rationalism-empiricism debate raises several subtle issues that are related to the ones we have been discussing in the last two paragraphs. As we noted earlier, Samuels' notion of being psychologically primitive is essentially the same as our idea of

Returning to the notion of an acquisition base, it is important to recognize that this notion is not indexed to any particular point in the human lifespan, and so in particular is not tied to the time of birth. This is because what it means for a psychological trait to be in the acquisition base has nothing to do with when during the course of development that trait appears. All that matters to a trait's being in the acquisition base is whether it is psychologically primitive. A trait doesn't have to be present at birth, for example, in order to be psychologically primitive and so to be part of the acquisition base. Psychologically primitive traits may come to be part of the mind at various times in life. There is nothing strange or mysterious about any of this. Just as some bodily traits develop later (e.g., teeth, secondary sexual characteristics), the same may be true of psychological traits. While the notion of an acquisition base doesn't mandate that this is the case for psychological traits, it is perfectly compatible with this possibility.

Similar considerations suggest that a commitment to there being an acquisition base doesn't entail that there is a fixed structure to the mind. There is nothing about a psychological trait's being part of the acquisition base that means that it cannot be altered or even eliminated later in development. A psychological trait might develop in stages, with earlier forms of the trait that were in the acquisition base being replaced by later forms either through learning or through maturation. And a psychological trait in the acquisition base might be altered, or simply replaced or overridden, through any number of psychological processes based on experience or cultural influences, just as for learned traits. The fact that a psychological trait is not itself learned or acquired via psychological processes is completely neutral with respect to changes of all these types—they are neither mandated nor excluded by the fact that the trait is part of the acquisition base.

a psychological trait being in the acquisition base. He uses this notion to address what he calls the *problem of special nativism*, which is about what it is for a psychological trait to be innate. Samuels contrasts this problem with what he calls the *problem of general nativism*, which is about "the general distinction between nativism and non-nativism in cognitive science"—in our terms, what the rationalism-empiricism debate is about. While we agree with the core of Samuels's account of what innateness is (see Chapter 6 for how our account differs from his beyond this core idea), we disagree with his understanding of how questions about innateness are related to the rationalism-empiricism debate. As he sees it, "claims about the innateness of specific (kinds of) cognitive structure…[are] equivalent to…[being] a nativist with respect to some specific (kind of) cognitive structure" (p. 234). This way of seeing things is problematic for at least two reasons. First, since, as we have just noted in the text, everyone must take something to be innate in the sense of being psychologically primitive, this means that all theorists—including empiricists—end up being rationalists/nativists (i.e., rationalists about the traits they take to be primitive). But it's just confusing to say that empiricists are rationalists/nativists. Second, and more importantly, it is a mistake to take being a rationalist about a psychological trait to be equivalent to taking that trait to be innate, since this neglects the crucial role of learning on rationalist accounts (this point will play a major role in our discussion below). In our view, neither of the two questions Samuels addresses should be framed as problems about nativism/rationalism, since neither is more about rationalism than it is about empiricism. And using the same term ("nativism") to get at what is at stake in both of these debates invites conflations between the very different issues that are at stake in the two debates (conflations of the sort that arise in some places in Samuels' discussion of what he describes to be "constraints on nativism in cognitive science").

So far, we have been discussing the rationalism-empiricism debate as if it were a single debate. However, in reality, there are many rationalism-empiricism debates pertaining to the origins of different psychological traits. For example, there is a rationalism-empiricism debate regarding the origins of language. There is a rationalism-empiricism debate regarding the origins of our understanding of geometrical concepts. There is a rationalism-empiricism debate regarding the development of personality traits. There is a rationalism-empiricism debate regarding the origins of aesthetic appreciation. There is a rationalism-empiricism debate regarding the origins of the capacity for walking. There is a rationalism-empiricism debate regarding the origins of a sense of fairness. And so on, for *many* other psychological traits. Some of these concern areas with a long history of controversy, some have only recently started to receive serious attention. But whether they are entrenched debates or just beginning, this list only scratches the surface of the range of possible debates. There are as many rationalism-empiricism debates as there are types of psychological traits whose origin can be debated.

One broad distinction among all these debates that is useful for the purpose of getting oriented is the distinction between what might be thought of as local rationalism-empiricism debates and global rationalism-empiricism debates. By a *local debate*, we mean one that is largely restricted to a specific type of psychological trait of interest. For example, a debate might focus just on the origins of language without paying much attention to the origins of musical cognition, numerical cognition, or any other psychological trait. Likewise, a debate might focus just on the origins of logical concepts (OR, AND, IF-THEN, and so on) without taking a stand on the origins of other types of concepts or, for that matter, on the origins of any other type of psychological trait.

In contrast, a *global debate* is concerned with the origins of a broad category of psychological traits. The most encompassing global rationalism-empiricism debate is about the totality of human psychological traits or structures. Ultimately, a truly encompassing debate of this kind would be informative about the origins of literally every psychological trait whose acquisition is explained in psychological terms, providing an exhaustive account of the human acquisition base. It would span everything from the origins of a simple fleeting preference (like preferring to eat out one night) to the origins of a fundamental psychological faculty (like a long-term memory system). It would cover the origins of psychological traits as diverse as those involved in a knowledge of Latin, the capacity to play a minor scale on the piano, a love of baseball, the ability to drive a car, an understanding of tort law, the ability to make coffee, the desire to imitate a celebrity's way of dressing, the ability to solve complex problems in thermodynamics, an understanding of the social conventions in a classroom, the concept FISH TACO, and so on for a truly vast

number of other psychological traits.[13] One less encompassing but still global rationalism-empiricism debate is the debate about the origins of human concepts. This debate—our central focus in the book—concerns the origins of concepts in general, not just the origins of concepts in a single conceptual domain or the origins of some particular set of disputed concepts. What is at issue is the developmental origins of the human conceptual system taken as a whole.

Box 2

Global Rationalism-Empiricism Debate—a global rationalism-empiricism debate is concerned with the developmental origins of all of the psychological traits of some broad type (e.g., the debate concerning the developmental origins of all concepts).

Local Rationalism-Empiricism Debate—a local rationalism-empiricism debate is concerned with the developmental origins of a particular type of psychological trait of interest (e.g., the debate concerning the developmental origins of the language faculty) or a narrowly circumscribed set of traits (e.g., concepts in a single domain, such as natural number concepts).

Note that a rationalist position in a global rationalism-empiricism debate is compatible with empiricist accounts in some or even many local rationalism-empiricism debates. This is true whether this is the global debate about the origins of the totality of all human psychological traits or a more restricted global debate like the debate regarding the origins of all concepts. For example, if a theorist were to adopt a rationalist account of the origins of language, personality traits, causal reasoning, emotions, social exchange, and normative concepts, this would be more than enough to make them a rationalist in the broadest global debate between rationalists and empiricists even if they didn't also adopt rationalist accounts of a fear of spiders, imitation, or musical ability. And a theorist who adopts a rationalist account regarding concepts in a number of core conceptual

[13] Of course, no theorists are actually trying to offer specific developmental accounts for every psychological trait—an impossibly immense undertaking. On the other hand, while it may be that no one is especially concerned with giving an account of how, say, the concept FISH TACO in particular is acquired, many theorists are interested in the general nature of the unlearned basis for acquiring the full range of concepts that humans possess or even more broadly the full range of psychological traits that human beings are capable of acquiring. Moreover, though no one has worked out views about the origins of every general kind of psychological trait, it's not at all uncommon for theorists to have relatively worked out views about a sufficiently large number of different types of traits to make it clear that their position in this global debate is one that is either rationalist or empiricist.

domains (e.g., pertaining to number, geometry, and agents) needn't also adopt a rationalist account of the origins of concepts in many other conceptual domains (such as animals, plants, tools, furniture, vehicles, games, occupations, etc.) in order to qualify as a rationalist in the global debate regarding the origins of concepts. A rationalist can adopt an empiricist account of the origins of *a great many* psychological traits and still count as a rationalist in a global debate.

It is also possible for an empiricist in a global debate to adopt a rationalist account of the origins of some psychological traits. However, there is a clear asymmetry between rationalists and empiricists that stems from the fact that the core feature of empiricist views is their frugality when it comes to the contents of the acquisition base. For this reason, it is effectively not possible for an empiricist in a global debate to adopt a rationalist account of the origins of a great many psychological traits. Doing so would just commit them to too rich an acquisition base to count as empiricist.[14]

The patterns for local debates are similar, but more pronounced. Consider, for example, the local debate concerning the origins of emotion concepts, like the concepts ANGER, JOY, FEAR, LOVE, and PRIDE. An account that adopted a rationalist treatment of even a very small number of such concepts would count as a rationalist account in this local debate. But an account that adopted an empiricist treatment of a very small number of such concepts—or even quite a large number of such concepts—wouldn't thereby count as an empiricist account in the local debate, simply because adopting a rationalist account for the remaining concepts would make the account a rationalist account overall. Again, this asymmetry stems from the differing approaches to the acquisition base taken by rationalists and empiricists. Rationalists see the acquisition base as rich and varied. Adding domain-general learning mechanisms to an acquisition base that contains many specialized components doesn't make the overall account less rationalist. But adding specialized learning mechanisms to an acquisition base that just contains domain-general components *does* make the overall account less empiricist. In fact, adopting a rationalist approach to even a small subset of the psychological traits that are at issue in a local debate means adopting a rationalist account in that debate (though, as noted earlier, this would be compatible with being an empiricist in a more global debate).

There is more to say about these patterns, and we will be returning to this issue later in the chapter. But what we have said so far will suffice for the moment. The notions of global and local debates are largely heuristic notions and not meant to bear serious theoretical weight (e.g., there is little point in trying say when a

[14] These points will become clearer once we outline the general account of when a view should be taken to be rationalist or empiricist which we are working toward. This general account will also address the question of what makes a view rationalist or empiricist to the extent that it is.

global debate becomes sufficiently local to qualify as a local debate rather than a global debate, or vice versa). Nonetheless, these notions are useful for getting oriented to general differences between rationalist and empiricist accounts in different types of debates.

There is one final pair of technical terms pertaining to the acquisition base that we need to introduce in this section. As we have seen, rationalists and empiricists differ in terms of the kinds of psychological traits or structures that they typically posit as being constituents of the acquisition base. While there is unquestionably overlap between the kinds of psychological structures that they may take to be part of the acquisition base, there are also types of psychological structures that are particularly characteristic of each of these approaches. We will use the term *characteristically empiricist psychological structures (or traits)* to refer to structures in the acquisition base which figure prominently in and are characteristic of empiricist accounts of the acquisition base. Similarly, we will use the term *characteristically rationalist psychological structures (or traits)* to refer to structures in the acquisition base which figure prominently in and are characteristic of rationalist accounts of the acquisition base.

What kinds of psychological structures are characteristically empiricist? Domain-general learning mechanisms are the core of empiricist accounts of the acquisition base. They are paradigmatic characteristically empiricist psychological structures, as are low-level sensorimotor representations. But there are other types of characteristically empiricist psychological structures as well. Another type that plays a major role in contemporary empiricist thinking involves psychological structures that bias attention towards certain types of sensory or perceptual properties. These low-level biases are often taken by empiricists to explain how a domain-general learning mechanism could come to process information in a particular content domain without having to posit domain-specific learning mechanisms as part of the acquisition base. For example, empiricists generally suppose that face recognition is grounded in a domain-general processing mechanism that is also involved in recognizing many other types of objects (bodies, houses, chairs, etc.). But on some empiricist accounts, a low-level perceptual bias in the acquisition base is taken to increase the tendency for this mechanism to process face-like stimuli by directing it to attend to a low-level perceptual property that is loosely correlated with face-like stimuli (e.g., curvilinearity). According to this type of proposal, although the processing mechanism is domain general, it develops the ability to process face-specific information (in addition to developing abilities to process information specific to many other domains) because this low-level attentional bias in the acquisition base ensures that it receives an ample supply of facial information.

What about characteristically rationalist psychological structures? Because rationalist acquisition bases are richer than empiricist ones, there are more types of characteristically rationalist psychological structures than there are types of

characteristically empiricist psychological structures. Paradigmatic characteristically rationalist psychological structures include domain-specific learning mechanisms, innate concepts and other abstract representations, innate knowledge structures, and a variety of other types of specialized psychological structures. Moreover, these categories themselves are much broader than the corresponding empiricist categories. For example, there is more variety among the sorts of innate domain-specific learning mechanisms envisioned by rationalists than among the innate domain-general learning mechanisms envisioned by empiricists. For empiricists, different proposed innate general-purpose learning mechanisms tend to cluster around mechanisms that engage in some form of statistical analysis, whereas for rationalists, different proposed innate special-purpose learning mechanisms can perform very different types of computations and can incorporate many types of data structures, as well as much domain-specific representational content in different types of representational formats.

Box 3

Characteristically Empiricist Psychological Structures—characteristically empiricist psychological structures are psychological structures that figure prominently in, and that are characteristic of, empiricist accounts of the acquisition base. These include:

- sensorimotor representations
- domain-general mechanisms
- low-level attentional biases

Characteristically Rationalist Psychological Structures—characteristically rationalist psychological structures are psychological structures that figure prominently in, and that are characteristic of, rationalist accounts of the acquisition base. These include:

- abstract representations and concepts
- domain-specific mechanisms
- knowledge structures

Though we may occasionally drop the "characteristically" from "characteristically empiricist psychological structures" or "characteristically rationalist psychological structures" for stylistic reasons, by and large we will use the full terms to emphasize that while these psychological structures are characteristic of empiricism or rationalism, they are not the exclusive property of empiricists on the one hand and rationalists on the other. As we noted earlier, empiricists can in

principle include concepts or even domain-specific learning mechanisms in their acquisition bases, and rationalists can and typically do include sensorimotor representations and domain-general learning mechanisms in their acquisition bases. The terms *characteristically empiricist psychological structure* and *characteristically rationalist psychological structure* highlight the fact that these structures are ones that figure prominently in and are characteristic of empiricist and rationalist approaches, respectively, while at the same time serving as a reminder that they are *only* characteristic of these approaches, not exclusive to them.

2.3 Learning Mechanisms and Their Local Acquisition Bases

In the previous section, we introduced the idea of an acquisition base, the distinction between local and global debates, and the idea of characteristically empiricist psychological structures and characteristically rationalist psychological structures. In this section, we turn to the next stage in developing our account of how the rationalism-empiricism debate should be understood, which is centred around the nature of rationalist and empiricist learning mechanisms and how they relate to the acquisition base.

Earlier, we emphasized that rationalists and empiricists don't disagree about the importance of learning. Instead, their disagreement is largely about *how* psychological traits are learned and consequently about the nature of the learning mechanisms that account for these traits. To a first approximation, learning mechanisms can be thought of as structures that implement psychological processes for acquiring new psychological traits. In one way or another, learning mechanisms bridge the gap between the psychological structures that are in the acquisition base and the psychological traits that are ultimately acquired on the basis of these structures. Since this is what the rationalism-empiricism debate is essentially about—how to explain the origins of psychological traits with reference to the contents of the acquisition base—learning mechanisms are absolutely central to this debate.

We will use the term *learning mechanism* in a deliberately broad way to cover essentially *any* collection of psychological structures that work together to acquire new psychological traits.[15] This means that any psychological trait that is not itself part of the acquisition base must be acquired via some learning mechanism or other. This applies equally to all such psychological traits, whether the trait is an individual representation, a body of knowledge, a psychological processing mechanism, an entire psychological faculty, an acquired link between psychological structures, or any other type of psychological trait. If it isn't part of the

[15] For stylistic reasons, we will occasionally use the term "system" instead of "mechanism", particularly when we are discussing the views of other authors who use the term "system" for the learning mechanism they posit.

acquisition base, then its acquisition will involve some type of psychological process. And whatever this psychological process is, it must be mediated by a learning mechanism of one sort or another, as we are using this term.

Learning mechanisms will vary enormously in terms of how simple or complex they are, the specific psychological structures they draw upon, the types of psychological processes they support, and how these processes unfold in the acquisition of new psychological traits. The advantage of using "learning mechanism" in this broad and inclusive way is that it gives us a single general term to refer to the psychological structures that are used to acquire any psychological trait at all that is acquired via a psychological process. While using the term *learning mechanism* in this way is not ideal, alternative terms like *psychological-level acquisition mechanism* or *mechanism by which psychological traits are acquired via psychological-level processes* would be needlessly cumbersome, as would *acquired via a learning mechanism or some other psychological-level acquisition mechanism*. In principle, we could use the term *acquisition mechanism*, but this term would be misleading, since it is usually understood to apply equally to traits that are acquired via psychological and non-psychological processes and so would obscure the essential fact that the mechanisms at issue are ones that involve *psychological* processes. Our use of the term "learning mechanism" is meant to be a compromise that avoids these consequences.

We do recognize that this way of way of talking about learning mechanisms may not conform perfectly to common-sense intuitions in some instances. Suppose, for example, that a composer imaginatively arrives at a musical motif, combining chords and rhythms in her head. A case like this arguably doesn't involve *learning*, even though the musical motif is a product of various psychological processes. Nonetheless, from the point of view of the rationalism-empiricism debate, such a case is relevantly similar to other cases that clearly do involve learning. Just as in clear cases involving learning, the representation of the musical motif in this sort of case is acquired via psychological processes which involve or depend upon a set of psychological structures in the acquisition base that ultimately explain its acquisition. And, just as in clear cases involving learning, one can ask whether these structures in the acquisition base tell in favour of rationalism or empiricism. In our view, the practical need for a general term for referring to the psychological mechanisms that bridge the gap between the acquisition base and psychological traits acquired via psychological processes outweighs the mild counterintuitiveness involved in treating "learning mechanism" as a technical term and extending its use to cover these sorts of cases as well.[16]

[16] Later, there will occasionally be times where we do need to distinguish learning from other types of psychological processes, especially when we come to Fodor's claim that learning is impossible (see Part IV). However, the context will make it clear that we are temporarily suspending our more inclusive usage to address concerns that are particularly directed at the ordinary notion of learning.

Let's turn now to consider the relations that learning mechanisms can have to the acquisition base. For all theorists, some learning mechanisms will be contained in the acquisition base. As we have seen, a typical empiricist view holds that the acquisition base contains at least one domain-general learning mechanism. Such a mechanism might initially take sensory or sensorimotor input and, on this basis, produce new psychological traits. Consider, for example, Elizabeth Ray and Cecilia Heyes's model of the origins of imitation (Ray and Heyes 2011).[17] Ray and Heyes reject the idea that an innate special-purpose learning mechanism is needed to explain imitation in infants (as when an infant sticks out their tongue in response to seeing an adult perform this action). Instead, they propose their *associative sequence learning (ASL) model*, which is grounded in a domain-general associative learning mechanism. They argue that given the right type of structured input, this mechanism can incrementally build up the needed associations between components of a perceived behaviour and the motor movements involved in producing this same behaviour:

> [T]he ASL model suggests that the correspondence problem [i.e., the problem of producing the behaviour that matches the behaviour of the person being imitated] is solved piecemeal and by a simple mechanism—associative learning. The success of this simple mechanism depends, not on powerful internal and specialized internal resources, but on the developing infant's environment, especially their sociocultural environment. (p. 97)

As we read Ray and Heyes, this simple mechanism for forming sensorimotor associations is psychologically primitive—it isn't acquired by some other psychological mechanism. In our terms, that means that for Ray and Heyes, it is part of the acquisition base.

Rationalists also posit learning mechanisms that they take to be part of the acquisition base. An example of such a learning mechanism is the mechanism for reorienting in an environment after becoming disoriented that was discussed in Chapter 1. As we noted there, the proposed mechanism in this case works by encoding certain geometrical properties of the environment. Unlike the mechanism posited by Ray and Heyes, this proposed mechanism is a domain-specific learning mechanism that only learns about one kind of thing. But like Ray and Heyes's ASL mechanism, this learning mechanism is also generally seen as not being acquired by some other psychological mechanism, making it part of the acquisition base.

[17] We will describe the sample learning mechanisms that we present in this section in very general terms since we are only interested in highlighting a few of their attributes which help to clarify the notion of a learning mechanism or are otherwise important for clarifying how the rationalism-empiricism debate should be understood. These mechanisms are only being used for illustrative purposes—we are not taking any stand on the existence of any of these mechanisms.

While the learning mechanisms in the two examples we have just given are taken to be part of the acquisition base, learning mechanisms can also be composed in part or entirely from psychological structures that are the products of other learning mechanisms. And they can also involve more complex structures and arrangements than in the examples we have just mentioned. Multiple learning mechanisms can work together in acquiring a new trait. The learning mechanisms in such complexes, along with other psychological traits they interact with, may be part of the acquisition base, may fall outside of the acquisition base (being products of psychological structures in the acquisition base, or products of such products, etc.), or may involve a of mix of these possibilities, with some components that are part of the acquisition base and some that lie outside of the acquisition base. Whenever a psychological process involving a collection of psychological structures of any of these kinds is responsible for acquiring new psychological traits, we will take the whole collection to be a learning mechanism as well.

For what follows, it will be useful to consider a few examples of learning mechanisms that involve a mix of psychological structures from inside and outside the acquisition base. Our first example again relates to the domain of geometry. The learning mechanism is from Elizabeth Spelke, Sang Ah Lee, and Véronique Izard's rationalist account of the origins of Euclidean geometrical concepts (Spelke et al. 2010).[18] First we need a bit of background. Spelke et al. note that in Euclidean plane geometry, any two forms are identical (or congruent) when they can be shown to coincide under rigid transformation. In essence, this means they are identical if one can be perfectly superimposed on the other through some combination of rotation (spinning it clockwise or counterclockwise in the plane) and translation (moving it across the plane). When two forms aren't found to coincide under these types of transformations, they are not identical and may differ in a number of ways. They may differ regarding what Spelke et al. refer to as *distance*, understood to be the length of a part of a form or the length of the space between two parts. They may differ regarding their *sense*, which refers to the left-right organization of their parts. For example, the letters "d" and "b" differ in sense, with one being the mirror image of the other. (Notice that a "d" cannot be superimposed on a "b" by any combination of rotation and translation within a plane; by contrast, a "d" can be superimposed on a "p" in this way). They may also differ regarding *angular* relationships, for example, one may have a more acute angle than the other. For this reason, Spelke et al. take the ability to represent Euclidean geometrical concepts as such to require a sensitivity to all three of these

[18] Although this example concerns the origins of a certain type of concept, the points we are making about learning mechanisms in this section apply to mechanisms for explaining the origins of any type of learned psychological trait (including such things as learned bodies of knowledge, learned processing mechanisms, and learned skills, for example).

properties (distance, sense, and angular relationships). On their view, the acquisition base does not contain any single psychological mechanism that is sensitive to all three of these properties. Instead, they hold that the acquisition base contains two distinct psychological mechanisms, each of which makes use of a subset of these properties in certain contexts, and that what happens is that these mechanisms are combined and augmented into what amounts to a further domain-specific learning mechanism that is capable of acquiring full-fledged Euclidean geometrical concepts.[19] To get a feel for how this is supposed to work, we'll first need a brief look at these two mechanisms and then we can turn to the account of how they are combined.

The first mechanism is the proposed reorientation mechanism that we discussed in Chapter 1 and mentioned just a moment ago. Spelke et al. point out that this mechanism only works at the level of the large-scale navigable environment, not at a smaller scale and not for two-dimensional geometrical forms. Moreover, although it is sensitive to differences in distance and sense in that it can represent a target as being, as Spelke et al. might say, *to the left of a long wall* or *to the right of a more distant wall*, it doesn't explicitly represent angular relationships. In support of this last claim, they cite a study that employed an interesting variation on the reorientation experiment with young children (Hupbach and Nadel 2005). Instead of using a rectangular shaped arrangement of walls, they disoriented children in a space with walls of equal length that formed a rhombus with two acute angles and two obtuse angles. They found that when an object was hidden at one of the corners, children searched at all four corners equally—they didn't seem to be able to use the information that the corner where the target object was hidden was obtuse (or acute) to narrow down their search to the two geometrically equivalent correct corners.[20] To Spelke et al., this suggests that the reorientation mechanism doesn't allow children to reorient themselves by using angular information.

The second of Spelke et al.'s two mechanisms has a different profile. This mechanism, which underlies the ability to represent the shapes of smaller, manipulable objects (like the difference between a box and a bowl) is equipped to distinguish between different angular relationships and does function for two-dimensional forms. However, this mechanism doesn't represent the shape of large-scale navigable areas and doesn't reliably distinguish between shapes that

[19] We should note that Spelke et al. do not employ the notion of an acquisition base, and so do not put their view in terms of what is in the acquisition base (they talk in terms of what is innate). But it is clear that they would accept that the psychological structures that they take to be innate are part of the acquisition base as they do not take them to be the product of a prior learning process. For convenience, we will often translate other authors' views into our framework and terminology in this way when doing so does no disservice to the views at issue.

[20] For other experimental work that bears on whether the reorientation system is limited vis-à-vis angular relationships, see Lee et al. (2012) and the discussion of the representation of places in Spelke (2022).

differ only with respect to sense (i.e., shapes that are mirror images of one another) (see Izard and Spelke 2009).

While neither of these two mechanisms on its own could acquire Euclidean geometrical concepts according to Spelke et al., combining the two mechanisms into a larger learning mechanism allows these concepts to be learned. Spelke et al.'s proposal is that Euclidean geometrical concepts are acquired as children simultaneously exercise both of these mechanisms when learning how to interpret and use cultural products, such as pictures, maps, and scale models, that exploit correspondences between arrangements of small-scale forms and large-scale features of the environment:

> through their experience with pictures, scale models, and maps, children may begin to view large-scale layouts not only as navigable surroundings but also as visual displays with forms that have distinctive angular relationships…through their experience with physical and mental rotation, children and adults may become able to treat small-scale objects and forms not only as visual displays with distinctive shapes but as layouts that can be exploited from different perspectives, by means of navigation systems that allow for stable representations of the distinction between leftward and rightward directions…By extending each of these kinds of geometrical analysis to new types of arrays, moreover, children may develop geometrical concepts that are more abstract and general than the concepts provided by their core systems. (Spelke et al. 2010, pp. 878–879)

Notice that the sort of learning mechanism for acquiring Euclidean geometrical concepts that Spelke et al. are pointing to here isn't a simple, prefabricated learning mechanism that is a part of the acquisition base. It is a more complex psychological mechanism that combines elements from both inside and outside of the acquisition base. This mechanism includes components that are themselves learning mechanisms in their own right, including the two domain-specific mechanisms that are from the acquisition base (the reorientation system and the system for representing the shapes of manipulable objects), along with further components, including some that are very likely domain-general learning mechanisms, which are involved in the interpretations of pictures, scale models, and maps.

Spelke et al.'s account involves two distinct systems in the acquisition base which themselves are concerned with broadly geometrical phenomena and which feed into a learning mechanism for acquiring concepts concerning a distinct geometrical domain, namely that of Euclidean geometry. But learning mechanisms that involve a mix of psychological structures drawn from both inside and outside the acquisition base needn't take this form. Another type of learning mechanism employing a mix of learned and innate structures builds on just a single system in the acquisition base that is related to the acquisition target, where the mixed

learning mechanism it feeds into results in a new capacity that transcends the limits of the innate system.

Stanislas Dehaene's rationalist account of the origin of natural number concepts (ONE, TWO, THREE, FOUR, FIVE, and so forth) is an example of this type (Dehaene 1997). His account is centred around a posited innate capacity to represent approximate numerical quantities, which has come to be known as the *approximate number system*.[21] Surprising as this may be, there is substantial evidence that the approximate number system is present in both infants and many kinds of animals.[22] This system represents numerical quantities specifically and not merely more concrete properties that tend to correlate with numerical quantity (such as the amount of area taken up by a collection of objects or the duration of a sequence of sounds or actions). For example, rats can be trained to tap a lever a number of times to receive a reward (say, eight times) and will tap this same number faster when they are hungrier, showing that they aren't just responding in terms of being rewarded for tapping for a certain amount of time—it's the (approximate) *number* of taps that controls this behaviour (Mechner and Guevrekian 1962). The approximate number system has several interesting features. One is that it is subject to the *distance effect* in that its ability to distinguish between two numerical quantities declines as the numerical distance between them decreases; for example, it's better at discriminating 3 from 10 than 3 from 4. Another is that it is subject to the *magnitude effect* in that its ability to distinguish between two numerical quantities that differ by the same amount declines as the numerical quantities become larger; for example, it's better at discriminating 6 from 4 (a difference of 2) than 16 from 14 (also a difference of 2).[23] These effects show that the posited approximate number system lacks the precision that is inherent to natural number concepts and instead involves a ratio-dependent approximate number representation. How then are precise natural number concepts learned? On Dehaene's account, what happens is that the approximate number system feeds into a more complex learning mechanism that recruits linguistic and other symbolic abilities and culturally inherited numerical technologies (such as tallies, number words, and counting procedures). Over time, these impose greater precision on the approximate number system's numerical content, allowing learners to form new numerical representations—precise numerical

[21] Dehaene called this system *the accumulator*, but we will use the now standard term *approximate number system*.

[22] In fact, there is reason to suppose animals can have an approximate system whose acuity is essentially on a par with—and perhaps sometimes even more discriminating than—the approximate number system in humans (Cantlon and Brannon 2006; Rilling and McDiarmid 1965).

[23] There is evidence the approximate number system is active not only when adults make rough judgements regarding the numerical size of a perceived collection or compare two perceived collections for which is numerically larger, but even when working with symbols for precise numerical quantities, such as Arabic numerals, and when performing precise calculations (Dehaene et al. 1990). Moreover, ability in formal mathematics seems to correlate with individual differences in the acuity of people's approximate number system (Halberda et al. 2008).

representations—which they initially associate with numerical symbols of one kind or another.

While Spelke et al.'s account of the origins of Euclidean geometrical concepts involves a learning mechanism that draws on two systems in the acquisition base each of which itself is concerned with a form of geometrical content, Dehaene's account of the origins of natural number concepts involves a learning mechanism that draws on just one system in the acquisition base which itself is concerned with a form of numerical content. In both cases, these critical domain-specific components from the acquisition base are only part of a more complex and encompassing learning mechanism, which crucially also draws upon psychological structures that aren't part of the acquisition base but instead are themselves the products of other learning mechanisms.

The final learning mechanism that we will mention here is Susan Carey's account of the origins of natural number concepts (Carey 2009). Like Dehaene and Spelke et al.'s accounts, Carey's proposed learning mechanism also makes use of a mix of psychological structures (some from inside the acquisition base, some from outside the acquisition base), and it is also a rationalist account. Importantly, though, it's one that doesn't rely on even a single mechanism from the acquisition base which is itself concerned with content that is closely related to that of the concepts acquired by the learning mechanism.

In broad outline, Carey's account is that young children start by developing abstract representations of small groups of individuals that are not specifically numerical representations and that these representations gradually acquire a numerical interpretation. The initial abstract representations are based on specialized mechanisms in the acquisition base for representing and tracking individuals, and mechanisms for creating and manipulating mental models. Among other things, the transition to numerical content is mediated by an innate understanding of quantification (understanding of contents like *some*, *all*, and *most*) and the ability to put collections into one-to-one correspondence with one another. Once the first few number concepts (ONE, TWO, THREE) are developed in this way, another process—which Carey refers to as *bootstrapping*—enables children to extrapolate beyond these first few number concepts to the full set of natural number concepts. The bootstrapping process in this case relies on further elements in the acquisition base, including an innate capacity for working with ordered lists, an innate capacity for drawing inductive inferences, an innate capacity for analogical reasoning, and innate linguistic and symbolic capacities. This account, while still clearly rationalist, posits rather different elements in the acquisition base than Dehaene's account, and notably, unlike Dehaene's account, does not draw on any resources in the acquisition base that are specifically numerical.

The examples that we have given of learning mechanisms that involve a mix of psychological structures drawn from both inside and outside the acquisition base

have been of rationalist learning mechanisms rather than empiricist ones. This was in part because we wanted to illustrate some of the variety that learning mechanisms can have, and in part because there is a greater diversity of rationalist learning mechanisms. But empiricist learning mechanisms can also have this mixed character, drawing on psychological structures from both inside and outside the acquisition base. In any case, as the sample learning mechanisms we have given illustrate, learning mechanisms come in a great many different shapes and sizes. They can be simple or complex. They can be part of the acquisition base, products of other learning mechanisms, or composed of a mix of psychological structures from inside and outside of the acquisition base. They can involve psychological structures which concern related contents or not, and when they do make use of structures with related contents, they can build on these in many different ways. Having seen some of the ways that learning mechanisms can vary, and having seen examples of learning mechanisms that have been proposed by both rationalists and empiricists, we are now ready to turn to the general question of how the types of learning mechanisms that play a central role in rationalist and empiricist accounts differ from one another.

Our discussion will build on the distinction between characteristically empiricist psychological structures and characteristically rationalist psychological structures, which was introduced in section 2.2. Recall that these terms refer specifically to psychological structures that are in the acquisition base, with the first picking out psychological structures which figure prominently in and are characteristic of empiricist theories, and the second picking out psychological structures which figure prominently in and are characteristic of rationalist theories. In much the same way, we can refer to the types of learning mechanisms that are representative of the empiricist approach as *characteristically empiricist learning mechanisms*, and the types that are representative of the rationalist approach as *characteristically rationalist learning mechanisms*. As a preliminary statement of what these come to, we can say that characteristically rationalist learning mechanisms are learning mechanisms that involve characteristically rationalist psychological structures, while characteristically empiricist learning mechanisms are learning mechanisms that don't and that instead only make use of characteristically empiricist psychological structures.[24] (We will provide a more precise formulation of what each of these kinds of learning mechanisms involves shortly).

As with characteristically rationalist and empiricist psychological structures, rationalists and empiricists can both make use of both characteristically rationalist and characteristically empiricist learning mechanisms, at least in some

[24] For ease of expression, we will often abbreviate these, and refer simply to *rationalist learning mechanisms* and *empiricist learning mechanisms*, but it should be kept in mind (as we explain below) that these aren't exclusive to the rationalist approach in the first case or to the empiricist approach in the second.

circumstances. Regarding any global rationalism-empiricism debate, it's possible for empiricists to hold that, in addition to domain-general learning mechanisms, some small number of characteristically rationalist learning mechanisms are also involved in learning, though empiricists will generally try to avoid postulating such characteristically rationalist learning mechanisms, and postulating more than a relatively small number of substantially different rationalist learning mechanisms simply makes a theory rationalist. In contrast, rationalists in a global rationalism-empiricism debate are free to hold that any number of characteristically empiricist learning mechanisms are involved in learning as long as they also posit a number of characteristically rationalist learning mechanisms. With respect to local debates, the options are even sharper. Empiricists are more or less unable to posit any rationalist learning mechanisms, as embracing a rationalist learning mechanism in a local rationalism empiricism debate typically makes a theory rationalist in that local debate. By contrast, rationalists are again free to accept the involvement of any number of empiricist learning mechanisms provided that they also take at least one rationalist learning mechanism to be involved.[25] These facts underscore an important asymmetry regarding the rationalism-empiricism debate, which is an outgrowth of the asymmetry we noted in section 2.2: Although rationalists and empiricists can both posit characteristically rationalist and characteristically empiricist learning mechanisms in explaining the origins of psychological traits, empiricists can posit characteristically rationalist learning mechanisms to only a highly limited extent without becoming rationalists, while there is no limit on the number of characteristically empiricist learning mechanisms that rationalists can posit without becoming empiricists.

As we noted a moment ago, there is also a sense in which there is a considerably greater variety of characteristically rationalist learning mechanisms than there is of characteristically empiricist learning mechanisms. Much of this greater variety stems from the fact that there is a greater variety of types of characteristically rationalist psychological structures than there is of characteristically empiricist psychological structures. This means that for rationalists there are generally many more types of structures in the acquisition base that can feed into different kinds of learning mechanisms than there are for empiricists. It also means that there is a sense in which there is also greater variety among competing rationalist accounts, since there is a wider variety of competing rationalist mechanisms that might be offered for acquiring the same trait. (We have seen some indication of this in the different accounts offered by Dehaene and Carey of the origins of natural number concepts.) This isn't to say that there aren't differences among competing empiricist accounts too. But given that empiricists generally rely primarily on domain-general learning mechanisms and low-level attentional biases,

[25] At the same time, as we explain in section 2.5, not all rationalist or empiricist accounts are equally rationalist or empiricist; one account can be rationalist (or empiricist) to a greater extent than another.

empiricist learning mechanisms, when viewed from a distance, often have much the same general shape, fundamentally turning on some form of statistical analysis or a data-driven learning process.

There is an important qualification that needs to be made here, however. While empiricists posit sparse acquisition bases, this does not necessarily mean that they are precluded from accepting a learning mechanism that makes use of some of the same kinds of resources which, on accounts like Dehaene's or Carey's, are taken to be characteristically rationalist psychological structures that are part of the acquisition base. It is just that they will typically need to suppose that, rather than being part of the acquisition base, these resources are themselves *learned* (ultimately on the basis of domain-general learning mechanisms in the acquisition base). For example, it's possible to imagine a theory that is very similar to Dehaene's account of the origins of natural number concepts but that supposes that, rather than itself being part of the acquisition base, the approximate number system is learned via an innate domain-general learning mechanism. On this type of account, once the approximate number system is learned, it could form part of a learning mechanism in more or less the same way as on Dehaene's account.[26] While not common, this type of empiricist learning mechanism is certainly possible. To address this type of case, we need to have a way of differentiating rationalist and empiricist learning mechanisms that employ much the same resources but that do not take these resources to have the same status in terms of whether they are part of the acquisition base or acquired through prior learning.

To differentiate between such accounts, we can highlight the fact that while rationalist and empiricist accounts of this type end up converging on more or less the same proximate learning mechanism for a trait, they nonetheless take this proximate learning mechanism to be based on different underlying structures in the acquisition base. It will help to introduce another piece of terminology to succinctly capture this type of difference. We will say that such accounts postulate different *local acquisition bases*, where a local acquisition base is the subset of the acquisition base that contributes to a learning mechanism in a local rationalism-empiricism debate. The local acquisition base for Dehaene's rationalist account of the origins of natural number concepts will then include the approximate number system. But the local acquisition base for the converging proximate learning mechanism on the empiricist account won't, because on this account the approximate number system is learned. Since it is learned and isn't part of the acquisition base on this account, it won't be part of the local acquisition base either.[27]

[26] For a proposal along these general lines, see Leibovich et al. (2017). On their model, the approximate number system isn't innate; it's acquired on the basis of a more general capacity for representing magnitude. But once it is acquired, their account of how concepts for natural numbers are learned is similar to Dehaene's.

[27] Local acquisition bases also give us a useful way of capturing differences among rationalist accounts. For example, on Dehaene's rationalist account, the most important component of the local acquisition base is the approximate number system, though other components will also contribute to

We have described the local acquisition base regarding a learned psychological trait as the subset of the acquisition base that contributes to the learning mechanism that is responsible for the acquisition of that trait. However, it turns out that what it means for something in the local acquisition base to *contribute* to a learning mechanism is more complicated than it might first appear. To see this, we need to recognize the importance of the entire learning mechanism involved in acquiring a psychological trait, not just the proximate learning mechanism. This comprises the full chain of psychological activity that is responsible for the acquisition of the trait. In different cases and for different learning mechanisms, this chain of psychological activity could have different characteristics. It could unfold in predictable stages or in a more haphazard way; it could involve a relatively isolated psychological change that comes about quickly or more extensive changes that are spread out across a long time span, years even.

Our question is, How can a psychological trait in the local acquisition base contribute to the proximate learning mechanism for a given trait? We can now see that, generally speaking, there are two different ways. One of these—the more direct way—is where the psychological structure in the local acquisition base itself plays an active role in the operations of the proximate learning mechanism. In this case, it may be that it is essentially identical to the proximate learning mechanism or alternatively that it is a component of the proximate learning mechanism. The other way in which a psychological trait in the local acquisition base my contribute to a proximate learning mechanism—the indirect way—is where the psychological structure in the local acquisition base is neither identical with or nor a component of the proximate learning mechanism but is instead part of the learning mechanism as a whole that ultimately produces the learned trait. In other words, the psychological trait in the local acquisition base is part of a chain of psychological activity in which it is the *products* of this activity (as opposed to the structure in the acquisition base itself) that are directly involved in the operation of the proximate learning mechanism. Consider again the empiricist account of the origins of natural number concepts that uses essentially the same proximate learning mechanism as that used in Dehaene's rationalist account. On this empiricist account, the proximate mechanism involved is not itself in the acquisition base. But what *is* in the acquisition base—domain general learning mechanisms—nonetheless contribute to the acquisition of the trait. It is just that they do so indirectly by producing this proximate learning mechanism.

It will be useful to have a single term that captures both of these ways in which a psychological structure in the local acquisition base may contribute to a

the acquisition of these concepts. By contrast, the approximate number system is not part of the local acquisition base for acquiring natural number concepts on Carey's account. Instead, the local acquisition base on her account contains, among other things, specialized mechanisms for representing and tracking individuals and for creating and manipulating mental models, a system for understanding quantification, and the ability to put collections into one-to-one correspondence.

learning mechanism—directly or indirectly. We will use the term *traces back to* for these purposes. So we will say that a given learning mechanism traces back to the psychological structures in its local acquisition base that contribute to this learning mechanism, either directly or indirectly. Likewise, we will also say that the traits acquired via this learning mechanism trace back to the psychological structures in the learning mechanism's local acquisition base.

We are now in a position to go beyond the preliminary characterization that we gave earlier for what makes a learning mechanism characteristically rationalist or characteristically empiricist. Earlier, we put this by saying that characteristically rationalist learning mechanisms in some sense involve characteristically rationalist psychological structures and that characteristically empiricist learning mechanisms only make use of characteristically empiricist psychological structures. We can now say that a *characteristically rationalist learning* mechanism is any learning mechanism which traces back to characteristically rationalist psychological structures in a local acquisition base, and that a *characteristically empiricist learning mechanism* is any learning mechanism which traces back to *only* characteristically empiricist psychological structures in the acquisition base.

To recap, characteristically rationalist psychological structures and characteristically empiricist psychological structures are ones that are in the acquisition base and that are the types of structures that are especially representative of the rationalist framework in the first case and the empiricist framework in the second. Learning mechanisms are collections of psychological structures that mediate the acquisition of new psychological traits. Although some learning mechanisms may be components of the acquisition base, many won't be or will have parts that

Box 4

Local Acquisition Base—a local acquisition base is a subset of the acquisition base that contributes to a learning mechanism in a local rationalism-empiricism debate (a debate that focuses on how a particular psychological trait or a narrowly circumscribed set of traits is acquired).

Characteristically Empiricist Learning Mechanism—a characteristically empiricist learning mechanism is a learning mechanism that traces back to a local acquisition base that is exclusively comprised of characteristically empiricist psychological structures.

Characteristically Rationalist Learning Mechanism—a characteristically rationalist learning mechanism is a learning mechanism that traces back to a local acquisition base that includes characteristically rationalist psychological structures.

aren't. Nonetheless, for any learning mechanism, we can ask about its psychological origins and what this tells us about its local acquisition base—the subset of the acquisition base that it traces back to. When a learning mechanism traces back to a local acquisition base that includes characteristically rationalist structures, it counts as a rationalist learning mechanism. When it traces back exclusively to characteristically empiricist psychological structures, it counts as an empiricist learning mechanism. Finally, in a global debate, empiricists and rationalists can both take on the psychological structures and learning mechanisms that are characteristic of the other's framework, but empiricists are far more limited in the extent to which they can do this owing to their commitment to a frugal acquisition base, whereas rationalists are not constrained in this way.

To close this section, we want to address two related issues regarding how best to understand competing claims about proposed learning mechanisms in any local rationalism-empiricism debate. The first issue concerns the status of theories of learning mechanisms. When a theorist puts forward an account of a learning mechanism for acquiring a psychological trait, this shouldn't be understood as saying that they take their proposed account to be the only possible way the trait could be acquired. Claims about learning mechanisms that are relevant to the rationalism-empiricism debate are not claims about how a trait *must* be acquired. They are claims about how the trait in question is *actually* acquired.

Of course, at this point in the development of the cognitive sciences, these claims (both rationalist and empiricist) should typically be taken as tentative hypotheses involving partial sketches of learning mechanisms intended to highlight certain critical aspects of the learning process. The aim is to make these explicit enough to be evaluated against the known facts about prior states of development, the learning environment, and any relevant findings about how the development of the trait actually takes place. Despite typically being only partial sketches of the origins of traits, both rationalist and empiricist theories positing learning mechanisms should be seen not as providing an account of how a given trait must be learned, but rather how it is in fact learned. This bears emphasizing because occasionally critics of rationalist approaches treat rationalist accounts as though they were claiming that the trait could only possibly be acquired through the rationalist mechanism, and then go on to argue that such an account can be rejected simply by showing that an empiricist alternative is possible in principle. However, it is a mistake to suppose that the mere possibility of an alternative account undermines any given proposal. What is at issue in the debate is what the learning mechanisms that we actually use to acquire traits are like (for more on this, see Chapter 17).

The second point concerns the possibility of there being multiple alternative paths to acquiring a trait. Given that competing proposals in a local rationalism-empiricism debate generally specify a single way that the trait is thought to be acquired, does this mean that their proponents are committed to there being no variation in how it is actually acquired? No, not at all. The point of advancing a given account is to offer what is thought to be an illuminating model for the

typical way that the trait is acquired. But it's to be expected that there will be a certain amount of variation. Such variation can come in different forms.

Some of this variation will involve fine-grained differences in learning mechanisms that will have little or no effect on the hypotheses at issue (e.g., small differences in memory capacity, attention, motivation, specific input, etc.) and so will not affect the overall local debate. In part, this is because learning mechanisms are typically specified at a level of generality that abstracts away from many of the fine details that are likely to differ across individuals.

The amount (and kinds) of variation present in learning mechanisms for a given trait will also be affected by the trait whose acquisition is at issue. Knowledge of a particular strategy in chess might be acquired through reading about the strategy, or alternatively from seeing a game played where the strategy was employed, or by discovering the strategy by thinking through possible moves and countermoves. On the other hand, the kinds of traits that are typically at issue in debates between rationalists and empiricists are not of this type. Instead, they typically involve relatively fundamental types of cognitive traits, such as the ability to speak a language, to recognize faces, to conceptualize oneself and others as possessing minds and mental states, to think in terms of numerical quantities, and so on. These kinds of traits are less likely to be subject to the sorts of variation in the types of learning mechanisms involved in their acquisition than something like a strategy in chess. It is even less likely that variation in learning mechanisms for such traits would affect whether the learning mechanism involved was rationalist as opposed to empiricist, or vice versa. Few if any theorists suppose that traits like the ability to speak a natural language are acquired via rationalist learning mechanisms for some individuals and via empiricist learning mechanisms for others. This is one of the reasons why such traits are of interest in this type of debate—because they are relatively fundamental traits, for which it is a reasonable assumption that the learning mechanisms involved are fairly uniform.

This is not to say that there will be no cases where fundamental traits of these sorts are subject to systematic variation. We know that there will be cases where there is substantial variation that directly affects the learning mechanisms that different individuals possess and make use of. Congenitally blind or deaf individuals, for example, will possess different types of acquisition bases than sighted and hearing individuals and will have substantially different patterns of input to their learning mechanisms due to their blindness or deafness. Such variation will no doubt lead to some variation regarding the traits that are learned and how they are learned. At the same time, it is very much an open empirical question how much of an impact this will have.[28] As we will see later in the book,

[28] For example, a now classic study examining the effect of congenital blindness on language acquisition found that, while it is often thought that congenitally blind children are at a great disadvantage for learning word meanings (given that they often lack perceptual access to the things being referred to in everyday conversations), this is not the case; vocabulary growth in congenitally blind children is on a par with other children's (Landau and Gleitman 1985).

sometimes individuals whose acquisition bases are affected in these and similar ways can end up with learning mechanisms and learned traits that are remarkably similar to those in other individuals (see especially Chapters 13 and 20).

2.4 Domain Specificity and Domain Generality

We have been building our account of what the rationalism-empiricism debate is really about in stages, introducing key distinctions and clarifying terminology as we go. In section 2.1, we argued that it is a mistake to see this debate as merely a confused way of expressing the trivial idea that infants have the capacity to acquire whatever psychological capacities they develop later in life. Rather, the debate should be seen as a substantive dispute which, in the first instance, is about the nature of the innate psychological structures underlying the development of psychological traits. In section 2.2, we refined this approach by introducing the idea of an acquisition base, which refers to the collection of psychologically primitive psychological structures—ones that are not acquired via any psychological process of acquisition. This allowed us to say that rationalist and empiricist accounts differ as to whether the acquisition base is largely restricted to what we are calling characteristically empiricist psychological structures (especially sensorimotor representations and domain-general learning mechanisms) or whether it also includes a significant number of characteristically rationalist psychological structures (especially more abstract representations and domain-specific learning mechanisms). Section 2.3 went on to explain how we will be using the notion of a learning mechanism in our account and how different theories of the way that a trait is learned are typically committed to learning mechanisms that trace back to different local acquisition bases. In this section, we turn to the next stage in our account—the distinction between domain specificity and domain generality.

Up to this point, we have relied on an intuitive understanding of this distinction, noting that domain-specific mechanisms figure prominently in rationalist theories. But we need to say a little more about what the distinction between domain specificity and domain generality comes to. Although these notions are widely relied on in discussions of the rationalism-empiricism debate and in other ongoing debates in philosophy and cognitive science, there is also much controversy about how to understand these notions and whether they hold up to scrutiny. In this section, our goal won't be to defend these notions or to grapple with the various puzzles and problems they have been thought to give rise to. We will simply explain how we understand these notions and will tease them apart from a related pair of notions that needs to be recognized in the rationalism-empiricism debate.[29]

[29] The notions of domain specificity and domain generality are often dismissed on grounds that they are confused notions that can't bear the theoretical weight that has been put on them. We think

To begin, we need to say something about what a domain is. A domain—or as we will often say, a *content domain*—is best thought of as a subject matter (Fodor 1983). It is the subject matter that a psychological structure is directed at. *Being directed at* is a relation much like *represents*, possessing the same idiosyncratic properties that are characteristic of a whole family of related notions, including *being about* and *having intentionality*.[30] One of these properties is *being perspectival*. This means that a content domain shouldn't be understood merely as a collection of entities. Built into the very idea of a content domain is that there is a way that the entities in that content domain are to be construed. As a consequence, two psychological structures could be directed at different content domains even if the two content domains contained exactly the same entities. To use a well-worn example, even if every creature with a heart is also a creature with kidneys, a mechanism that is specialized for representing and reasoning about creatures with hearts would be directed at a different content domain than a mechanism that is specialized for representing and reasoning about creatures with kidneys. Or, to use another example, a mechanism could be specialized for representing three-angled closed polygons as opposed to three-sided closed polygons, even though all triangular polygons are trilateral polygons, and vice versa. The perspectival nature of *being directed at* also means that the same entity can belong to many different content domains. For example, the same entity can be in the domain of physical objects, the domain of animals, the domain of agents, and so on. This entity would be represented in different ways and for different purposes by different domain-specific cognitive mechanisms directed at these different content domains.

A second and related feature of the relation *being directed at* is that subject matters needn't correspond to objective categories discovered by science and may even involve things that don't actually exist. Just as a story could have unicorns as its subject matter or have as its subject matter how things might have gone if the dinosaurs hadn't become extinct, a cognitive mechanism could have as its content domain mythical or fictional creatures or be concerned with counterfactual events. For much the same reason, the content domain which a cognitive mechanism is directed at could diverge from the categories that are recognized and investigated by science even when the domain involves real as opposed to fictional entities. For example, while standard biological taxonomies don't recognize categories like *tree* or *fish* (because these categories don't involve groupings of organisms that include all of the descendants of a common ancestor), a cognitive mechanism might well range over these categories and have them as part of its content domain. This would be true, for example, of a mechanism that is

that these concerns are misplaced. Here we just briefly outline what we take to be the best understanding of these notions without addressing the many puzzles and confusions that have been thought to undermine these notions. We address these concerns in Margolis and Laurence (2023) while also providing a much more detailed discussion of our account and of the general theoretical context.

[30] We will have more to say about intentionality and related notions later, in Chapter 6.

responsive to the biological realm as it is conceptualized in everyday thinking (folk biology) as opposed to how it is conceptualized in scientific biology (Medin and Atran 1999).

As we move forward, it will be useful to have a term for the concepts that are associated with a given content domain through being directed at that domain. We will use the term *conceptual cluster* for this purpose. Suppose, for example, that there is a learning mechanism that is specialized for acquiring animal concepts. Given the right types of experience, it generates specific animal concepts—ZEBRA, BOA CONSTRICTOR, FALCON, and so on. Taken together, all of these concepts would constitute a conceptual cluster which is directed at, or has as its subject matter, the content domain (or conceptual domain) *animals*.[31] Having this terminology in place helps to keep clear whether one is referring to the subject matter (content domain) or to the psychological structures that are directed at the subject matter (in this case, the conceptual cluster that is directed at the content domain). A similar issue regarding potential unclarity also arises for more complex informational states, where psychologists often use the term *body of knowledge*. This term could be taken to refer to the subject matter (what it is knowledge of) or to the psychological states that encode and process information pertaining to this subject matter. To be clear, when we speak of something as a body of knowledge, we are referring to psychological states that are directed at a subject matter, rather than the subject matter itself. For example, a body of knowledge that is specific to physical objects and core physical interactions between such objects is a psychological structure that is directed at the content domain *physical objects*.[32]

To sum up the terminology so far: a content domain is a subject matter. In contrast, conceptual clusters and bodies of knowledge aren't subject matters; they are psychological structures which are directed at particular subject matters.

Box 5

Domain (or Content Domain)—a domain is a subject matter that a psychological structure can be directed at.

Conceptual Cluster—a conceptual cluster is a collection of concepts which are directed at a particular domain.

[31] We will sometimes use the term *conceptual domain* as an alternative term to refer to a content domain that a particular conceptual cluster is directed at.

[32] We will follow the common practice in cognitive science in using the term *body of knowledge* in a way that has no implications regarding the truth of the representations involved or the type of justification or warrant the individual has for them (in contrast to the way that the term *knowledge* is used in most areas of philosophy).

Having clarified what a domain is (that is, a content domain), we can now turn to the question of what makes for domain *specificity* or domain *generality*. To a first approximation, domain specificity is a matter of being directed at a particular domain, whereas domain generality is a matter of being directed at a number of distinct domains. What makes it the case that a domain-specific psychological mechanism is directed at a given content domain? One common answer to this question is that it is a matter of the input to the mechanism. Carruthers (2006) sees domain specificity in these terms, distinguishing a mechanism's input from other information it may access in the course of its operations. Input for Carruthers is understood in terms of what "turns on" the mechanism. For example, supposing a cognitive mechanism were only turned on by linguistic input, then this mechanism would be considered to be a domain-specific mechanism that is directed at the content domain of language.

The idea that input is what makes a domain-specific cognitive mechanism be directed at a given content domain isn't the only possibility, however. Other theorists emphasize the nature of the computations that take place within the mechanism (e.g., Cosmides and Tooby 1994; Gallistel 2003). Consider, for example, a cognitive mechanism that Cosmides and Tooby have proposed which is dedicated to determining whether those receiving benefits in social exchanges are entitled to them (often referred to as a *cheater detection module*). The proposed mechanism is taken to employ distinctive processes that are specially tailored to determining the legitimacy or illegitimacy of received benefits in social exchanges. Yet the input to this mechanism can be highly diverse. The relevant benefits might involve financial gain, admittance to a desirable school system, the right to drive someone's car, or any number of other things, and the legitimacy of such benefits might turn on a huge variety of factors. This way of determining what makes a cognitive mechanism specific to a particular domain allows for domain specificity in cases where the mechanism may have diverse inputs but is nonetheless directed at a particular content domain in virtue of the fact that the processing mechanism is specifically tuned to processing content from the content domain that it is directed at.

There is also a third factor that should be considered regarding what makes a domain-specific cognitive mechanism be directed at a given content domain, one that has been largely neglected in the literature on domain specificity. This has to do with the *output* of the mechanism. Let's consider again a hypothetical domain-specific mechanism that is solely devoted to acquiring concepts of animals. Arguably, a key feature that makes such a mechanism specialized for the content domain of animals is the fact that the output of this mechanism is the conceptual cluster that is directed at the content domain of animals. One of the advantages of using output to determine what makes a cognitive mechanism domain specific is that in many cases we may not have much information about how a mechanism works or exactly what type of input it is restricted to, and it is more straightforward to simply consider its output. For example, if a learning mechanism just

produces representations of faces, then it is specialized for the content domain *faces*, and we do not need to know whether its internal computations are uniquely suited to faces in order to see that the mechanism is domain specific.[33]

In our view, rather than trying to decide which of these three factors—distinctive input, specialized internal processes, or distinctive output—is most important, or trying to distinguish different senses of domain specificity linked to these factors, it is better to understand domain specificity as a function of all three factors. In particular, while we think that any of the three factors suffices for domain specificity, we see domain specificity as involving all three factors.

Take, for example, Chomsky's classic proposal that there is an innate language faculty, an innate domain-specific mechanism for acquiring natural language syntax. By hypothesis, this mechanism produces just one thing—a grammar that specifies the syntactic properties of the local natural language. If a young learner is exposed to more than one natural language, their language faculty may produce further grammars for each of these languages, but it can't do much else. It can't acquire knowledge of the rules of chess; it can't figure out how to navigate through a maze; it can't help you balance your chequebook. It can't even produce a grammar for some other type of system of communication, a "language" whose structural properties substantially deviate from those of human natural languages. This is because it is directed at languages that conform to the principles of Universal Grammar.[34] Clearly, then, this mechanism is quite limited regarding its output. It is also limited regarding its input. The language faculty, on this proposal, is selective regarding the information it is responsive to and uses when forming a grammar. It doesn't respond to sounds in general or even more narrowly to the vocalizations emitted from other individuals. Its input consists of *linguistic* expressions (words, phrases, sentences), which it represents specifically as linguistic data. Finally, the language faculty exploits this incoming information in a distinctive manner. On one such proposal, the language faculty embodies a set of parameters each of which has just a few options regarding some critical syntactic property. For example, a parameter may determine whether a language is *head-initial* or *head-final* (the head of a phrase being the word that establishes the phrase's syntactic category, such as the verb in a verb phrase). In head-initial languages, the head appears before its complements; in head-final languages, it appears after its complements. The point is that a mechanism that incorporates a number of parameters of this kind, which are specific to structural features of

[33] Such a mechanism might employ a form of statistical analysis that could be equally used in mechanisms that are provided with different types of content. Nonetheless, as we see it, if it were part of an overall cognitive architecture in which it was positioned to only receive input involving facial stimuli and, as a consequence of this arrangement, delivered output of just one type of content (facial representations), it would still count as a domain-specific mechanism.

[34] Universal Grammar refers to a set of principles that apply to all human natural languages and which play a critical role in the domain-specific language acquisition device envisioned by Chomskyan accounts of language acquisition. (For an introduction to Universal Grammar, see Cook and Newson 2007.)

natural language, is uniquely suited to acquiring languages that conform to Universal Grammar and hopelessly unsuited to doing anything else. So, this prototypical case of a domain-specific mechanism clearly involves all three factors. The language faculty is specialized for this one domain because of the type of input it relies on, the way it processes this input, and the type of output it can produce.

In our preliminary characterization of domain specificity and domain generality, we made the simplifying assumption that domain-specific mechanisms are directed at only a single domain. In fact, however, it seems reasonable to suppose that both domain specificity and domain generality are graded phenomena that come in degrees. In particular, we will assume that a domain-specific mechanism is one that is directed either at a single domain or just a few domains, especially when these are closely related in content. One cognitive mechanism will be more domain specific than another to the extent that it is directed at a smaller number of closely related content domains (where being directed at only one content domain counts as being maximally domain specific).

In contrast, a domain-general mechanism is one that is directed at more than just a few domains (especially when these are diverse domains that are not closely related in content).[35] A domain-general mechanism of this sort is directed at these various domains not by collapsing them into a broader domain but rather by being directed at them in a differentiated way. It is *multiply directed*—directed at each of the various domains it concerns separately—by being successively directed at each of these domains when it is processing information pertaining to that content domain. For example, a domain-general concept learning mechanism would be capable of acquiring concepts in a variety of content domains not in virtue of properties that all these concepts have in common as members of a single larger content domain, but rather in virtue of properties that they each have that make them members of their respective different content domains. When such a domain-general learning mechanism acquires concepts in the tool domain, it is directed at the content domain *tools*. When it acquires numerical concepts, this very same mechanism is directed at the content domain *number*. And so on for other conceptual clusters and their content domains.[36] One cognitive mechanism will be more domain general than another to the extent that it has a higher degree of multi-directedness: that is, it is directed at a larger number of different domains (especially when they are diverse in content), being

[35] A domain-general mechanism will be directed at multiple domains in virtue of how it relates to the same three factors that determine the domain that a domain-specific mechanism is directed at. In particular, a domain-general mechanism will be directed at multiple domains in virtue of it taking input from these multiple domains, producing outputs in these multiple domains, and having a processing mechanism that is not specialized for processing content from any particular content domain.

[36] In their initial state, domain-general processing mechanisms may not yet treat inputs and outputs from particular domains as belonging to distinct domains but can be seen to be domain general in virtue of taking inputs from a range of domains and having a processor that is not specialized for a particular domain. Relatedly, there is a derivative sense of domain generality associated with a type of processing, as opposed to a processing mechanism, where a type of processing counts as domain general to the extent that it is not specialized for a particular domain.

successively directed at each of these domains as such when it is processing information pertaining to that content domain.

Much more could be said about the notions of domain specificity and domain generality, but what we have said will suffice to clarify our use of these terms. We will end this section by introducing a related distinction that is easy to conflate with the distinction between domain specificity and domain generality. This related distinction concerns a different sense in which a mechanism may be special purpose or general purpose, but unlike the distinction between domain specificity and domain generality, it doesn't concern the range of content domains that the mechanism is directed at. Instead, this distinction has to do with the range of functions a psychological mechanism has, that is, the range of cognitive operations or computations it can perform, such as computing the similarity to a prototype, drawing inductive inferences, or rehearing information in working memory. When a mechanism only has one kind of function or a small range of closely related functions, then we will say it is *functionally specific*, and when it has more than a small range of functions, especially when they are diverse functions, we will say it is *functionally general*.[37]

Mechanisms that are both functionally specific and domain specific have been central to rationalist theorizing and include mechanisms like the reorientation system and the approximate number system. And mechanisms that are both functionally general and domain general have played an important role in accounts in cognitive science from its earliest days (Newell et al. 1958; Newell and Simon 1972) to the present (e.g., LeCun et al. 2015).

Crucially, however, the question of what range of functions a mechanism has is distinct from, and independent of, the question of what range of content domains a mechanism is directed at. This means that a functionally-specific mechanism needn't also be domain specific. A mechanism can be functionally specific and *domain general*. To take a simple example, consider a cognitive mechanism that only performs one type of inference, drawing logical inferences in accordance with modus ponens (inferring *Q*s from premises of the form *if P, then Q* and *P*). It would be domain general in that it can perform this kind of computation on content drawn from any content domain, but that is all that it can do—there are no other types of inferences it can handle. In that case, it would be a general-purpose mechanism vis-à-vis content domains (making it domain general) but special purpose vis-à-vis its range of cognitive operations (making it functionally specific).[38]

[37] This distinction is similar to ideas raised in Barrett (2009), which examines the notion of domain specificity in connection with a commitment to an adaptationist perspective, and in Sperber (1994) and Carruthers (2006), which are primarily concerned with offering an account of what modules are. For our purposes, we can remain neutral as to whether any of the traits in question are adaptations or what exactly makes a cognitive mechanism a module.

[38] Carey's rationalist account of the origins of number concepts, briefly discussed in the previous section, draws on several functionally-specific domain-general mechanisms. For example, a specialized mechanism for keeping track of the positions of items in ordered lists would be functionally

Box 6

Domain Specific—a domain-specific learning mechanism is one that is directed at just one or a very small number of domains, especially when these are closely related in content.

Domain General—a domain-general learning mechanism is one that is directed at more than just a few domains, especially when these are not closely related in content. (Such a mechanism is directed at multiple domains not by collapsing them into a broader domain but by being successively directed at each of the domains when it is processing information pertaining to that domain.)

Functionally Specific—a functionally-specific learning mechanism is one that only has one kind of function or a very small range of closely related functions.

Functionally General—a functionally-general learning mechanism is one that has more than a small range of functions, especially when these are not closely related.

Domain specificity and domain generality play a crucial role in the rationalism-empiricism debate in light of the fact that innate domain-general learning mechanisms are prototypical characteristically empiricist psychological structures that are at the very heart of empiricist accounts, while innate domain-specific learning mechanisms are prototypical examples of characteristically rationalist psychological structures that, likewise, are at the very heart of rationalist accounts. As we noted earlier, there is an asymmetry in the way that rationalists and empiricists can make use of such structures in their respective accounts. Positing any number of domain-general psychological structures as part of the acquisition base does not change a rationalist account into an empiricist account. However, positing more than a few domain-specific psychological structures as part of the acquisition base (especially in a local rationalism-empiricism debate) typically means that an otherwise empiricist account would no longer be empiricist.[39]

specific—only encoding and recovering ordinal relations within such lists—but would be domain general in that it could perform this function for any content domain.

[39] The relationship between functional specificity and functional generality and rationalism and empiricism is broadly similar, though whereas positing more than a few domain-specific psychological structures as part of the acquisition is incompatible with empiricism, empiricists can accept a larger number of functionally-specific psychological structures as part of the acquisition base without effectively becoming rationalists.

2.5 What Makes One Account More Rationalist (or More Empiricist) Than Another?

Up to this point, we have largely focused on understanding what it is that makes an account fall within one or the other of the two frameworks of rationalism and empiricism, and on clarifying key distinctions and terminology needed to fully understand debates between rationalists and empiricists regarding the origins of psychological traits. But there are many possible theories within the frameworks of rationalism and empiricism, both for global and for local debates, and it will sometimes be useful to be able to compare some of these to determine whether, and the extent to which, one is *more rationalist* or *more empiricist* than another. In this final stage of developing our account, we will explore how to make these comparisons.[40]

In particular, we will be highlighting a number of factors that effectively provide independent dimensions along which one account might be more or less rationalist or empiricist than another. We will also briefly consider how these dimensions interact and how trade-offs among different dimensions affect the overall profile of how rationalist or empiricist an account is. The factors that we identify don't allow for fine-grained comparisons of the extent to which different accounts are rationalist or empiricist. But making such comparisons isn't something that we see there being much point to doing in any case. Ultimately, the real value of highlighting and clarifying these factors is that doing so leads to a deeper understanding of rationalism-empiricism debates and the range of positions available in such debates. The set of factors that we identify will also allow us to provide a more precise statement of what makes an account fall within one or the other of the overall frameworks of rationalism and empiricism.

Dimension 1: Quantity. The first factor concerns the number of psychological structures that an account posits as part of the acquisition base. In some ways, this is probably the most obvious factor involved in determining how rationalist or empiricist an account is. Just as positing a highly limited number of innate psychological structures has long been taken to be the hallmark of empiricism, positing a greater quantity of innate psychological structures has likewise been taken to be a paradigmatic feature of rationalism. So variation along this dimension is clearly a factor for what makes one account either more or less rationalist or empiricist than another. Other things being equal, a commitment to more structures in the acquisition base makes a view less empiricist and more

[40] Up until now, we have been adopting the simplifying assumption that individual learning mechanisms are either rationalist or empiricist without qualification. But the considerations that we will discuss below, which are concerned in the first instance with how different rationalist and empiricist accounts can vary in the extent to which they are rationalist or empiricist, also provide a framework for understanding how individual learning mechanisms can vary in the extent to which they are rationalist or empiricist.

rationalist, and a commitment to fewer structures in the acquisition base makes a view more empiricist and less rationalist.

The first point to note regarding the contribution of quantity to these comparisons is that quantity alone is typically sufficient to make one view more rationalist or empiricist than another. Notice that quantity can make a difference even when the only psychological structures that are being considered are characteristically empiricist psychological structures. For example, if two accounts are otherwise alike but one holds that the acquisition base contains just a single domain-general learning mechanism and another holds that it contains many domain-general learning mechanisms, the second would count as empiricist to a lesser extent (and rationalist to a greater extent), even if they are both empiricist accounts.

So, there is clearly a sense in which the quantity of psychological structures in the acquisition base is an important factor regarding how rationalist or empiricist an account is. That said, there are complications in assessing the contribution of quantity in determining the extent to which accounts are rationalist or empiricist. One reason for this is that even an empiricist might posit a large number of psychological structures as being part of the acquisition base. For example, an empiricist might posit a great many fine-grained low-level sensorimotor representations of different types. In principle, the numbers here might be extremely large. Estimates of the number of different shades of colours that are discriminable in human vision are in the millions (Pointer and Attridge 1998). And estimates of the number of different kinds of olfactory stimuli humans are capable of discriminating are over a *trillion* (Bushdid et al. 2014). These two examples only scratch the surface of the full range of types of sensorimotor representations that might be taken to be part of the acquisition base, even by a staunch empiricist. Moreover, if different theorists were to have different estimates of the number of discriminable colours but didn't otherwise differ regarding the acquisition base, this wouldn't seem to have very much at all of an impact on where they stand in the rationalism-empiricism debate. Even if one of these theorists posited twice as many discriminable colours than the other, this wouldn't necessarily make their view substantially more rationalist.

A related issue concerns the fact that different approaches might count what is effectively the same innate endowment in different ways. For example, two theorists might both suppose that the ability to represent different levels of brightness is innate, but the first might see this as involving a large number of representations, each a separate psychological structure in the acquisition base (corresponding to each discriminable level of brightness), while the second might see this ability as involving a relatively small number of psychological structures that have different settings (with different combinations of settings representing different levels of brightness). These theories may not differ in terms of how rationalist they are in any meaningful way. Still, they might be taken to differ in terms of the number of psychological structures they claim to be innate for what are essentially book-keeping reasons. In short, while the quantity of psychological

structures posited as being part of the acquisition base is clearly an important factor in comparing how rationalist or empiricist different accounts are, quantitative comparisons are not always entirely straightforward.

One way of partially addressing this complication is to recognize the interaction between the quantity of psychological structures posited and the types of psychological structures posited. Other things being equal, increasing the number of characteristically rationalist psychological structures that are posited in the acquisition base has a greater effect in terms of making one account more rationalist than another than does increasing the number of characteristically empiricist psychological structures that are posited in the acquisition base. Two accounts that are otherwise alike but where one posits even a small number of characteristically rationalist psychological structures while the other posits none at all will generally differ more strongly in terms of how rationalist they are than other accounts that differ only regarding the number of characteristically empiricist psychological structures they posit. Regardless of differences in the number of characteristically empiricist psychological structures that the first two accounts posit, the one that posits characteristically rationalist psychological structures will not only generally count as more rationalist but may in fact no longer be an empiricist account at all.

Dimension 2: Complexity. While the quantity of psychological structures in the acquisition base clearly matters to how rationalist an account is, it is not the only factor. Another important factor is the internal complexity of the innate psychological structures that are posited, particularly for characteristically rationalist psychological structures. For example, there is a notable difference between positing an innate concept (e.g., the concept OR) and positing an innate domain-specific faculty (e.g., a Chomskyan language faculty). While these both count as characteristically rationalist psychological structures, one is vastly more complex than the other. Of course, complexity here will correlate to some extent with quantity (a language faculty will involve a greater quantity of subcomponents than a single concept), but it seems clear that even when quantity is controlled for, greater complexity of the innate structures posited will make an account more rationalist.

One way to see this is by considering a case involving differences of complexity associated with competing views of a given type of proposed psychological structure. For example, a number of different theories might all posit an innate domain-specific language faculty but differ dramatically in terms of the complexity that they associate with this faculty. One theory might posit a more complex system with detailed information about numerous syntactic properties and constructions, while the other posits a less complex system that embodies just a few very general linguistic principles. If we compare this to the example we mentioned just a moment ago, where different theories take the acquisition base to contain very different numbers of sensorimotor representations of a given type,

we can see that things play out very differently in this case than in that one. Even if theories that posit a language faculty with greater complexity also end up positing greater numbers of psychological structures in the acquisition base, unlike the case of quantity of sensorimotor representations, the difference here really *does* make a difference for how rationalist the account is. Positing a substantially more complex innate language faculty makes an account substantially more rationalist.

We should also note that while we have used the example of the language faculty as an illustration, these points about the complexity of the structures in the acquisition base are entirely general. They apply not just to language, but to structures pertaining to memory, vision, emotion, personality, or any other aspect of cognition. Complexity is a further factor, in addition to quantity, that makes its own distinct contribution to how rationalist or empiricist a theory is.

Dimension 3: Degree of articulation. Related to the complexity of the posited psychological structures in the acquisition base is a further factor that can affect how rationalist or empiricist an account is, which we will refer to as their *degree of articulation*. To see what we mean by this, it will help to back up a bit first.

In our initial characterization of the rationalism-empiricism debate, we particularly focused on the general contrast between an emphasis on domain-general learning mechanisms in empiricist theories and domain-specific learning mechanisms in rationalist theories. However, as we saw in section 2.3, learning mechanisms needn't be fully formed in the acquisition base. Learning mechanisms, including rationalist learning mechanisms, can be constructed from a mix of innate and learned components. The degree to which a given learning mechanism is preformed in the acquisition base is a paradigmatic example of what we mean by the degree of articulation of a characteristically rationalist psychological structure.

Notice that degree of articulation is independent of complexity in that, for a learning mechanism of any given degree of complexity, there is a separate question regarding the degree to which it is preformed in the acquisition base. In order to attain its fully articulated state, a complex learning mechanism might require anything from needing no further elaboration to a modest amount of fine-tuning, to acquisition of a few additional critical components, to assembly from scratch from a mix of components drawn from both the acquisition base and a pool of previously learned traits. And different theories will posit different types and different amounts of learning (and other types of psychological processes) to achieve the fully articulated learning mechanism based on what they take it to trace back to in the acquisition base.[41]

A related sense in which there can be differences in the degree of articulation of psychological structures in the acquisition base concerns not the articulation

[41] Degree of articulation is perhaps most clearly associated with learning mechanisms, but it's worth noting that, in principle, any type of complex psychological structure can come in varying degrees of articulation.

of their internal structure, but rather the degree of articulation of their relations to other psychological structures; that is, the extent to which, in the acquisition base, they are already embedded in a larger network of structures.[42] Consider, for example, a cognitive mechanism for recognizing faces. A rationalist account might take there to be an innate system of some degree of complexity and (internal) articulation which will become the mature face recognition system. This mature system is embedded in a network of structures that includes, among other things, relations to an information store regarding the identities of known individuals, a capacity for recognizing emotions based on facial features, a capacity to track and monitor direction of gaze and extract information about what others are attending to, and much else. If a rationalist account of the origins of these further capacities was given, this would of course make the overall account more rationalist. But there is also a question of the extent to which the connections among these different mechanisms are already established in the acquisition base. In principle, such connections might be learned or unlearned. To the extent that they are unlearned, this too would make the overall account more rationalist.

Dimension 4: Diversity of content domains. Another factor which can affect how rationalist or empiricist an account is has to do with the characteristically rationalist psychological structures in the acquisition base—not how many of these there are but how diverse they are in terms of the content domains they are collectively directed at. Of course, diversity will correlate to some extent with quantity. But they are distinct factors as can be seen from the fact that one theory might posit a number of distinct characteristically rationalist structures in the acquisition base that are all directed at the same content domain (e.g., language), whereas another theory might involve a comparable number of characteristically rationalist structures of comparable complexity in the acquisition base that are respectively directed at content domains concerning quite different contents (e.g., objects, emotions, geometry, and moral norms). Despite having much the same number and kinds of characteristically rationalist structures in the acquisition base, the second kind of theory is clearly more rationalist than the first as a result of the greater diversity of content domains its characteristically rationalist structures are respectively directed at.

While the importance of diversity of content domains seems clear as a general factor that should be taken into account, there are questions about the best way of understanding content diversity. For example, an account that posits a given number of characteristically rationalist psychological structures respectively directed at different but closely related content domains seems like it should count as less diverse—and less rationalist—than one that posits the same number of characteristically rationalist psychological structures respectively directed at

[42] Since connected clusters of structures can be taken to effectively constitute learning mechanisms in their own right, the distinction between internal and external articulation can't bear a great deal of theoretical weight. But it is useful for heuristic purposes to highlight some of the different forms articulation can take.

less closely related content domains. So, an account that posits these structures in the domains of number, morality, and language would be more rationalist that one that posits them in the domains of propositional attitudes, emotions, and sensations. Issues like this complicate precisely how diversity should be understood, but the core idea seems clear enough. For present purposes, all that matters is that diversity of content domains, understood in broadly the sense we have outlined constitutes a further dimension in determining how rationalist or empiricist an account is.

Dimension 5: Abstractness. Another dimension of variation concerns the degree of abstractness of the psychological structures in the acquisition base. In the first instance, abstractness applies to representations. And while it can be hard to quantify, for present purposes we can think of it roughly in terms of the semantic distance between any given representation and the lowest-level sensorimotor representations that form the mind's most basic point of contact with the world. A theory that posits only sensorimotor representations in the acquisition base minimizes the abstractness of its innate representations. One that posits innate concepts like INFINITY, GOD, POSSIBILITY, or TRUTH would be abstract to a considerably greater degree and, other things being equal, would consequently also be more rationalist than the first type of account.

Like some of the other factors we've discussed, abstractness will also correlate with quantity to some extent. Theories that posit representations with highly abstract content as part of the acquisition base will typically posit these in addition to the representations with less abstract content, which other accounts might be restricted to. Nonetheless, abstractness is a further factor that goes beyond quantity as such. Two theories might posit the same overall number of psychological structures in the acquisition base, with one positing representations that are considerably more abstract than the other. In that case, the theory positing the representations with greater degree of abstractness would count as more rationalist.

Dimension 6: Degree of domain specificity. Domain specificity plays a significant role in determining whether an account is rationalist, since innate domain-specific mechanisms are paradigmatic characteristically rationalist structures. At the very least, the number of posited domain-specific psychological structures in the acquisition base contributes to how rationalist an account is—other things being equal, the greater the number of such structures in the acquisition base, the more rationalist the view. But it's not just the number that matters. Degree of domain specificity is a factor that affects the degree to which an account is rationalist or empiricist as well. In particular, all else being equal, an account that includes innate domain-specific structures that are domain specific to a greater extent is more rationalist than an account that also includes domain-specific structures but ones that are domain specific to a lesser extent.

As an example, consider two hypothetical innate learning mechanisms that might be involved in acquiring representations of moral norms. Both are domain

specific, but they differ in one important respect. The first is solely directed to moral norms. In contrast, while the second acquires moral norms and takes these norms to form a distinctive category, it also acquires non-moral conventional norms (e.g., norms about what it is appropriate to eat for breakfast or whether it is appropriate to wear shoes in the house), taking these to form a separate distinctive category.[43] Although both of these types of mechanisms are domain specific, the first is directed solely at the domain of moral norms, whereas the second is directed at two domains: the domain of moral norms and the domain of conventional norms. Since the second mechanism is directed at just these two domains, it is still relatively domain specific. But in being directed at two domains rather than one, it is less domain specific than the first type of mechanism, making it rationalist to a lesser degree.

So, degree of domain specificity is a further contributing factor in determining how rationalist or empiricist an account is. Surprisingly, however, it turns out that the degree of domain specificity of a psychological structure is less of an important factor in contributing to the extent to which an account is rationalist or empiricist than it might initially appear to be.

Degree of domain specificity on our view is, as in the example we just gave, a matter of the number of domains a mechanism is directed at. The fewer domains that a domain-specific mechanism is directed at, the more domain specific it is, where being directed at a single domain is being maximally domain specific. It is tempting, however, to think of degree of domain specificity not in terms of the number of domains a mechanism is directed at, but rather in terms of the *breadth* of the domain that a mechanism is directed at. On this alternative way of thinking about degree of domain specificity, a domain-specific mechanism with a narrower domain would be more domain specific than a domain-specific system with a broader domain. For example, consider again Dehaene's account of how concepts of natural numbers are acquired, which is rooted in the approximate number system. This is a domain-specific system that is directed at the content domain of approximate numerical magnitudes. One type of alternative to Dehaene's account, which is widely understood to be more empiricist, is an account which is organized around a system that is directed at several types of approximate magnitudes—spatial, temporal, and numerical—in a way that does not differentiate among them (e.g., Walsh 2003). This type of *spatial-temporal-numerical magnitude system* is directed at a single domain, just as Dehaene's approximate number system is, but the domain that it is directed at has a broader content domain than Dehaene's, encompassing temporal and spatial magnitudes in addition to numerical ones, and treating these magnitudes in an undifferentiated way as all spatial-temporal-numerical magnitudes.

[43] In other words, this mechanism represents moral and non-moral norms as distinct types of norms and exhibits systematic differences in how it functions when dealing with these two types of norms; it doesn't represent them in an undifferentiated manner.

It may seem that the fact that this alternative system is directed at a broader domain makes it less domain specific than the approximate number system and that this explains why it is less rationalist. But as tempting as this view may be, it can't be right (Margolis and Laurence 2023). If it were, then a mechanism with a very broad domain wouldn't be domain specific; it would be domain general. However, a mechanism that is directed at, say, the natural numbers—an *infinite* domain—isn't domain general. It needn't even be more general than some other mechanism that is directed at a finite domain, such as a mechanism for representing a finite number of types of emotions (happiness, anger, fear, and so on). In fact, one domain-specific mechanism can be directed at a domain whose members (considered extensionally) constitute only a very limited subset of the domain that another domain-specific mechanism is directed at without the first being any more domain specific than the second. For example, although every tool is a physical object but not vice versa, a mechanism for acquiring just tool concepts isn't inherently more domain specific than a mechanism for acquiring just physical object concepts (that is, concepts like OBJECT or PHYSICAL SUPPORT, which apply to physical objects in general in virtue of their being physical objects). Neither of these mechanisms is more rationalist or more domain specific than the other—they are both directed at a single domain, and so equally domain specific, even though the domains have very different breadths. So, something else must account for the difference between the approximate number system and the spatial-temporal-numerical magnitude system, explaining why these systems (and those in similar cases) don't seem to be equally rationalist. That something is degree of alignment.

Dimension 7: Degree of alignment. The last factor that we will highlight, which affects how rationalist or empiricist an account is, turns on the relationship between two domains that are associated with a learning mechanism—its *target domain* (the domain that the learning mechanism as a whole is directed at) and its *resource domain* (the domain that the innate resource which the learning mechanism traces back to is directed at). In particular, the more closely related these two content domains are—or as we will often put it, the more closely *aligned* they are—the more rationalist the account is (other things being equal).[44]

Put in these abstract terms, the notion of alignment can be difficult to grasp. But we can see how it works by looking at a couple examples. Consider again Spelke et al.'s (2010) learning mechanism for acquiring Euclidean geometrical concepts, which we discussed in section 2.3. For Spelke et al., possession of Euclidean geometrical concepts requires the capacity to represent distance, direction, and angle. Spelke et al.'s learning mechanism traces back to two critical innate domain-specific resources, each of which contributes some, but not all, of

[44] When there is more than one innate resource that the learning mechanism traces back to, degree of alignment will be determined by the most closely related resource domain.

these representational capacities. One is the reorientation system (which applies to the large-scale navigable environment, which the system represents in terms of the features of distance and direction). The other is a system for representing shapes (which applies to smaller and manipulable objects, which the system represents in terms of the features of distance and angle). Each of these resource systems is a domain-specific system that is directed at a domain concerning broadly geometrical phenomena. So the content domains that these innate resources are directed at are closely aligned with the target domain of the learning mechanism for Euclidean geometrical concepts. Any mechanism whose domain-specific resource is directed at a domain that is broader than just geometrical phenomena, or is not directed at a domain concerning geometrical phenomena per se at all, would not be as closely aligned. On the other hand, the innate resources in Spelke's account are not *perfectly* aligned with the target domain of Euclidean geometry either, since neither represents all three features of distance, direction, and angle.

Given the importance of the notion of alignment, it will be useful to briefly work through a couple more examples, where different learning mechanisms exhibit different degrees of alignment.[45] As a first example, consider some different ways of learning about dangerous animals. One possibility, which we will discuss in Chapter 14, is that there is an innate system that is specifically geared towards learning about dangerous animals as such. But putting this possibility aside, there are other kinds of rationalist learning mechanisms that could be involved in learning about dangerous animals, ones which trace back to other types of domain-specific resources in the acquisition base. One is that the innate resource it traces back to is a system that is directed at the domain of animals more generally (not the domain of dangerous animals). A different possibility is that the innate resource it traces back to is a system that is directed at all natural kinds (not just animals).[46] The point of interest here is that mechanisms for learning about dangerous animals that respectively trace back to these two different types of resources would differ in terms of the degree of alignment between the domains that their respective innate resources are directed at (animals vs. natural kinds) and the target domain that the learning mechanisms as a whole are directed at (in both cases, dangerous animals). The one involving an innate system for representing animals is clearly more aligned with an overall learning mechanism for learning about dangerous animals than the one for representing natural kinds is. It is also the more rationalist account for precisely this reason (see Figure 2.1).

[45] See Margolis and Laurence (2023) for a more detailed example and further discussion of alignment.

[46] A natural kind may be understood here as roughly any category that is conceptualized as having a hidden essence that supports inductive inferences from one instance of the category to others irrespective of how perceptually similar they are. In addition to animals, these include natural phenomena as diverse as other types of living kinds (e.g., plants and fungi), substances (e.g., gold and water), and processes (e.g., lightning). For more on the psychology of natural kinds and especially work at the interface of philosophy and developmental psychology, see Keil (1989) and Kornblith (1993).

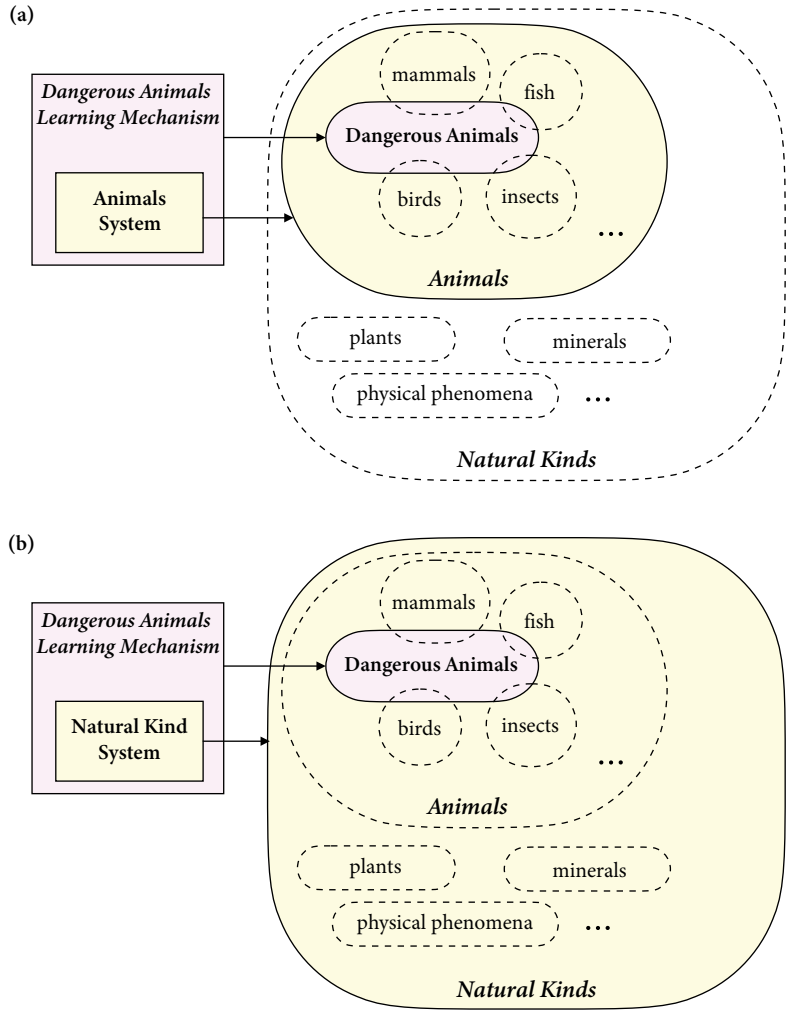

Figure 2.1 Variation in the extent of alignment with a target domain. An example of how learning mechanisms that differ in their degree of alignment differ in the extent to which they are rationalist. Two schematic models of the acquisition of knowledge about dangerous animals trace back to different critical innate domain-specific resources: (a) a system for learning about animals in general, (b) a system for learning about natural kinds more generally. These postulated resources are equally domain specific, but they differ nonetheless regarding their degree of alignment to the domain targeted for learning by the learning mechanisms in question (the target domain being the domain of dangerous animals in both cases). This difference contributes to how rationalist the account is—the greater the degree of alignment, the more rationalist the account.

Neither of the innate resources in this example (or in the case of Spelke et al.'s learning mechanism for acquiring Euclidean geometrical concepts) are perfectly aligned with the target domain of the learning mechanism as a whole. There is no reason why this cannot happen, however, and when it does, other things being equal, the learning mechanism is more rationalist as a result. Take, for instance, two learning mechanisms for acquiring language which trace back to different sorts of domain-specific resources in the acquisition base. In one, the learning mechanisms for acquiring language traces back to an innate system that incorporates the principles of Universal Grammar (which are specific to natural language). In the other, the learning mechanisms for acquiring language trace back to a general communication system that incorporates principles pertaining to communication in general (which are not in any way specific to natural language) (see Figure 2.2). The second of these would be like the other examples we have considered so far, where the domain-specific resources were closely, but imperfectly, aligned with the target domain (which in this case is natural language). But the first would be a case in which the target domain and the domain of the innate resource it traces back to are identical—both are the domain of language—exhibiting the maximal degree of alignment.

Regardless of the further details for spelling out how each of these theories works, both would be deemed rationalist learning theories in virtue of their commitment to key domain-specific resources in the acquisition base that the learning mechanisms trace back to. Nevertheless, it would be wrong to suppose that they are all on par simply because they make use of an innate domain-specific mechanism. Clearly, the first learning mechanism is a more rationalist account of how language is learned. What's more, this can't be because the first learning mechanism traces back to a more domain-specific innate resource than the second. These are equally domain specific, each being specific to a single content domain—the first to the domain of language, the second to the domain of communication. What makes the first learning mechanism more rationalist, then, isn't an *intrinsic* feature of the resource it traces back to. It's a matter of how this resource is related to the domain targeted for learning. The first learning mechanism is more rationalist because the target domain and the domain of the resource the learning traces back to are in greater alignment with one another.

Where does this discussion leave us? Rationalist and empiricist theories do not differ from one another in just a single way. Instead, there are a number of dimensions along which such accounts can differ from one another, each of which can vary independently of the others. One account may be more rationalist and less empiricist—or more empiricist and less rationalist—than another in light of any of these factors. But trade-offs among the different factors are also a possibility, such that being more rationalist with respect to one dimension can mean that an account is more rationalist than another overall, even if it is less rationalist with respect to other dimensions. All of this is true for both rationalist and empiricist

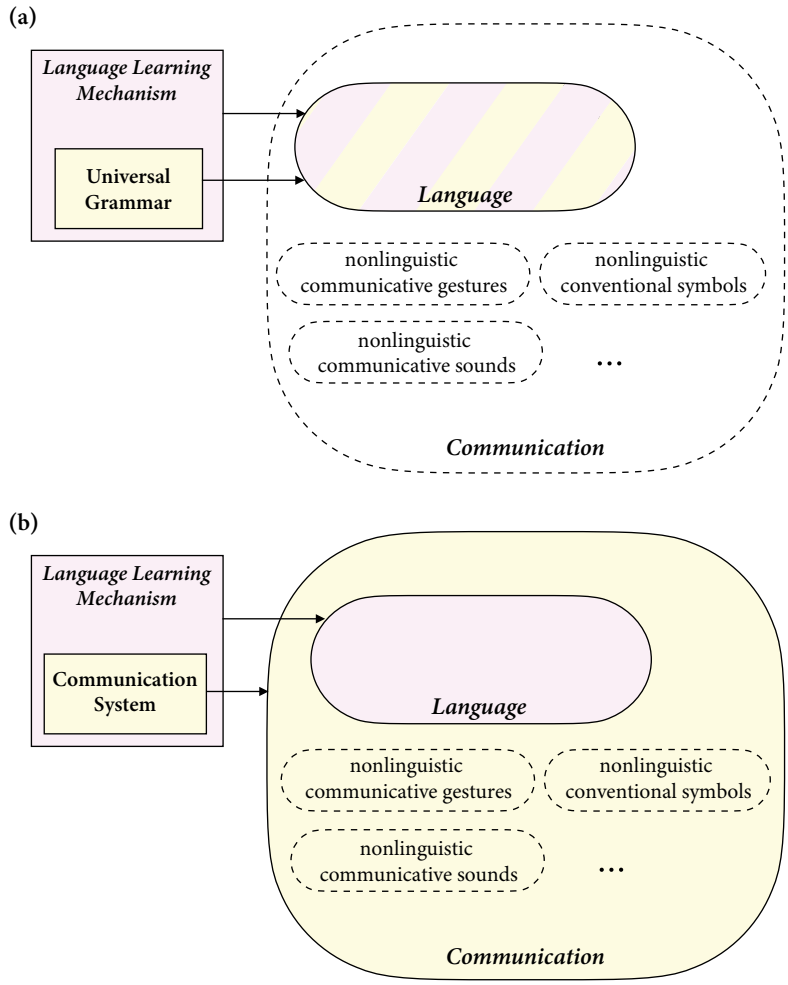

Figure 2.2 Alignment to an extent vs. perfect alignment. (a) A learning mechanism for acquiring language which traces back to an innate resource that is also directed at the domain of language, making the resource domain and the target domain perfectly aligned. (b) A learning mechanism for acquiring language which traces back to an innate resource that is directed at the domain of communication, making the resource domain and the target domain closely, but not perfectly, aligned.

views in any type of rationalism-empiricism debate, local or global. At the same time, most—perhaps all—of these factors are not amenable to precise, fine-grained determinations. As a result, they can only contribute in a coarse-grained manner to the extent to which any given account is rationalist or empiricist.[47]

[47] As we mentioned earlier, we think that there is little if anything to be gained from trying to make fine-grained comparisons in any case. The main purpose in distinguishing and characterizing these factors is to clarify different facets that are in play in the rationalism-empiricism debate and deepen

Box 7

Dimensions of Variation for Positions in Rationalism-Empiricism Debates

1. **Quantity**—quantity concerns the number of psychological structures in the acquisition base (particularly the number of characteristically rationalist psychological structures).
2. **Complexity**—complexity concerns the complexity of psychological structures in the acquisition base (particularly of characteristically rationalist psychological structures).
3. **Degree of Articulation**—degree of articulation concerns the extent to which psychological structures are already elaborated into their full mature form in the acquisition base (particularly for characteristically rationalist psychological structures).
4. **Diversity of Content Domains**—diversity of content domains concerns the set of domains targeted by all of the domain-specific psychological structures in the acquisition base taken together. Each domain-specific psychological structure in the acquisition base targets just one or a small number of domains, but collectively these domain-specific structures may target a wider range of domains. The diversity of content domains is the extent to which this full set of targeted domains is diverse.
5. **Degree of Abstractness**—degree of abstractness concerns the semantic distance between the content of a given representation in the acquisition base and that of the mind's lowest-level sensorimotor representations.
6. **Degree of Domain Specificity**—degree of domain specificity concerns the extent to which characteristically rationalist psychological structures in the acquisition base are domain specific.
7. **Degree of Alignment**—degree of alignment concerns the extent to which two domains are aligned with one another, namely, the target domain (the domain that a learning mechanism is directed at) and the resource domain (the domain that the innate resource which the learning mechanism traces back to is directed at). The more closely related the target domain and the resource domain are, the greater the extent of alignment between them.

our understanding of the various ways in which rationalist and empiricist accounts differ from one another. We will, however, occasionally use these factors to draw broad and general comparisons between our own view of concept nativism with other well-known alternative rationalist accounts, most notably, Fodor's radical concept nativism (particularly in Chapter 26).

Although these factors, in the first instance, concern how accounts can vary in the extent to which they are rationalist or empiricist, they can also be seen as factors that contribute to making an account simply fall either within the overall framework of rationalism or within the overall framework of empiricism. Because there are a number of different dimensions involved and because there are trade-offs to be made among them, what makes a view count as rationalist or empiricist overall is somewhat complex. The short story is that a view counts as rationalist if the combination of weights across different dimensions is sufficiently in the direction of being more rationalist; likewise, a view is empiricist if the combination of weights across different dimensions is sufficiently in the direction of being more empiricist. But these can be spelled out in different ways. For example, a view might be rationalist in light of there being characteristically rationalist psychological structures pertaining to *many domains* even when these are relatively lacking in complexity, relatively unarticulated, or not especially closely aligned with the target domains. At the same time, a view might also be rationalist in light of there being characteristically rationalist psychological structures pertaining to *just a few domains* but where these are relatively complex, richly articulated, or closely aligned with the target domains.

Much the same applies in characterizing what makes a view a version of concept nativism—that is, a rationalist as opposed to an empiricist account of the origins of concepts. For expository ease, we will often describe what makes an account a version of concept nativism in an abbreviated (and admittedly less accurate) way by saying that *there is a rationalist account of the origins of concepts in more than just a few content domains*, or by saying that a view holds that *concepts in more than just a few content domains are either innate or else acquired via rationalist learning mechanisms*. These ways of describing concept nativism have the advantage of succinctly conveying the general shape of what concept nativism claims. However, they sacrifice much of the nuance of the fuller account we have sketched here, neglecting the many possibilities for trade-offs of various types among the different dimensions that are relevant to whether, and the extent to which, a view is rationalist or empiricist. Accordingly, these glosses, which gesture towards a prototypical form of concept nativism, should not be read as a full and complete description that captures the entire framework of concept nativist positions. Whenever we speak this way, it is simply intended to provide a convenient shorthand for the fuller picture that we have presented in this section, which embraces the many potential trade-offs among the factors we have outlined that are consistent with an overall rationalist view about the origins of concepts.

2.6 Conclusion

The purpose of this chapter has been to provide a more detailed version of our account of what the rationalism-empiricism debate is about and, at the same

time, to clarify some key distinctions and introduce some terminology that will be useful throughout the book. At the most basic level, we take the rationalism-empiricism debate to be about the fundamental psychological structures that form the ultimate basis for learning. These structures, which aren't themselves acquired via psychological processes, comprise what we have called the *acquisition base*. Generally speaking, the empiricist vision of the acquisition base is a frugal one, holding that it is largely restricted to domain-general learning mechanisms, sensorimotor representations, and other characteristically empiricist psychological structures. In contrast, the rationalist vision of the acquisition base takes it to also include many characteristically rationalist psychological structures, paradigmatically including domain-specific learning mechanisms, concepts, and other types of characteristically rationalist psychological structures that rationalist learning mechanisms trace back to.

We have seen that rationalist and empiricist views differ along at least seven dimensions. This is essential to keep in mind since many theorists often end up being overly focused on a single dimension to the exclusion of all others. The reality is that the rationalism-empiricism debate is considerably more complex—and more interesting—than accounts that focus on a more narrow range of dimensions recognize. More generally, though, the perspective on the rationalism-empiricism debate that we have presented in this chapter also serves as an antidote to many unproductive ways of understanding this debate that have led those who see the debate in these terms to think that it should simply be abandoned. We saw this at the start of this chapter with the views of Goodman and Strawson. In Chapter 3 we will see that many theorists understand the debate differently than we do, often, like Goodman and Strawson, taking the debate to be fundamentally confused. Comparing these alternative ways of understanding the debate with our own way of understanding it will help to clarify why we have framed the debate the way that we have and will put us on a solid footing for exploring rationalism's prospects.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0002

3

Why the Rationalism-Empiricism Debate Isn't the Nature-Nurture Debate

Many theorists see the rationalism-empiricism debate as deeply problematic. They think that we should not only reject specific views about the origins of particular traits but that we need to reject the entire rationalism-empiricism debate, which they see as rooted in confusion. In this chapter, we will argue that this rejection of the debate as fundamentally confused is misguided and that the rationalism-empiricism debate should continue to play a major role as theorists of all stripes try to understand the origins and workings of the mind.

Why is it that the rationalism-empiricism debate is often dismissed? Many of the arguments against the rationalism-empiricism debate revolve around variations on the theme that it should be understood as a debate about nature and nurture, where it is then argued that the nature-nurture debate is fundamentally untenable and hence so is the rationalism-empiricism debate. But it is a mistake to equate the rationalism-empiricism debate with the nature-nurture debate. Instead, the rationalism-empiricism debate should be interpreted in terms of the account that we elaborated in Chapter 2, where rationalism and empiricism involve competing views regarding the character of the acquisition base. Given our account of the rationalism-empiricism debate, the arguments that this debate is untenable because of confusions about nature vs. nurture don't raise any substantive difficulties for the debate at all. The fact that our account of what is at issue in the rationalism-empiricism debate renders it immune to these challenges can itself be seen as an argument in support for our account of the debate. And, as we will argue in section 3.2, further support is provided by consideration of how participants in the rationalism-empiricism debate understand the debate in practice—when they are engaged in the details of arguing for or against specific experimental results or particular accounts of the origins of a given psychological trait—and by the fact that it is only by interpreting the rationalism-empiricism debate as a debate about the acquisition base that we can see why it has led to so much productive research in cognitive science.

3.1 Is the Rationalism-Empiricism Debate Fundamentally Confused? Nature, Nurture, and Related Issues

In this section we look at a series of arguments for the view that the rationalism-empiricism debate is fundamentally confused. Many of these stem from interpretations of the debate which conflate it with the nature-nurture debate, where *nature* is understood as concerning the contribution of genes to cognitive and conceptual development and *nurture* is understood as concerning the contribution of the environment to such development.

Everyone is an interactionist. The first of these arguments has to do with the way that genes and the environment interact in producing phenotypic traits. It appears less often in scholarly publications than in popular ones, but it is nonetheless an argument that we have repeatedly encountered in conversations with philosophers and scientists. The argument begins by noting that everyone has to accept that genes all by themselves can't produce an organism or any of its traits, and likewise that the environment all by itself can't produce an organism or any of its traits. Genes and the environment work together; they have to interact. Sometimes this is put forward as a theory of development dubbed *interactionism*. But really it is not so much a theory as it is a truism that no one disputes. In any event, the argument we have in mind takes this truism to offer a damning perspective on the nature-nurture dispute and, by extension, on the rationalism-empiricism debate. The thinking here is analogous to that in the argument from Goodman and Strawson in Chapter 2. It's that if everyone must accept that cognitive traits owe their existence in part to nature (genes) and in part to nurture (environment), then there isn't anything for rationalists and empiricists to disagree about.

As before, the conclusion doesn't follow. Just because rationalists and empiricists agree to the truism doesn't mean that there isn't anything substantive for them to disagree about. In particular, they can still disagree about the way that psychological traits are acquired—that is, about the character of the acquisition base that is involved in the acquisition of various psychological traits. And if you look at the sorts of theories of the origins of psychological traits that rationalists and empiricists actually offer and at the kinds of critiques regarding one another's theories that they make in practice—something we will briefly do in section 3.2 and in far more detail later in the book—there can be little doubt that the character of rationalist theories of acquisition (and their associated acquisition bases) systematically differs from the character of empiricist theories of acquisition (and their associated acquisition bases). Thus, to argue that rationalists and empiricists have nothing to disagree about because they both accept interactionism only serves to obscure the very real differences between rationalists and empiricists. Perhaps the truism of interactionism undermines the nature-nurture debate, but if it does, this would only show that the rationalism-empiricism

debate shouldn't be understood in terms of the debate about nature and nurture. Rather, as we argued in Chapter 2, it should be understood in terms of a highly substantive disagreement about the character of the acquisition base.

Genes and the environment do not make separable contributions to development. A related argument begins with the same starting point, the truism of interactionism, but proceeds in a more interesting and sophisticated manner. The argument is premised on an interpretation of the rationalism-empiricism debate according to which it should be understood as offering competing stands on the relative contributions of nature (genes) and nurture (environment) to development. On this view, rationalism is seen as claiming that nature and nurture interact but nonetheless nature is more important (at least more important in those cases where a rationalist account of a given trait is presumed correct), and empiricism is seen as claiming that nurture is more important. Or, if you like, rationalists place more weight on genes, empiricists on the environment.[1] From this the argument takes it to follow that the debate between rationalists and empiricists is spurious. This is because, given the way that genes and the environment work together, it turns out to be impossible for either to make a more significant contribution than the other to any given trait. As Evelyn Fox Keller explains:

> the Swiss primatologist Hans Kummer remarked some years ago—and Frans de Waal (2002) reminds us—trying to determine how much a trait is produced by nature and how much by nurture, or how much by genes and how much by environment, is as useless as asking whether the drumming that we hear in the distance is made by the percussionist or his instrument. Richard Lewontin offered another metaphor: "If two men lay bricks to build a wall, we may quite fairly measure their contributions by counting the number laid by each; but if one mixes the mortar and the other lays the bricks, it would be absurd to measure their relative quantitative contributions by measuring the volume of bricks and of mortar" (1974, 401). (Keller 2010, p. 7)

These metaphors are meant to convey that genes and the environment are equal partners in development. To even make a single protein requires not just the bits of DNA that, in some sense, code for the protein but also many other cellular materials and environmental conditions. The protein building process breaks down into at least two major subprocesses: a transcription process in which

[1] Unlike some of the other arguments in this section, the type of view criticized by this argument is one that is sometimes—confusedly in our view—endorsed by rationalists and empiricists alike. In some cases, theorists endorse this type of view while simultaneously holding a view somewhat like the one that we advocate in Chapter 2, without recognizing that the two views aren't equivalent. More generally, rationalists and empiricists often indiscriminately and confusedly endorse several different and incompatible understandings of the debate.

messenger RNA is created on the basis of a portion of DNA and a translation process in which the messenger RNA is then used as a template for forming the protein (see Chapter 4 for more on how all of this works). Among other things, the raw materials that RNA transcripts are constructed from must be present together with many different specific proteins, multi-protein complexes, and other kinds of cellular machinery, all working together within an intricate series of processes. And of course the cell's overall physical and chemical conditions must be such that they will maintain the integrity of all these elements and support the chemical transitions involved in all of these processes. In this way, genes are very much like Frans de Waal's drummer. A protein-coding segment of DNA (drummer) can only lead to the production of a protein when the many cooperating and enabling elements of cellular machinery are present and the cell's overall physical and chemical conditions are just right (drum). These further entities and cellular conditions are just as important to outcomes that are informally said to be genetic.

Moreover, because genes and environmental factors have distinct yet complimentary causal roles in the formation of any particular trait, there is no common scale to independently measure their relative contributions. This means that it doesn't make sense to try to quantify how much of a trait is caused by one and how much by the other. For example, it doesn't make sense to say that a person's IQ is owing X% to her genes and Y% to her environment (Sober 1988; Block 1995).

If it isn't obvious that we can't say that a person's genes or environment is more important for a given trait, this is partly because of a common misunderstanding about the sorts of heritability statistics that are routinely reported in behavioural genetics and often widely publicized. Heritability statistics concern the degree to which variation in a trait (e.g., height) in a population correlates with (or in the language of behavioural genetics, is accounted for by) genetic variation in a given environment. A high heritability estimate would indicate that, in the population and environment studied, a considerable amount of the difference in the measured trait correlates with a genetic difference. In contrast, a low heritability score would indicate that, in the population and environment studied, this variation does not strongly correlate with a genetic difference. Suppose, then, that researchers arrive at a high heritability estimate for a trait, such as height in a given species of plant. Doesn't that mean that height in these plants is "more genetic", that nature (as opposed to nurture) should be given more credit for explaining their height? Not at all. Heritability statistics don't say anything about what *causes* a given trait—in this case, what causes the plants' height. All they do is measure a correlation regarding the amount of *variation* in a trait (again, variation in a population in a given environment).

Moore (2001) offers a nice analogy that illustrates how little a measure of such correlation says about causation. Snowflakes can only form when both the temperature and the humidity meet certain conditions. There has to be enough humidity for precipitation to take place, and also, the temperature has to be below

freezing. Suppose one day that the humidity is high at the North Pole and low at the South Pole, but in both places it is well below 0° Celsius. Then the variation in snowfall across these two locations is completely accounted for by the variation in the humidity—it correlates with this variation 100%. But obviously the temperature isn't any less important in the causation of the snow. On the contrary, it is extremely important. If the temperature weren't below freezing, the North Pole would only see rain. Indeed, we may suppose that, on the same day, the humidity at the North Pole is identical to the humidity in a forest in a temperate zone, where the temperature is well above freezing. In this case, the variation in snowfall is entirely accounted for by the variation in the temperature—it correlates with this variation 100%. But again, that doesn't make the humidity any less important regarding what actually causes the snow.

The argument we are considering begins with the fact that, like the drummer and the drum, genes and the environment don't make contributions to the development of traits that can be quantified (X% from the genes, Y% from the environment) and concludes that there is a deep problem with the rationalism-empiricism debate, understood as a disagreement about the relative importance of nature and nurture to development. But the conclusion that ought to be drawn, we'd suggest, is that this just goes to show that the rationalism-empiricism debate shouldn't be identified with the nature-nurture debate or understood in terms of the relative contributions of genes and the environment to development. Notice that our own account of what is at stake in the rationalism-empiricism debate in Chapter 2 makes no mention of genes per se. It is framed in terms of a question about the character of the acquisition base. We take it that rationalists and empiricists can all agree that the development of the acquisition base depends on interaction of genes and the environment and that, like the bricks and mortar or the drum and drummer, it makes little sense to say that one is more important than the other. Still, this doesn't mean that there isn't room for there to be systematic and substantial disagreement between rationalists and empiricists—in particular, there is room for disagreement about the character of the psychological structures that are in the acquisition base.

Rationalism (or empiricism) is manifestly wrong. The next argument rejects the rationalism-empiricism debate—once again understood as a disagreement about nature and nurture—on the grounds that at least one side in the debate is manifestly wrong. On this argument, the relevant side in the debate is thought to be so off the mark that it shouldn't be taken at all seriously and that consequently any debate in which its status is at issue isn't a debate worth having.

For example, in his Presidential Address to the International Conference on Infant Studies, David Lewkowicz argues that the rationalism-empiricism debate, which he identifies with "the nature-nurture dichotomy", is "biologically implausible" (Lewkowicz 2011, p. 331). Research into developmental processes, he claims, "renders simplistic questions such as whether a particular behavioural capacity is

innate or acquired scientifically uninteresting" (p. 331). The views he rejects are indeed simplistic. Here is his core argument against rationalism:

> the rationalists' assumption that structure and function are predetermined by genes is a *non sequitur* because no organism can possibly develop in a vacuum; its environment must in some measure contribute to its development. Certainly, everyone would agree that no organism can develop in the absence of oxygen, proper nutrition, and the correct temperature, never mind the usual stimulation that organisms receive from their caregivers. (p. 344)

And here is his argument against empiricism:

> The empiricists' assumption that structure and function are fully determined by environmental influences is equally problematic in that an organism's biological endowment (however, loosely it might be defined) obviously contributes in a major way to its development. (pp. 344–345)

These doubts are closely related to the first argument we looked at, which aimed to use interactionism to deflate the rationalism-empiricism debate. However, rather than taking the truism of interactionism to show that there can be no substantive difference between rationalism and empiricism (since both must endorse interactionism), Lewkowicz's argument takes rationalism and empiricism to be views that in one way or another *reject* interactionism.[2] However, this line of thought crucially depends on a markedly uncharitable reading of the rationalism-empiricism debate, as if rationalists and empiricists think that the source of psychological traits can only be credited to one thing, genes or the environment. No one actually holds such a view; certainly none of the theorists that Lewkowicz mentions by name do.[3]

A similar argument to Lewkowicz's identifies empiricism with the view that the mind begins as a blank slate in the sense that it has no innate structure whatsoever. The problem with this view, as we have already noted, is that a blank slate cannot learn anything. So if the rationalism-empiricism debate were to turn on whether or not the mind is initially a blank slate, then it would hardly be worth pursuing; one side of the debate would be a non-starter. Spencer, Blumberg, et al. (2009) cite just this rationale for their negative assessment of the debate.

> We reemphasize that a developmental systems view [the view they endorse] is not the classical counterpoint to the nativist [i.e., rationalist] program—we are not arguing for a return to empiricism and notions of a "blank slate." After all,

[2] For similar arguments, see Elman et al. (1996) and Moore (2001).

[3] Lewkowicz's cited rationalists couldn't be more explicit about rejecting the views that he deems biologically implausible. See, e.g., Spelke and Newport (1998) and Marcus (2004).

> the notion of a "blank slate" is just as poorly grounded as claims about "primitives" and "essences". (p. 84)

They go on to conclude that "it is time to retire the nativist-empiricist dialog and encourage a new dialog" (p. 85).[4]

In much the same vein, Newcombe (2002) identifies empiricism with the position that the mind is a blank slate and rationalism with the position that the environment plays no role in development.[5] She then points out that even the theorists who are considered to be the most forthright proponents of rationalism and empiricism don't actually hold these patently indefensible views and suggests that this is good reason to abandon the debate:

> the more one considers the debate between nativism and empiricism, the more one concludes that neither extreme possibility is viable. John Locke and Noam Chomsky are two thinkers often presented as clear examples of empiricist and nativist approaches to the origins of knowledge. However, Locke recognized that infants are innately endowed with sensory equipment and a propensity for forming associations, and Chomsky was certainly aware that exposure to a particular language in the environment is vital for becoming, for example, a Chinese speaker rather than a speaker of Swahili. So each man, in his own way, is a type of interactionist, if interactionism is simply defined as recognizing a role for both nature and nurture in development. Rather than endlessly replaying the empiricist-nativist debate, researchers need to get on with the detailed work of proposing exactly how starting points in infancy—stronger than those postulated by Piaget—are transformed into mature competence—perhaps not quite in the way Piaget imagined, but nonetheless in generally interactional ways. (Newcombe 2002, p. 400)

Spencer, Blumberg, et al. and Newcombe, like Lewkowicz, would be right to dismiss the rationalism-empiricism debate if rationalism and empiricism were the views they take them to be. There is certainly no interest in a debate about cognitive development in which one side holds that the environment has no role to play whatsoever and the other holds that the mind has no innate structure of any kind. But this is a fundamental misunderstanding of what rationalism and empiricism

[4] A number of quotes in this chapter employ the term "nativism" or its cognates and refer to the "nativism-empiricism debate". As we noted in Chapter 1, "nativism" is one of several terms that is used in a way that is equivalent to our use of "rationalism" (others include "innatism" and "innativism").

[5] The view that rationalism (nativism) holds that the environment plays absolutely no role in development is nearly as common a misunderstanding of the rationalism-empiricism debate as the view that empiricism holds that the mind is a blank slate. (For one of many examples of this misunderstanding, see the popular textbook *An Introduction to Developmental Psychology* (Slater and Bremner 2017)). Since the environment plays at least some role in the acquisition of literally every trait, this construal of rationalism also has the same detrimental effect as the blank slate construal of empiricism, draining the debate of any possible interest by making one side of the debate unsustainable.

hold, so any problems these views have don't argue for abandoning the rationalism-empiricism debate; they only argue for abandoning these mistaken understandings of what the debate is about.[6] As we argued in Chapter 2, a much better interpretation of the rationalism-empiricism debate is available in which neither side is committed to such manifestly false views.

The rationalism-empiricism debate (or the rationalist side of this debate) is undermined by problems with the notion of learning. As we have seen, the rationalism-empiricism debate is often identified with the nature-nurture debate, particularly by critics. Some critics take both debates to be undermined by their dependence on a problematic distinction between psychological traits that are learned and psychological traits that are not. According to these critics, we should reject this distinction and its attendant notion of learning and hence we should reject the rationalism-empiricism debate too.

For example, Elman et al. (1996) write that "nature is usually understood to mean 'present in the genotype,' and nurture usually means 'learned by experience'" and suggest that both of these views are faulty:

> The difficulty is that when we look at the genome, we don't really see arms or legs (as the preformationists thought we might) and we certainly don't see complex behaviors.
>
> Learning is similarly problematic. We know that learning probably involves changes in synaptic connections, and it is now believed that these changes are effected by the products of specific genes which are expressed only under the conditions which give rise to learning. (p. xi)

Proponents of developmental systems theory often speak of a related problem with the idea of learning. The criticism, in this case, is that the traditional understanding of learning—the one that appears in the rationalism-empiricism debate—is supposed to be too narrow to do justice to the full range of experiences that matter to development. Lewkowicz (2011) expresses the point this way:

> the learning part of the nativist dichotomy only refers to the traditional concept of learning that includes classical or operant conditioning, training, practice, and imitation through observation. It misses all the other forms of external and internal stimulation and its developmental trace effects that do not qualify as traditional learning effects but that can have profound effects on organisms and their development. All of these effects, together with traditional learning effects, are part of the broader concept of experience. (p. 337)

[6] The views that these critics attribute to rationalists and empiricists don't sit well with rationalists' or empiricists' self-characterizations or with how they argue with one another in practice (see section 3.2). Even without looking at these examples, however, one ought to be deeply suspicious of a characterization of the rationalism-empiricism debate that has the consequence that Chomsky, of all people, isn't a rationalist.

We certainly agree that learning includes much more than conditioning, training, practice, and imitation. It also includes processes of belief formation through instruction and reasoning, various types of perceptual learning, and numerous processes in which the acquisition of a psychological trait is mediated by special-purpose learning mechanisms. At the same time, we don't think the theoretical distinction between *developmental processes that are involved in learning* and *developmental processes that aren't* should be collapsed in favour of a single broad notion: *development that is responsive to experience*. Such a notion is so broad that it would apply equally to human cognitive development and the growth (i.e., development) of daffodils. That's just too broad.

Much the same could be said of an undifferentiated notion of brain activity. Some processes in the brain psychological the neural computations that are involved in cognition, while others are non-psychological even if they support cognition in indirect ways (e.g., cellular respiration). Lewkowicz is of course right to emphasize that a diverse range of processes at multiple levels of organization influence development and leave their mark on the brain. These include gene expression (including cases where gene expression is influenced by the regulatory effects of other genes), neural activity in response to external stimuli, interneural stimulation, effects of changes in hormone levels, cell growth and cell death, the formation of new synapses, biochemical reactions to pheromones, immune response to foreign substances, and so on. No one should deny that these (and many other) processes are part of the full story about the many changes that take place in development. But we see no reason to suppose that all such activity must be conceptualized in an undifferentiated way under a "broader concept of experience". On the contrary, recognizing that there are potentially important differences between, say, learning to read and an immune response to meningitis or a brain haemorrhage caused by a blow to the head is simply to recognize that we need more than one way of accounting for the diverse effects that comprise all of the changes that take place in the brain. Rationalists and empiricists aren't unaware of the full range of potential causes that contribute to development; they are just particularly focused on certain types of causes—ones that cluster around paradigmatic instances of learning and psychologically-mediated developmental processes more generally.[7]

A similar problem affects the concern registered by Elman et al. (1996), noted earlier. Although they are right to suggest that gene expression may play a crucial role in the changes that take place in learning, this is hardly a reason to abandon the notion of learning itself. If anything, it is a reason to not characterize learning in a way that excludes the possibility of genes playing a role in learning. What's more, although rationalists and empiricists don't often talk about the low-level physical details that implement learning, most theorists in the rationalism-empiricism debate—rationalists and empiricists alike—would happily grant that

[7] We will return to the question of what counts as a learning process in Chapter 25.

gene expression plays an important role, just as they would be happy to grant the significance of such basic neurological processes as myelination and long-term potentiation. There is nothing in rationalist or empiricist views of cognitive development that should be thought to discourage their proponents from taking on discoveries about cellular mechanisms and processes involved in learning.

The objections by Lewkowicz and by Elman et al. represent one type of concern about the notion of learning which is supposed to undermine both rationalism and empiricism. But it's worth mentioning another potential concern about the notion of learning, one that is directed particularly to rationalism and that is motivated by the opposite perspective. In this case, it's assumed that learning is a perfectly coherent notion and that learning *does* take place in development. Then the problem with rationalism is supposed to be that rationalism is inherently opposed to learned psychological traits.

The problem with this objection is that, as we saw in Chapter 2, what is characteristic of rationalism, as opposed to empiricism, isn't that rationalism is anti-learning but rather that rationalism sees some learning as being mediated by importantly different types of mechanisms than those employed by empiricists. In particular, rationalist accounts prominently involve what we have been calling characteristically rationalist learning mechanisms in addition to the types of learning mechanisms characteristically employed by empiricists. For example, a rationalist view of language acquisition might hold that the learning process involves a set of constrained choices that are specific to language, so that much of what occurs in learning a grammar amounts to choosing between a relatively small number of alternatives (Yang 2006). In much the same way, and for a wide range of acquired psychological traits, rationalist theories embody proposals for how learning unfolds, albeit proposals that may be counterintuitive to theorists who are fundamentally committed to empiricist principles.

While the compatibility of rationalism and learning may be surprising to some, it isn't an idea that is particularly new among rationalists. In fact, one of the foundational documents for contemporary rationalism is C. R. Gallistel's aptly titled book *The Organization of Learning*. In the introduction, Gallistel announces: "My purpose is to sketch a new framework for the understanding of animal learning and the investigation of its cellular basis" (Gallistel 1990, p. 3).[8] The framework he proposes is plainly rationalist in crediting animals with innate specialized computational learning systems that underlie such things as navigation and foraging. If Gallistel is right, it is because animals possess these sorts of innate specialized learning systems that they are able to learn such things as the way home from their current location and the optimal strategy for obtaining food in a given region.

[8] For similar rationalist expressions of the centrality of learning to rationalism which are more focused on human learning, see, among others, Gelman (1990); Wynn (1992); Cosmides and Tooby (1997); Pinker (1997); Keil (1999); Spelke and Newport (1998); Leslie (2004); Sperber and Hirschfeld (2006); Landau (2009); and Marcus (2009).

Developmental theorists should focus on processes, not origins. Another argument put forward by critics of the rationalism-empiricism debate is that this debate gives rise to a problematic focus on the wrong type of explanatory project. The charge is that these views (rationalism and empiricism, or rationalism in particular) are too wrapped up with efforts to explain the origins of psychological traits in development, resulting in a "static" view of the mind. Instead, critics argue researchers should adopt a more "dynamic" approach by focusing on the more valuable project of explaining developmental processes.

Lewkowicz (2011) develops this charge particularly with rationalists in mind:

> nativists motivate their experiments in terms of the nature-nurture dichotomy and ask origins-oriented rather than process-oriented questions. The problem is that the dichotomy ignores the fact that developing organisms are fused systems wherein organismic and environmental factors are in such continuous interaction that it makes no heuristic sense to treat them as separable influences. (p. 345)

Lewkowicz takes himself to be speaking for a diverse group of theorists who broadly subscribe to a developmental systems perspective: "I echo the many prior calls to abandon dichotomous developmental thinking and its focus on the origins question. It is time to shift our focus to the processes question" (p. 355).[9]

In an important and influential discussion of how to understand conditions that result in atypical development, Karmiloff-Smith (1998) offers a related criticism:

> For both the strict nativist and the empiricist, the notion of "environment" is a static one, whereas development (both normal and atypical) is of course dynamic. The child's way of processing environmental stimuli is likely to change repeatedly as a function of development, leading to the progressive formation of domain-specific representations. (p. 390)

Notice that the charge, put this way, is meant to apply to empiricists as much as rationalists. In both cases, Karmiloff-Smith would claim that the theoretical framework is static because it doesn't take into account the possibility that

[9] Lewkowicz cites, among others, Lehrman (1953, 1970); Schneirla (1957); Gottlieb (1997); Oyama (2000); Thelen (2000); Griffiths and Gray (2004); Bateson (2005); Sameroff (2005); and Overton (2006). The latter two are singled out for presenting their views, like Lewkowicz, in their own presidential addresses to the International Conference on Infant Studies: "Thelen offered a framework that enables us to ask how the moment-moment fluctuations in an organism's sensorimotor activity are linked to emerging perceptions, actions, and cognitive structures. She offered this as an alternative to static views of the mind…For Sameroff, outcome depends on the transaction between the organism and its environment where individuals are constantly being changed by and changing their environments" (Lewkowicz 2011, p. 357). Similar sentiments are expressed by Lerner (2015); Witherington et al. (2018); and many others.

domain-specific systems are formed in response to the particularities of experience. Karmiloff-Smith maintains that small differences in an infant's input can lead to attention being selectively applied in ways that cause significant cognitive changes, which may feed into cycles of development that amplify these changes further or set up the possibly of subsequent new types of cognitive processing. Thus, as Karmiloff-Smith sees it, these small differences in early experience can result in divergent ways of processing information even if, at a coarse level, people have what looks like the same cognitive capacity (e.g., people with Williams syndrome, who appear to have strong linguistic skills, may have developed ways of processing language that are very different than those of neurotypical adults, despite surface similarities). Importantly, this view of development doesn't deny the existence of domain-specific cognitive systems. Rather, for Karmiloff-Smith, domain-specific cognitive systems should be understood to be the product of domain-general learning and subject to ongoing changes, as learning is itself a continuous process.

Although Karmiloff-Smith contrasts her view with both rationalism and empiricism, it wouldn't be unreasonable to read her as offering an empiricist framework for modelling development. On our characterization of rationalism and empiricism, we were careful to note that empiricists needn't oppose the existence of domain-specific learning systems as long as, for the most part, they maintain that these are acquired on the basis of more fundamental domain-general systems—in other words, as long as they maintain that the acquisition base is largely domain general. To the extent that Karmiloff-Smith's views fit with this understanding, she may be counted as an empiricist in our sense even if she isn't an empiricist in her own sense of the term. Then the question of whether this brand of empiricism is successful would need to be addressed by looking at particular domain-specific cognitive capacities on a case-by-case basis. As we will be examining a range of examples later, for now it will suffice to say that, although we think Karmiloff-Smith's proposal is a serious empiricist contender, the weight of empirical evidence favours a rationalist treatment of many domain-specific learning systems all the same.[10]

Things are trickier when we turn to the more radical proposal represented in Lewkowicz's remarks. This is because it isn't clear whether rationalism is as static as critics like Lewkowicz maintain, nor is it clear that it would be so bad if it were. For one thing, rationalist learning systems do make use of environmental input. All proposed rationalist language learning systems, for example, are sensitive to the language in a learner's environment. There is also no reason why innate domain-specific learning systems can't make use of environmental input in highly

[10] See Parts II and III below for our case for a rationalist account in many of the domains that Karmiloff-Smith is concerned with. And see Chapter 20 for a detailed examination of her views and arguments.

interactive ways, with feedback loops affecting their development and their subsequent operations. Critics like Lewkowicz place a lot of weight on the claim that development and behaviour are dependent on complex causal interactions operating simultaneously and across different levels of organization, from the small-scale level of genes and cellular processes, to organs, systems of organs, and the external environment. But as a general principle, this isn't anything that rationalists would deny.[11] And, of course, the rationalist acquisition base itself has to be acquired; rationalists would certainly agree that this depends on highly complex interactions between organismic and environmental factors (see, e.g., Marcus 2004 for a rationalist perspective on the interactive nature of the processes involved in establishing the brain's wiring). So the charge of stasis really comes down to the impressionistic claim that rationalist models are insufficiently interactive. Perhaps. But then, perhaps they identify just the right amount of interaction regarding the cognitive capacities they aim to explain. Ultimately, this is a question of which sort of account—rationalist or empiricist—is best supported by the full weight of empirical evidence. Accordingly, it does not provide general grounds to dismiss either the rationalism-empiricism debate or rationalist accounts independent of a detailed look at the evidence for specific rationalist and empiricist models.

Still, we think it might be useful to say something about the sorts of examples that frequently come up in discussions of so-called process-oriented development. A representative and widely cited example is Gottlieb's (1997) study of the mallard duck's imprinting response to maternal calls, which had previously been found to appear in newborn ducklings that had been deprived of prenatal experience of adult mallard vocalizations. As Moore (2001) explains, "Gottlieb understood that this does not mean that the environment in general is unimportant in the development of the trait! So, he began to look for other environmental factors involved in the trait's development" (p. 122). The factor he came to focus on was the vocalizations made by duck embryos themselves in the days that preceded hatching. Interestingly, he found that when embryos are permitted to hear their own vocalization, the postnatal response to maternal calls develops normally, but that embryos that are "devocalized" fail to develop the normal response. Spencer, Blumberg, et al. (2009), who also claim that we should "no longer…abide the nativist-empiricist debate and nativists' ungrounded focus on origins" (p. 79), explain the upshot of this work by Gottlieb as follows: "self-stimulation from embryonic vocalizations tunes the auditory system and establishes a bias that shapes the latter preference for the maternal call" (p. 81).

But why should the mere fact that experience plays a role in development undermine rationalism? The simple answer is that it doesn't. What the rationalist

[11] For example, Steven Pinker, the quintessential rationalist, says that "[t]he development of organisms must use complex feedback loops rather than prespecified blueprints" (2004, p. 12).

is committed to isn't the idea that environmental interaction is unnecessary or unimportant but rather the view that, in addition to the types of psychological structures empiricists posit, characteristically rationalist psychological structures also figure prominently in the acquisition base. And, as should be clear from the discussion in both this chapter and Chapter 2, there is no reason whatsoever why rationalist domain-specific acquisition mechanisms can't make use of experiential input as part of a characteristically rationalist story about cognitive development. If examples like Gottlieb's are to be brought to bear on the evaluation of rationalism, the question shouldn't be whether experience is somehow involved in development. It needs to be about the nature of the mechanisms that transform this experience into the traits whose acquisition they support. If, for example, it turns out that the effect of hearing the self-generated vocalization does its work through the activity of a very general preference-forming system that is capable of forming preferences across a wide range of different domains based on diverse types of sensory inputs, then this would be more congenial to empiricism. On the other hand, if it does its work through the activity of a special-purpose system that employs an auditory template that the experience serves to calibrate and maintain, then this would be more congenial to rationalism.[12]

Empiricism by another name. Given the numerous calls to abandon the rationalism-empiricism debate (often conflated with the nature-nurture debate) and given the widespread sentiment that rationalism and empiricism aren't viable, one might wonder what sort of alternative these critics would put in their place. The simple answer is that all too often they want to replace rationalism and empiricism with…*empiricism*. True, it's not empiricism as they understand the term. But it is, nevertheless, empiricism according to what we have been arguing is the best way to construe the rationalism-empiricism debate. The situation, in other words, is that these critics identify rationalism and/or empiricism with a highly implausible view and then proceed to argue that we should reject both of these—that we should abandon the rationalism-empiricism debate—in favour of a far more reasonable alternative. Such alternatives go by a number of names—*constructivism*, *neoconstructivism*, *neuroconstructivism*, among others—but, in the end, the position these critics settle on is simply their preferred form of

[12] There are other possibilities as well. We mention these two simply to illustrate that the search for a process in development doesn't undermine the rationalism-empiricism debate, as there are questions about the mechanisms involved in these processes, and the rationalism-empiricism debate is about the character of these mechanisms. As it happens, the true story about what is going on in the mallard duck example is unclear. We ourselves are struck by the fact that the vocalizations that the embryos hear are not especially similar to maternal vocalizations. As Moore notes, "there is almost no resemblance at all between the peeping of unhatched ducklings and the calls produced by mature mallard ducks" (Moore 2001, p. 122). This would suggest that the mechanism isn't an instructive experience-driven empiricist one, but this leaves open many possibilities regarding the nature of the mechanism and how it operates. See Chapter 10 for further discussion of Gottlieb's study.

empiricism, as we understand this term. The trouble is that this isn't a legitimate way to argue for empiricism. Showing that a critic's preferred view is better than caricatures of rationalism and empiricism hardly shows that it is the best view; there will be *many* views that are better than these caricatures. And if empiricism (properly understood) is still on the table, then so is rationalism and the rationalism-empiricism debate (properly understood).

An illustrative example can be found in the influential book *Rethinking Innateness* (Elman et al. 1996).[13] As we saw above, Elman et al. argue that the nature-nurture debate is fundamentally confused since neither the nature nor the nurture position makes sense (they take nature to entail preformationism and nurture to exclude genetic influences on learning).[14] From this, they conclude that we should abandon this debate and adopt an interactionist perspective, which they refer to as *constructivism*:

> The obvious conclusion is that the real answer to the question, *Where does knowledge come from*, is that it comes from interaction between nature and nurture, or what has been called "epigenesis". Genetic constraints interact with internal and external environmental influences, and they jointly give rise to the phenotype. Unfortunately, as compelling and sensible as this claim seems, it is less a conclusion than a starting point. The problem does not go away, it is simply rephrased. In fact, epigenetic interactions must, if anything, be more complicated than the simpler more static view that x% of behavior comes from genes and y% comes from the environment. For this reason, the interactionist (or constructivist) approach has engendered a certain amount of skepticism on the part of developmentalists…In fact, we believe that the interactionist view is not only the correct one, but that the field is now in a position where we can flesh this approach out in some detail. (pp. xi–xii)[15]

In spelling out their positive model of development, Elman et al. make extensive use of domain-general connectionist networks:

> throughout this book we advocate that a developmental perspective is essential to understanding the end state [the adult mind], and that the connectionist framework, with its focus on learning rather than on-line steady-state computations, is especially relevant to that endeavor. (p. 109)

[13] For other examples of this form of argument, see among others, Dupré (2003); Karmiloff-Smith (2009b); Spencer, Blumberg, et al. (2009); Stiles (2009); and Churchland (2012).

[14] See again the section *The rationalism-empiricism debate (or the rationalist side of this debate) is undermined by problems with the notion of learning.*

[15] We addressed the view that Elman et al. call the "static view" ("that x% of behaviour comes from genes and y% comes from the environment") in the section *Genes and the environment do not make separable contributions to development*, above.

The beauty of the connectionist framework, according to Elman et al., is that connectionist networks can achieve their results with only the most general constraints on their overall organization and the learning environments in which they operate.

In short, Elman et al. argue for what is essentially an empiricist view—in which domain-general learning mechanisms play a large role in cognitive development—after mistaking rationalism for a kind of preformationism and empiricism for the doctrine that genes are irrelevant to development. The problem with this approach is that rationalism and empiricism (properly understood) are both perfectly compatible with Elman et al.'s driving motivation, which is to recognize that theories of cognitive development have to be interactionist. To the extent that their interactionism allows for domain-general connectionist learning mechanisms, it allows for domain-specific learning mechanisms too. There may be further reasons to question whether rationalism offers the better explanation of how particular cognitive capacities are acquired—we will come to these later—but the present point is that we have ample reason to reject the claim that the rationalism-empiricism debate itself is bankrupt and to reject the idea that this should warrant our adopting what is in fact a version of empiricism (albeit under a different name).

Many philosophers and scientists claim that there is something deeply wrong with the rationalism-empiricism debate—so wrong that the only reasonable response is to simply abandon this debate. We have been arguing that this scepticism is not warranted and that these alleged problems with the rationalism-empiricism debate are often really nothing but artefacts of misguided and counterproductive ways of interpreting the debate. Fortunately, there is a better way of understanding the rationalism-empiricism debate—namely, the approach that we offered in Chapter 2. On this understanding, the rationalism-empiricism debate is about the character of the acquisition base—the psychological structures whose acquisition is not mediated by more fundamental psychological acquisition systems, and which ultimately explain the origins of all other psychological structures. This interpretation makes clear that rationalists and empiricists aren't arguing over truisms (e.g., that infants have the capacity to acquire the cognitive capacities that they acquire, or that genes and the environment interact). And it offers an illuminating framework in which many of the debate's critics, when they aren't arguing against a straw man, maintain recognizable views *within* the rationalism-empiricism debate (i.e., they are actually opposed to rationalism, not to the coherence of the rationalism-empiricism debate).

3.2 The Rationalism-Empiricism Debate in Practice

Coupled with the discussion in Chapter 2, the arguments in the previous section show that the interpretations of the rationalism-empiricism debate that its many

critics take for granted are not mandatory and that there is a perfectly coherent alternative—the one given in Chapter 2—for which the problems that allegedly undermine the debate simply don't arise. This in itself provides strong grounds for adopting our account of the rationalism-empiricism debate. But are there other grounds for the interpretation we gave in Chapter 2 which should be at least briefly noted? In fact there are three: (1) Framing the debate in terms of competing views regarding the character of the acquisition base does justice to what rationalists and empiricists actually say about the debate. (2) Crucially, this way of framing the debate is at the heart of *the arguments* that rationalists and empiricists give when evaluating specific theories and experimental results regarding the origins of any given psychological trait. (3) Finally, framing the debate in these terms and not as a debate about nature versus nurture (or genes and the environment) makes sense of why the debate has proven to be so productive.

Let's begin with what rationalists and empiricists themselves say about the debate and about their own positions in this debate. In many cases, both rationalists and empiricists are explicit about the key element of the disagreement being about the nature of the acquisition base. The attention to characteristically rationalist psychological structures, such as innate representations and innate domain-specific learning mechanisms, can be seen in early contemporary discussions of these matters, particularly in Chomsky's writings.

For example, in *Aspects of the Theory of Syntax*, Chomsky says:

> The empiricist approach has assumed that the structure of the acquisition device is limited to certain elementary "peripheral processing mechanisms"…Beyond this, it assumes that the device has certain analytical data-processing mechanisms or inductive principles of a very elementary sort, for example, certain principles of association…A rather different approach to the problem of acquisition of knowledge has been characteristic of rationalist speculation about mental processes. The rationalist approach holds that beyond the peripheral processing mechanisms, there are innate ideas and principles of various kinds that determine the form of the acquired knowledge in what may be a rather restricted and highly organized way. (1965, pp. 47–48)

Steven Pinker, one of the foremost public spokespersons for rationalism, also writes that:

> Everyone [in the rationalism-empiricism debate] acknowledges that there can be no learning without innate circuitry to do the learning…The disagreements…, though significant, are over the details: how many innate learning networks there are, and how specifically engineered they are for particular jobs. (2002, pp. 35–36)

Contemporary rationalist developmental psychologists express similar thoughts:

> nativists and empiricists primarily disagree on the extent to which pre-existing biases for specific domains of information go beyond those in effect at the levels of sensory transducers. (Keil 1999, p. 585)

> By core cognitive architecture I mean those human information processing systems that form the basis for cognitive development rather than its outcome (Leslie 1988). Understanding this core is the primary aim of all theories of cognitive development. One view of the core is that it is essentially homogeneous and that any differentiation of its architecture is the product of development. The general all-purpose learning device of classical associationism is an elegant and influential example of this view. An alternative view of the core is that it contains heterogeneous, task-specialized subsystems. (Leslie 1994, p. 120)

> The substantive issue concerning a given body of knowledge is, "what is the nature of the built-in mental mechanisms that are responsible for the emergence of the knowledge?" With regard to this question, the term 'empiricist' typically applies to accounts positing a general-learning mechanism (classically, the laws of association), while "nativist" applies to accounts involving domain-specific mechanisms. (Wynn 1992, p. 378)

And rationalist evolutionary psychologists characterize the debate in similar terms:

> Historically, there have been two basic conceptions of human nature: the empiricist conception, in which the brain is thought to comprise only a few domain-general, unspecialized mechanisms; and the nativist conception, in which the brain is thought to comprise many, domain-specific, specialized mechanisms. (Symons 1992, p. 142)

> the real nature-nurture debate is between those who believe the human mind has many psychological mechanisms that are domain-specific and special-purpose (e.g., mate-choice mechanisms), and those who believe human behavior is the product of a few global, domain-general mechanisms (e.g., the culture theorists' hypotheses about culture-learning, norm imitation, etc.). (Tooby and Cosmides 1989, p. 36)

> The genuine disagreement is not about the relative importance of "nature vs. nurture" in development (an inane formulation that has spectacularly impeded progress; one might as well ask whether hemoglobin or air is more essential to human survival). The difference is simply this: Those who derive explicit inspiration from selection thinking commonly expect the evolved mechanisms of the human mind to be numerous and specialized, whereas most psychologists and social theorists seem to believe that relatively few general-purpose mechanisms will do the job. (Daly and Wilson 1988, p. 9)

Nor is it only rationalists who see the debate in these terms. Here, for example, is the empiricist philosopher Jesse Prinz's overall characterization of the debate:

> For Empiricists, the crucial thing is generality. We have innate resources that help us acquire knowledge, but the very same resources are used to learn about very different kinds of things…These resources include our senses, some general-purpose learning rules and perhaps even a few innate concepts, like those on Kant's list. For Rationalists, the innate machinery is much more specialized…In the lingo, these resources are "domain-specific". (Prinz 2012, pp. 85–86).

Similarly, empiricist developmental psychologists have highlighted the disagreement regarding domain-specific learning systems. For example, in an article with the subtitle "Infant Rule Learning Is Not Specific to Language", Saffran et al. (2007) write:

> A central issue in cognitive neuroscience concerns how the brain is functionally organized. One view is that discrete systems exist in the human brain for solving specific problems facing the organism, such as learning language or processing faces. Alternatively, learning mechanisms may operate more generally, with similar processes underlying multiple functions. (pp. 669–670).

Empiricists also explicitly take issue with rationalists in just these terms. For example:

> within the last few years these nativist views increasingly have come under fire, and alternative explanations are appearing in the literature…In each case, the newer studies indicate that simpler perceptual and attentional processes can explain the apparent precocious performance of young infants…The view of infant cognitive development that we propose depicts infants developing their knowledge about the world by way of a continuous interplay between a set of domain-general learning mechanisms and changing environmental experiences. (Cohen et al. 2002, p. 1324)

> we suggest that a conventional nativist picture, stressing domain-specific, innately specified modules, cannot be sustained. (Chater and Christiansen 2010, p. 1132)

> despite using only domain-general constraints, the connectionist model of semantic learning explains evidence others use to argue that children rely on innate domain-specific constraints. (McClelland et al. 2010, p. 353)

In short, rationalists and empiricists both frequently and explicitly describe the content of the rationalism-empiricism debate and their own views within this debate in terms that are broadly in agreement with the basic characterization of the debate we outline in Chapter 1.[16]

This brings us to the second reason for favouring our interpretation of the rationalism-empiricism debate. It is not only when they are reflecting on their respective conflicting general theoretical commitments that rationalists and empiricists highlight their differing views regarding the character of the acquisition base. Even more significantly, we think, competing perspectives on the character of the acquisition base are at the heart of the *arguments* they use when they are involved in the nitty-gritty business of evaluating specific theories and experimental results. In domain after domain, rationalists and empiricists can be seen to be arguing with one another regarding the psychological structures that underlie particular aspects of cognitive development, with rationalists claiming that empiricists' proposed domain-general mechanisms cannot explain development in particular domains as well as rationalist accounts that posit innate domain-specific elements (and characteristically rationalist psychological structures more generally), and empiricists claiming that rationalists' proposed domain-specific mechanisms are not warranted or are otherwise inferior to empiricists' domain-general ones.

Later we will see many examples of these types of arguments (especially in Parts II and III). Here we will point to just a few instances, without elaboration.

- Spencer, Blumberg, et al. (2009) writing about aspects of language acquisition claim that "statistical learning provides a clear alternative to nativist views" (p. 82) and that domain-general "connectionist networks can capture statistics of sequences and contextual dependencies (e.g., Elman, 1990)" (p. 83).
- Elman et al. (1996) argue that domain-general connectionist models provide a better account of cognitive development for the representation of objects than rationalist models, such as Spelke (1991).
- Scarf et al. (2012) argue that simple domain-general processes of association may explain the data reported in Hamlin et al. (2007), which Hamlin and her colleagues have taken to argue for a rationalist treatment of social evaluation and the origins of moral judgement in terms of innate concepts and innate domain-specific learning mechanisms.

[16] Admittedly, however, as we have noted earlier, it is not uncommon for theorists to also characterize the debate in other terms that are incompatible with this understanding, including in terms of the relative contributions of genes and the environment. But, tellingly, when it comes to actually arguing for or against particular rationalist or empiricist accounts, these authors focus almost exclusively on the kinds of considerations that speak to what may or may not be part of the acquisition base (as we'll see in a moment).

- Ray and Heyes (2011) argue against rationalist accounts of imitation and contrast them with domain-general learning accounts, arguing for the view that "natural selection has shaped the human mind, not by producing complex, specialized cognitive 'modules' (e.g., Cosmides and Tooby 1994), but by favouring relatively simple behaviour-control mechanisms that channel the effects of domain- and taxon-general cognitive processes (Heyes, 2003; Sterelny, 2003)" (p. 102).
- Rogers and McClelland (2004) argue that domain-general connectionist models can explain not only how learners detect correlations among a category's features but also which features are more diagnostic for categories in different domains. They put this forward as one of a number of examples in which a domain-specific aspect of mature semantic memory can be acquired on the basis of domain-general principles.
- Perner and Ruffman (2005) argue that domain-general processes of association account for the results in Onishi and Baillargeon (2005), which rationalists have cited as demonstrating the ability of infants to represent false beliefs and have viewed as a key finding in the case for a rationalist theory of mentalizing abilities in terms of innate concepts and innate domain-specific learning mechanisms.

These examples are all of empiricists arguing against rationalists, not in abstract terms, and not in terms of the relative importance of genes or the environment, but regarding ongoing research and concrete proposed models of development, where domain-general learning mechanisms are argued to provide a better explanation of experimental results than competing accounts involving innate representations or characteristically rationalist learning mechanisms. We could give numerous examples of rationalists arguing against empiricists in similar fashion—in this case, claiming that concrete proposed empiricist models don't fare well compared to models that make use of innate representations or characteristically rationalist learning mechanisms. We won't belabour the point any further, however, as we will see many examples like this later in the book.

Finally, there is a third reason for interpreting the rationalism-empiricism debate as we do. This is simply the observation that it makes sense of the fact that the rationalism-empiricism debate has been so productive. Consider the case of the study of language acquisition and understanding. Chomsky's work in linguistics was not motivated by a view about the relative importance of genes versus the environment in the acquisition of language. Rather, it was directly motivated by the perceived inadequacy of existing domain-general theories of language acquisition and their accompanying theories of language. This work, and the subsequent rationalist research programme it led to, generated a wealth of new data and theoretical insights and inspired much further theorizing that made use of

innate representations and characteristically rationalist learning mechanisms. It also inspired others to try to find ways to resist these rationalist developments and the rich rationalist acquisition base that they entailed, leading to the discovery of further invaluable data and the development of new and important contrasting empiricist perspectives (e.g., giving rise to significant advances in connectionist modelling and statistical learning theory). All observers should agree that these competing perspectives have enormously benefited the study of language, with each side continually forcing opposing theorists to refine and elaborate their views about the character of the acquisition base and its relation to the acquisition of traits to deal with new findings and ever more sophisticated theories.

Similar dynamics have played out in almost every area of cognitive science and in relation to an increasingly wide assortment of cognitive capacities—from face perception and object representation to social reasoning and moral motivation. These advances have not been driven by general claims about the importance of genes versus the environment. Instead, in each case, researchers have been guided by considerations bearing directly on the character of the innate psychological structures ultimately responsible for the origins of psychological traits (i.e., the acquisition base)—precisely in line with our understanding of the rationalism-empiricism debate. And they have responded to each other's findings and proposed psychological models with counterproposals that have helped all involved to attain a better understanding of the cognitive capacities at issue. It would hardly be an overstatement to say that the rationalism-empiricism debate has been the driving motivation behind some of the most ingenious and enduring theorizing in cognitive science.

3.3 Conclusion

Many contemporary theorists hold that the rationalism-empiricism debate is fundamentally flawed and that it should be abandoned altogether. They interpret this debate as a debate about nature versus nature and argue that the nature-nurture debate is riddled with confusion. We have seen, however, that the charge of confusion is misplaced and that the arguments against the rationalism-empiricism debate don't undermine the debate once it is understood in the way that we have suggested it should be, as a debate about the character of the acquisition base. If anything, their arguments bring out many of the advantages of our interpretation of the rationalism-empiricism debate since, unlike views that equate it with a debate about nature and nurture, the interpretation in terms of the acquisition base is immune to the difficulties these arguments raise. We have seen as well that further support for our account comes from the way that participants in the rationalism-empiricism debate understand their own positions, from how they argue with one another when evaluating specific claims and findings, and from

the fact that understanding the debate in these terms explains the enormous fruitfulness of the debate across the cognitive sciences. The rationalism-empiricism debate is not fundamentally flawed. On the contrary, not only shouldn't it be abandoned, it should take centre stage in the vast ongoing interdisciplinary project of trying to understand how the mind works.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0003

4
The Viability of Rationalism

Chapters 2 and 3 focused on what is at stake in the rationalism-empiricism debate, presenting a detailed account of what we take the debate to be about and addressing a variety of challenges that have been taken to undermine it. In this chapter, we turn to rationalism itself. Rationalism is widely regarded to be an antiquated and profoundly flawed view that can be safely discarded. Critics have claimed that it is unscientific and theoretically lazy (in avoiding the real work of explaining where psychological capacities come from), overly intellectualist (for positing too many complex psychological processes), and excessively speculative (for its reliance on evolutionary "just-so" stories)—and that such failings warrant simply rejecting out of hand the entire rationalist framework (i.e., rationalism in general) as a way of addressing the origins of psychological traits in any type of rationalism-empiricism debate, whether local or global. According to such critics, rationalist accounts are simply non-starters across the board. We disagree. Despite these and other charges aimed at undermining all forms of rationalism, we think that contemporary rationalism is a robust and powerful explanatory theoretical framework and that rationalist accounts cannot simply be dismissed in this way. However, we are cognizant of just how widespread these anti-rationalist sentiments are in philosophy and cognitive science.[1] For this reason, before we can begin to look at the more specific issues that arise for rationalist accounts of the origins of concepts, we need to say something about the viability of rationalism in general. This chapter will argue that none of the theoretical challenges purporting to undermine rationalism in this general way are successful. As a result, the merits of rationalist accounts of the origins of any given type of psychological trait must be evaluated through a detailed consideration of the arguments and evidence for and against such accounts.

4.1 A Preliminary Case for Rationalism

Much of this chapter will be devoted to critically examining a range of objections that aim to show that rationalism is essentially a non-starter—that rationalism is

[1] Rationalism has long been out of favour in philosophy. It is noteworthy, for example, that virtually every major philosopher of the twentieth century who had anything at all to say about psychological and social phenomena sided with empiricism or adopted theories that are plainly anti-rationalist. This includes not just philosophers in the analytic tradition (Russell, Carnap, Ayer, Wittgenstein, Goodman, Quine, Putnam), but also luminaries in the continental tradition (Heidegger), phenomenology (Merleau-Ponty), and postmodernism (Foucault).

so deeply flawed that it is not a viable theoretical option. Before we get to those objections, however, it will be helpful to have some sense of what motivates rationalism and why it is a position that is at least worthy of detailed critical attention. In making a preliminary case for rationalism, we should be clear that we are not arguing for rationalist accounts across the board for every type of psychological trait. We think that rationalist accounts are highly plausible for a number of important psychological traits. But empiricist learning mechanisms are also important and play a role in the acquisition of many psychological traits too. As should be clear from Chapter 2, rationalism is perfectly compatible with the existence of domain-general learning. It should also be kept in mind that no rationalist accepts every rationalist account, just as no empiricist accepts every empiricist account. Rationalism and empiricism are broad theoretical frameworks, and for any psychological trait there will be many possible rationalist accounts of the origin of that trait and many possible empiricist accounts of the origin of that trait. And of course, since these accounts are in competition with one another, most of them, both rationalist and empiricist, will turn out to be mistaken. And as noted in Chapter 2, global rationalism is perfectly compatible with local empiricism. With these caveats out of the way, we can begin by taking a brief look at two of the most important arguments that rationalists have employed in arguing for rationalist accounts of the origins of a range of psychological traits: the *poverty of the stimulus argument* and the *argument from animals*.

The poverty of the stimulus argument is undoubtedly the most famous and widely cited argument for rationalism. It is important to keep in mind, however, that it isn't so much a single argument but a family of arguments. What binds them together is the observation that, given only general-purpose learning mechanisms, the information in a learner's environment is inadequate to account for the fact that some type of psychological trait is reliably acquired by learners. Since the input is inadequate, the difference has to be made up somewhere. According to the poverty of the stimulus argument, what is required is an acquisition base for learning that is richer than empiricists suppose—a set of innate psychological structures that isn't restricted to domain-general learning mechanisms.

Empiricists have been highly critical of this form of argument. Some of these criticisms have been based on empirical claims. It is said that proponents of the poverty of the stimulus argument haven't provided enough evidence to establish that the environment is as impoverished as rationalists claim it is (e.g., Putnam 1967; Cowie 1999; Pullum and Scholz 2002). Or considerations are offered that are meant to suggest that general-purpose learning mechanisms might be capable of accomplishing the learning task after all, for instance by raising the possibility that rationalists have underestimated the power of statistical methods (e.g., Lewis and Elman 2001; Prinz 2002).

Philosophical critics of the poverty of the stimulus argument have also put forward more principled, theoretical objections. One of these is to insist that learning routinely allows us to go beyond the stimulus (i.e., beyond the evidence

supplied by experience) and that such learning is undoubtedly accomplished by general-purpose learning mechanisms of the sort that empiricists endorse. For example, in learning what a curry is or about a given style of art, we can readily determine that something is or is not in the extension of the relevant concept even if we haven't encountered it before and haven't been given explicit evidence about its status. Yet to require a special-purpose mechanism for learning about vindaloo or impressionist painting is absurd (Cowie 1999; see also Goodman 1967, Sampson 2005). More generally, it is evident that paradigmatic instances of general-purpose learning, such as inductive inference, are in the business of using limited, finite data to draw conclusions about a vastly larger range of cases. But if we don't need a specialized rationalist learning mechanism for deducing that all swans are white, why do we need a specialized system for acquiring the rules of English syntax?

All of these objections are based on what we take to be oversimplified accounts of the poverty of the stimulus argument.[2] One of the main problems is the failure to recognize that poverty of the stimulus arguments specifically contrast the outcome of general-purpose learning mechanisms with the outcome of the more specialized learning mechanisms posited by rationalist theories. Poverty of stimulus arguments don't merely argue that what is learned goes beyond the input to learning in some way; rather, they argue that what is learned goes beyond the input to learning *in a way that purely general-purpose learning mechanisms cannot account for*. So it is no objection to proponents of the poverty of the stimulus argument that induction goes beyond the evidence of experience.[3]

One especially vivid type of poverty of the stimulus argument draws upon the results of what are known as isolation (or deprivation) experiments. These are empirical studies in which, by design, the experimental participants are removed from all stimuli that are related to a normally acquired trait. For example, Irenäus Eibl-Eibesfeldt showed that squirrels raised in isolation from other squirrels, and without any solid objects to handle, spontaneously engage in the stereotypical squirrel digging and burying behaviour when eventually given nuts. Eibl-Eibesfeldt notes that,

[2] See Laurence and Margolis (2001) for an extended discussion of the logic of poverty of the stimulus arguments and for theoretical and empirical support for the applicability of such arguments to the acquisition of natural language.

[3] This point explains why some poverty of the stimulus arguments from the history of philosophy are less compelling than others. For example, no contemporary theorist would be moved by Descartes' argument (mentioned in Chapter 1) in which he claims to show that ideas are innate simply by pointing out that these ideas aren't found in the sense organs. This is because it does not adequately argue for domain-specific learning mechanisms over domain-general ones that also go beyond the input in learning. Descartes' argument is therefore ineffective as a critique of empiricist models of concept acquisition even if it is effective as a critique of scholastic Aristotelean views of perception in which a form is supposed to be literally transmitted from an object to a sense organ.

> the stereotypy of the movement becomes particularly obvious in captivity, where inexperienced animals will try to dig a hole in the solid floor of a room, where no hole can be dug. They perform all the movements already described, covering and patting the nut, even though there is no earth available. (Eibl-Eibesfeldt 1989, p. 21)

Since the squirrels were kept apart from their conspecifics, they had no exposure to this stereotypical behaviour prior to exhibiting it themselves—the stimulus was about as impoverished as can be. Reliable acquisition of such a complex and idiosyncratic behaviour under these circumstances provides extremely good evidence against the view that in such cases acquisition is mediated by a general-purpose learning mechanism.

We should stress, however, that a successful poverty of the stimulus argument doesn't require anything as stark as removing learners from their normal environment. Consider an example from the study of language acquisition (Crain and Thorton 1998; Crain and Pietroski 2001). It turns out that English-speaking children sometimes go through a peculiar stage as they learn how to form certain types of questions. They insert an extra wh-word (e.g., "what" or "who"), saying things like:

(1) What do you think what Cookie Monster eats?
(2) Who did he say who is in the box?

These sentences are, of course, ungrammatical in adult English. But it's not as if these children randomly insert extra wh-words any which way. On the contrary, their speech exhibits a systematic and predictable pattern in which they only place an extra wh-word in a specific location in a question, they don't insert extra wh-*phrases* in these locations, and they don't place an extra wh-word in constructions involving subordinate clauses with infinitive verbs. So the following are ungrammatical both to adults *and* to children passing through this phase:

(3) *What do you what think is in the cupboard?
(4) *Which Smurf do you think which Smurf is wearing roller skates?
(5) *Who do you want who to win?

What's more, though adult English speakers don't say any of (1)–(5), the pattern found in their children's speech does appear in other natural languages, including German, Irish, and Chamorro. For example, in German an extra wh-word is used, but not a wh-phrase, making (6) grammatical but (7) ungrammatical:

(6) Wer_i glaubst du wer_i nach Hause geht?
Who do you think who goes home?

(7) *Wessen Buch$_i$ glaubst du wessen Buch$_i$ Hans liest?
Whose book do you think whose book Hans is reading?

Since the English-speaking children who ask questions with extra wh-words grow up in perfectly normal linguistic environments, this case isn't like the isolation experiments we have just considered—they have been exposed to many English sentences. Nevertheless, it satisfies all the conditions necessary for a successful poverty of the stimulus argument. In particular, two facts make it highly unlikely that the learning process underlying these linguistic patterns is based solely on general-purpose learning mechanisms, such as ones that rely on domain-general processes of statistical analysis. First, these English-speaking children aren't emulating a linguistic pattern that is present in their linguistic environment. After all, adult English speakers don't say things like (1) and (2) any more than they say things like (3)–(5). So it is highly unlikely that these children are picking up on a pattern that is data driven. Second, since children's selective use of extra wh-words conforms to the (grammatical) pattern for using extra wh-words in other natural languages—languages these children *haven't* been exposed to—it is also highly unlikely that it amounts to a random deviation of the sort that would be expected of a general-purpose statistical analysis of a noisy signal. Rather, it appears to be a principled deviation that reveals certain assumptions that constrain the learning process. The most natural explanation of these facts is that principles specific to language are part of the acquisition base, as rationalists generally maintain. On this approach, in learning language, children are working with a highly constrained hypothesis space, and this leads some children to temporarily adopt a set of syntactic rules that are laid out in this space even if these rules aren't attested to in the data.

The examples so far have been of poverty of the stimulus arguments in which it is patent that the stimulus is impoverished. But poverty of the stimulus arguments don't always work in such a blatant manner, where an acquired behaviour fails to appear in the learning environment altogether. For example, Chomsky and others have argued for a rationalist treatment of the principles governing natural languages even though these principles are instantiated in countless utterances that are available to language learners. The key point once again is that the sense in which the stimulus is impoverished is *relative to general-purpose learning mechanisms*. The idea is that such mechanisms cannot reliably produce the learned outcome on the basis of the utterances that children hear.

One reason why they can't do this is that the correct hypotheses are not at all the most natural ones for a learner without domain-specific learning biases employing only empiricist learning strategies (Laurence and Margolis 2001). Indeed, there are numerous alternatives that would be more natural to such a learner but that would lead the learner astray. A related major problem for empiricist models of language learning is that children don't just need there to be

suitable examples of a linguistic principle present in their environment. They also need to be able to represent the examples in the right way (i.e., in terms of their grammatically relevant properties) to ensure that the utterances they hear can serve as meaningful data for them (Laurence and Margolis 2001). A poverty of stimulus argument of this kind doesn't turn on the absence of a form to be learned. It relies instead on the reasoning that learners without domain-specific learning biases who are granted only general-purpose learning mechanisms would lack the materials that would even allow them to entertain the evidence that points to the linguistic principles to be acquired. Relevant data may be present but inaccessible to learners if learners do not have the cognitive resources to see the data for what it is.

The poverty of the stimulus argument in its different forms offers a forceful reason to posit rationalist learning mechanisms regarding a diverse range of psychological traits. But the case for a rationalist approach to the mind doesn't end with the poverty of the stimulus argument. Given the characterization of the rationalism-empiricism debate that we argued for in Chapter 2—focusing on the disagreement between rationalists and empiricists regarding the acquisition base—it should be clear that adjudicating between the two, in the most general terms, takes the form of an argument to the best explanation. So a wide variety of explanatory factors may come into play. Any type of evidence at all that increases the likelihood that the acquisition base contains a significant number of characteristically rationalist psychological structures that contribute to rationalist learning mechanisms would count in favour of rationalism.

One noteworthy implication of this fact—that rationalism is to be argued for on explanatory grounds—is often lost in philosophical discussions. It is that psychological traits can be rooted in special-purpose learning mechanisms even if the stimulus is *not* impoverished, that is, even if there is no poverty of stimulus argument to be had. Indeed, this situation may be fairly common. Consider again the nut burying behaviour of squirrels. Though there is an especially compelling poverty of the stimulus argument regarding its acquisition under experimentally controlled conditions of isolation, squirrels rarely find themselves in this situation. Their normal environment isn't so impoverished, as other squirrels' nut burying behaviour is readily observable. So squirrels might acquire knowledge of this behaviour through observation. Nonetheless, it looks as though they are equipped with a special-purpose learning mechanism that can operate in the absence of the sorts of evidence a general-purpose learning mechanism would require.

Why might there be rationalist special-purpose learning mechanisms even when the experiences relevant to an acquired trait aren't particularly impoverished? There are a number of reasons. One is that the trait may be important enough that it can't be left to a less reliable means of acquisition. In fact, sometimes when a trait is important enough, there may even be multiple independent

specialized mechanisms involved. This appears to be the case for chicks in how they are able to come to identify their mother. They have one mechanism for detecting a large moving object and another that relies on a shape template (Johnson et al. 1985; Carey 2009).[4] A second reason for the possession of a specialized learning mechanism in the absence of an impoverished environment is the potential benefit of acquiring a trait rapidly. This is the likely explanation of infants' instinctive avoidance of visual cliffs. It would be easy enough to learn about the danger of cliffs through experience, but this way of doing things would be much slower—and rather more dangerous. Yet another reason why a rationalist learning mechanism might be involved in cognitive development even if the environment isn't impoverished is the cognitive cost of using a domain-general learning mechanism. For example, less cognitive effort is needed, on average, for a learner to acquire a principle based on a forced choice of two options (as in linguistic parameter setting) than having to choose from an immense space of possibilities (which a truly general-purpose learning mechanism with no domain-specific biases would face).

Although the poverty of the stimulus argument is the most famous argument for rationalism, it is just one part of a much larger set of explanatory considerations that argue for rationalism.[5] The second major argument for rationalism that we want to discuss here—the *argument from animals*—further strengthens the rationalist's case by placing the rationalism-empiricism debate in a broader context. The argument from animals is grounded in the fact that animals have a plethora of special-purpose learning mechanisms. Some of these are shared across species (even widely shared across distantly related species), while others are unique to a given species. But *human beings are animals too*. Hence, the argument concludes, we should also expect some of the ancient and widely shared mechanisms to populate the human mind, and we should also expect the human mind to have other less widely shared mechanisms, including some special-purpose learning mechanisms of its own—ones that are geared towards our own particular needs as animals.[6]

The number of examples of special-purpose learning mechanisms in the animal kingdom is large and impressive. It includes learning mechanisms associated with the ability to communicate with conspecifics, to identify which items to treat as food and which items to avoid eating, to select the location of a defended territory or a home site, to identify and respond to predators and aggressive animals,

[4] Carey (2009) uses this example to illustrate that it isn't problematic to attribute special-purpose learning mechanisms to animals. We wholeheartedly agree. See the *argument from animals* below and in Chapter 10.

[5] We limit ourselves to two types of arguments here in discussing rationalism as a general approach to the origins of psychological traits. We will be discussing a broader range of arguments for rationalist accounts when we turn to our positive case for rationalism regarding the origins of concepts in Part II.

[6] For related arguments, see Gallistel et al. (1991) and Carruthers (2006).

to identify potential mates and make effective mating decisions, to identify and respond selectively and appropriately to kin, to establish and maintain knowledge of the dominance relations in one's group, to determine the time of day that an event occurs, to determine the temporal interval separating a series of events, to estimate the numerical quantity of a perceived group or collection, to determine the numerical order in a series (e.g., in a series of landmarks), to determine which among alternative foraging locations produces the higher rate of return, to learn the distance and direction of one's home location or a point of origin, and much else.[7]

However, it is easy to overlook the specialized nature of the learning mechanisms that go with these abilities if you don't attend to the subtle behavioural details in which they are expressed. For instance, one might suppose that learning to avoid poisonous food is simply a matter of associating a negative event (such as becoming ill) to the sensory event that precedes it. The quicker the onset of the negative event, the easier it is to learn to avoid food of that type. But in fact, food aversions are not acquired via a general-purpose learning mechanism for associating sensory events with negative events that closely follow them. Rats that are rendered ill after ingesting flavoured water learn to avoid that taste, ignoring equally highly correlated visual or auditory cues, and the association can be effectively formed after a long latency period—on the order of hours, not seconds. In contrast, rats that are punished via a shock can learn the visual and auditory cues, but the punishment does have to occur within seconds of ingesting the water (Garcia and Koelling 1966; Revusky and Garcia 1970). What's more, the predisposition to link illness with taste rather than with visual cues isn't universal. Bobwhite quail, whose food choices rely a great deal on vision, favour colour (Wilcoxon et al. 1971), and Vampire bats, who are monophagous feeders (they just eat blood), don't form taste aversions at all, even though closely related bat species that are food generalists behave much as rats do (Ratcliffe et al. 2003). It would be easy to miss these patterns regarding the acquisition of food aversions. But once one attends to these patterns, it is clear that the learning processes that underlie food aversions are distinct from those that subserve arbitrary learned associations and that they are subject to different predispositions according to the feeding strategy of a species.[8]

[7] For overviews of related empirical work, see Gallistel (1990); Olmstead and Kuhlmeier (2015); and Bueno-Guerra and Amici (2018).

[8] In a defence of the role of associative learning in animal cognition, Heyes (2012) uses the example of learned food aversions to illustrate the power of general-purpose learning and the danger of what she refers to as "association-blindness" or "the failure to consider associative learning as a candidate explanation for complex behaviour" (p. 2695). While we agree with Heyes that associative learning hypotheses should be considered as candidate explanations for behaviours, we don't share her view that cognitive science as a field is guilty of association-blindness. In fact, food aversions, in our view, argue for precisely the opposite moral: researchers face the danger of seeing association everywhere, and of being all too willing to offer associative accounts of complex behaviour when the full range of data shows that they are inappropriate.

Consider another example of a specialized learning mechanism that might easily be overlooked unless one attends to subtle behavioural details. After meandering in search of food, desert ants can find their way home by following a straight line back to their nest. Incredibly, the ants are able to do this, in spite of the fact that their environment is relatively featureless (and so lacking in landmark cues), because of a mechanism for *path integration* (also known as *dead reckoning*), that is, one that keeps track of the cumulative changes in their direction of movement and the distance covered. Determining the position of the nest is a matter of combining this information periodically, just as sailors compute their position by noting each change in direction and the speed and duration of the ensuing segment of their journey (using speed and duration to compute distance).

The postulation of a special-purpose learning mechanism for path integration may at first seem unlikely for an animal as primitive as an ant. Couldn't ants be using a simpler means, such as a scent emanating from the nest, or a chemical signal left on the trail? In principle they could be; however, these alternatives have been tested empirically and found not to work. Experiments reveal that if a desert ant is displaced from a food source on a foraging run, it will attempt to return to its nest by following a direct path in the direction that would have led back to its nest had there been no intervention. After travelling in a straight line the corresponding distance home, it enters into a characteristic search pattern for where its nest ought to have been (Wehner and Srinivasan 1981; Gallistel 1990). Likewise, if an ant's legs are made longer or shorter (by the addition of stilts or through partial amputation), it will systematically overshoot or undershoot the distance to its nest (Wittlinger et al. 2006). Ants are simple creatures, but packed into their tiny brains is a mechanism for path integration (just one of the many types of specialized learning mechanisms that insects use for navigation). Once again, without carefully attending to the behavioural details, it would be easy to posit a broadly empiricist account of these abilities. But on closer examination, the behaviours turn out to be far more complex than one might initially have imagined, and the learning mechanisms underpinning these behaviours need to do justice to this complexity.

In our initial presentation of the argument from animals, we put the argument in terms of two equally important claims. The first is that special-purpose learning mechanisms are a characteristic feature of animal minds. The second is that humans are animals. Now no one would deny the second claim. But some may be reluctant to draw any inferences from our being animals, since human beings may be very different kinds of animals than non-human animals. Perhaps it is not so surprising that non-human animals have special-purpose mechanisms that support their limited forms of behaviour, but why think that applies to *us*? After all, human behaviour is strikingly flexible and open-ended compared to the behaviour of all other extant animals. So there may be doubts about whether it would be warranted to draw any conclusions about the human mind from facts

about the minds of insects and rodents, or even from closer relatives, such as monkeys and chimpanzees (Brown 2019). Or to put much the same point in terms of the evolution of the human mind, it is certainly true that traits are sometimes retained through the course of evolution, but sometimes they are lost as well (see, e.g., Wilson 2008). Maybe what happened in the case of the human lineage was the disappearance of ancient special-purpose learning mechanisms and, in their place, one or more powerful general-purpose learning mechanisms emerged, effectively reconfiguring the organization of the human mind.

This is a possibility, to be sure. Humans are of course a remarkable species. And there should be no doubt that general-purpose learning mechanisms are a significant feature of the human mind. Nonetheless, many of the special-purpose learning mechanisms that appear in other animals would have continued to be of enormous value to the extent that our ancestors faced the very same problems these mechanisms evolved to address and that they address with remarkable efficiency and reliability. Thus it is reasonable to suppose that these wouldn't be replaced wholesale even if some of their functions could be handled in other ways by additional general-purpose learning mechanisms (e.g., the learning of a technology that is dependent on cultural transmission).

Take navigation. There are many ways of navigating in both familiar and unfamiliar environments, and it turns out that evolution has produced a variety of special-purpose systems to this end. We have just seen that path integration is one of these. As long as an animal can keep track of the distance traversed for each directional change, it is possible to compute a direct path to its starting position. This makes path integration particularly valuable in situations where landmark information isn't available or easy to access. Notice, though, that this isn't a circumstance that is peculiar to locating a nest in the middle of a featureless desert. Path integration is potentially valuable in any type of terrain where landmarks can't readily be exploited—because of recent changes to the landscape, because an animal is new to the area and hasn't had the opportunity to note its salient stable features, or even for the simple reason that the environment's features are perceptually inaccessible (e.g., it is too dark for them to be seen). Path integration may also have a role to play even when an animal *is* in a position to identify landmarks in that it can support learning about landmark information (Müller and Wehner 2010). And it is common for animals to possess a number of redundant and overlapping systems for navigation. It shouldn't be all that surprising, then, that a wide range of animal species that inhabit environments very different from featureless deserts make use of path integration in navigation—for example, cockroaches, crabs, geese, hamsters, naked mole rats, dogs, and primates (including humans).[9]

[9] See Gallistel (1990) and Etienne et al. (1998) for path integration in animals, and Loomis et al. (1999) and Smith et al. (2013) for path integration in humans. Wolbers et al. (2007) show that the cortical system used in path integration in humans is similar to that used by rodents and non-human primates.

Another navigational technique, which is associated with a distinct special-purpose learning mechanism, is one that we saw in Chapter 1. It keeps track of the geometrical properties of an area's boundaries and surfaces to assist an animal in finding its way after becoming disoriented. This too is especially helpful in cases where landmark information isn't available, can't be used, or wouldn't be expected to be reliable. Like path integration, there are many situations in which a mechanism of this kind would be valuable if not essential, situations that are hardly peculiar to non-human animals. As we noted in Chapter 1, this mechanism, too, is widespread among animals and is present in humans.

Other ways of navigating make use of landmarks and feature information. For example, these can be used as a beacon that an animal can home in on or use to fix a compass bearing. More interestingly, landmarks can be plotted in a representation that also includes the distances and geocentric relationships among differing locations in an area, essentially creating a cognitive map of the environment. Cognitive maps are an especially powerful navigational tool because they allow an animal to infer a novel and efficient path between previously encountered locations, as when a bee flies directly from one flower patch to another even if previously it had only visited those sites on independent excursions from its hive (Gould and Gould 1995; Cheeseman et al. 2014). Like path integration and geometry-based reorientation, the use of landmarks in navigation has also been demonstrated in a wide variety of species, from invertebrates to birds and mammals. Given how widespread all of these navigational resources are among animals, the widespread presence of multiple and redundant systems of navigation in animals, and the obvious selection pressures on the formation and maintenance of navigational abilities, it would actually be rather surprising, we think, to discover that humans alone lack these sorts of mechanisms.

This isn't to say there aren't differences of detail. For example, one would also expect that animals' navigational abilities and proclivities would be influenced by particular features of the environment in which they live and forage, by their patterns of migration, their breeding schedule and mating strategies, their lifespan, and so on. They should also differ according to a species' sensory systems and mechanisms of locomotion. For example, a compass bearing can be fixed in various different ways including by visually locating the sun (taking into account both the time of day and the time of year), by responding to the polarization of the sun's light (on cloudy days and close to dusk or dawn), by detecting changes in the earth's magnetic field, and by locating the Pole point in the night-time sky. But the claim that humans share some of the same basic navigational systems as other animals doesn't mean that they share all navigational systems that other animals possess—something not true of other animals either—or that they rely upon them to the same extent, in exactly the same situations, or that the sensory and perceptual input that they draw upon has to be the same.

A lot more could be said on behalf of the poverty of the stimulus argument and the argument from animals. When we turn to the case for concept nativism (in Part II), we will discuss a number of related arguments that pick up on similar themes. For the moment, however, we hope that this brief sketch indicates some of the motivations for adopting a rationalist view of the mind in general and why, at the very least, rationalism is a position worthy of detailed critical attention. The next step is to examine some of the influential objections to rationalism that are often cited by its critics.

4.2 Objections and Replies

Many theorists—philosophers and cognitive scientists—hold that rationalism as a perspective on the origins of psychological traits of any sort is not merely mistaken, but fundamentally and fatally flawed. The feeling is that there are a number of general objections to rationalism that are so basic and so forceful that rationalism can be rejected without having to take a hard look at the evidence that might be marshalled in its favour. If rationalism's prospects are as dim as these critics maintain, then the project of this book—a full-scale defence of concept nativism—would be a pointless endeavour. For this reason, it is essential for us to address these general objections and to show why they don't work. Addressing them will also help to further clarify the rationalist framework and the rationalism-empiricism debate.

Rationalism is lazy science. One common complaint against rationalism is that it exhibits a kind of theoretical laziness for merely postulating complex innate structures, or an excessively rich acquisition base, rather than taking up the challenging task of truly explaining where psychological traits come from. For example, Churchland (2012) characterizes the rationalist tradition as holding that "since one has no idea how to explain the origin of our concepts, one simply pronounces them innate" (p. 15). In a similar vein, Karmiloff-Smith (2009b) praises Piaget for stressing "that nativism was a theoretical cop-out" (p. 99).

In our view, this type of criticism of rationalism would be valid if rationalists blindly and uncritically posited innate representations and rationalist learning mechanisms to explain the acquisition of cognitive traits. But they don't. In all of the examples we have discussed so far—and in numerous examples that we will encounter in later chapters—rationalists are clearly engaging in a substantial explanatory enterprise. They aren't unthinkingly taking psychological structures to be innate.

Notice as well that a similar charge of lazy science would extend to *empiricists* who blindly and uncritically suppose that most cognitive traits are acquired by a small number of domain-general learning mechanisms, that is, empiricists who "simply pronounce" that concepts or other psychological structures are acquired by domain-general learning mechanisms without truly explaining how they are

acquired. Clearly, then, if there is a problem here, it isn't with rationalism per se but with any cursory assertions about where concepts and other psychological structures come from.

Rationalists and empiricists have the same burden of showing that their view provides the best explanation of the acquisition of given psychological traits. To put this in the terms laid out in Chapter 2, rationalists and empiricists both need to show why their view of the acquisition base should be accepted. Good rationalists (like good empiricists) take this responsibility quite seriously.

Finally, as also noted in Chapter 2, rationalism isn't primarily about identifying innate concepts or fully formed innate psychological mechanisms. Much of what rationalism is about is explaining how concepts, mechanisms, and other types of psychological traits are the product of rationalist learning mechanisms and psychological development, where the rationalist learning mechanisms involved are often themselves the product of learning and development, tracing back to a mix of characteristically empiricist and characteristically rationalist psychological structures in the acquisition base. These rationalist learning mechanisms are accepted precisely because they can explain aspects of development that are mysterious on empiricist theories (or, more generally, because they otherwise provide a better overall account of the origins of certain psychological traits than competing empiricist accounts do). Such work hardly amounts to simple pronouncements. It requires painstaking experimentation. All this objection shows, then, is that rationalists and empiricists alike must provide arguments and evidence to support their claims, a point that was never in dispute.

Rationalism is unparsimonious. Another common objection to rationalism is that it is unparsimonious. Many take it to be simply obvious that empiricism is more parsimonious than rationalism and hold that this has the consequence that empiricism should be regarded as the default view in the rationalism-empiricism debate. It is also often assumed that empiricism's default status means that all empiricist alternatives need to be ruled out before rationalism can even be considered a live option. For example, Prinz (2012) says "Empiricism is a more economical theory. There is no reason to postulate innate machinery without powerful evidence. Thus, Empiricism should be the default position until evidence weighs in favour of Rationalism" (p. 90). Prinz further claims that rationalists "need to show that the kind of knowledge they attribute to infants is of a type that would be impossible to learn by observation" (2012, p. 110). In much the same spirit, Haith (1998) writes that rationalists who "pursue high-level cognitive constructs [in explaining developmental data] must play the default game. That is, one must fend off every possible perceptual interpretation of differences to entertain default cognitive interpretations," adding that "even when an immediate perceptual explanation is not obvious, there is the danger that one will come along" (p. 170).

This objection raises a number of important issues, and we will discuss it in greater detail in Chapter 17. Here we will focus on just two questions: (1) Is empiricism truly more parsimonious than rationalism? (2) Supposing it is, does this secure empiricism's default status and give us reason to reject rationalism without further ado? It turns out that both of these issues are more complicated than they might first appear.

Regarding the first, there is certainly one sense in which empiricism is more parsimonious than rationalism: empiricism posits fewer (and fewer kinds of) innate psychological structures in the acquisition base (see Chapter 2). However, comparing numbers of innate psychological structures isn't the *only* way to measure the relative parsimony of a psychological theory. In fact, this measure captures just a single facet of the ontogenetic process—the psychological structures that psychological accounts of development begin with. But as Spelke (1998) points out, development is not simply a matter of what's in the acquisition base; we need to attend to the *full developmental process* and ask which account of this full process is more parsimonious (see Figure 4.1).

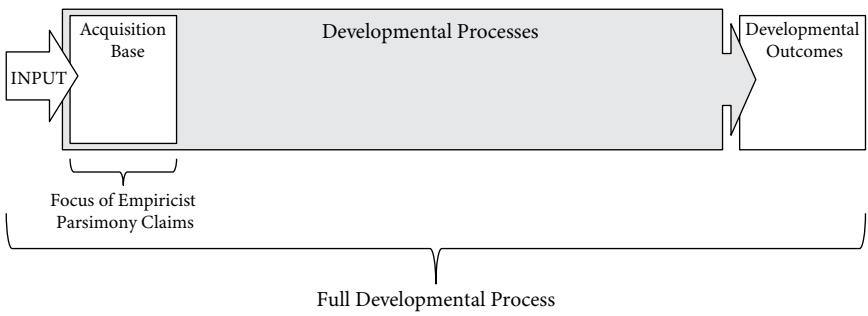

Figure 4.1 Parsimony in cognitive development. When evaluating a theory of development for how parsimonious it is, empiricists typically focus exclusively on the psychological traits that are claimed to be in the acquisition base. But the full developmental process needs to be considered.

For example, we need to compare proposed learning mechanisms in terms of the quantity of input they require to achieve a learning outcome. An account may be more parsimonious to the extent its learning mechanisms require less input, a dimension of parsimony we might call *data set parsimony*. This opens up the possibility that rationalism could be more parsimonious in some cases, since rationalist theories posit domain-specific learning mechanisms that build in assumptions about the domains in which they operate. As a result, they tightly constrain the space of options that a learner has to consider and, if successful, promise to get by with a learner having to make a smaller number of observations. Notice that this may well reduce demands on a learner's memory and attention. This, in turn,

means that a rationalist account might be more parsimonious in another way. It might be more *computationally parsimonious*, requiring fewer computations in order to achieve a learning outcome than competing empiricist accounts. What's more, parsimony in both of these senses might have the further effect of making a rationalist account more metabolically efficient—fewer observations and a quicker outcome means fewer expended calories—so that a rationalist account might be more *energetically parsimonious* too. Also relevant are considerations having to do with phylogenetic development, or development on an evolutionary timescale. The argument from animals suggests that rationalist theories, in some cases, require fewer modifications leading to particular psychological traits, as certain acquisition systems may be evolutionarily ancient (e.g., acquisition systems related to spatial representation and navigation). In this situation, a rationalist account could turn out to be more *phylogenetically parsimonious* (see Fitzpatrick 2008 for related discussion).

Where does this leave us with the question of whether empiricism is more parsimonious than rationalism? If considerations having to do with parsimony are going to be of any use, then we shouldn't focus on just a single facet of development (numbers of innate traits, the quantity of data required, etc.); we need to consider the full developmental process. When we take this broader perspective, there will inevitably be trade-offs between different types of parsimony. One theory might be more parsimonious in terms of the number of innate learning mechanisms it postulates but at the expense of making its learning processes computationally unparsimonious. Another might be more parsimonious in terms of the quantity of data its learning mechanisms require but at the expense of postulating learning mechanisms that are phylogenetically unparsimonious. And so on. Since it is hardly obvious how to adjudicate between such trade-offs, there simply is no way to pre-empt having to actually compare competing empiricist and rationalist theories. Considerations of parsimony are no reason for dismissing rationalism out of hand and forgoing a serious consideration of the arguments and evidence bearing on particular rationalist and empiricist accounts.

Still, suppose for the sake of argument that empiricists' theories, in some instances, are more parsimonious than rationalist theories according to some agreed upon ideal regarding these trade-offs. Our second question is whether this fact on its own would constitute good grounds for dismissing rationalism. One might think so, since parsimony is widely considered a significant factor in the evaluation of scientific theories. It turns out that even under these circumstances, parsimony cannot be applied in this way.

One complication is that rationalism and empiricism are not individual theories, but rather large-scale frameworks for explaining the development of psychological traits. As a result, it is not clear what it would even mean to say that one of these frameworks is simpler than the other. Would *every* empiricist theory have to be more parsimonious than every rationalist theory? (But why couldn't there

be at least some empiricist theories that were less parsimonious than at least some rationalist theories?) Would it be enough if *most* empiricist theories were more parsimonious than most rationalist theories? (But then why prefer the empiricist theories over their more parsimonious rationalist counterparts in the minority of cases where the rationalist theory is more parsimonious?)

Even if there were a principled way of answering these questions, there is a much deeper problem with the idea that empiricist theories being more parsimonious than rationalist ones (according to some agreed upon ideal of parsimony) would provide sufficient grounds to accept empiricist theories over rationalist ones. This is that parsimony considerations are standardly taken to apply only in cases where the theories being compared are otherwise equally good at explaining the existing evidence. And there is a good reason why this is so. After all, it is not hard to find a more ontologically parsimonious theory to replace almost any current theory if the alternative theory needn't be equally good at capturing existing data. This means that parsimony isn't relevant until competing explanations prove to be equals in other respects (this is essentially the moral of the argument from Chomsky discussed in Chapter 1). Parsimony is not the sort of consideration that *can* be used to establish that one theory (or theoretical framework) wins by default; rather, we would need to look at the relative explanatory merits of competing empiricist and rationalist theories first.

So even if some empiricist theories were more parsimonious than competing rationalist theories in some agreed upon sense, and even if the more parsimonious approach should be accepted (other things being equal), this would still give us no reason at all to dismiss rationalism prior to a detailed examination of the evidence to which empiricist and rationalist theories are accountable.

Rationalism is overly intellectualist. Another charge that is often raised against rationalism is that rationalism overly intellectualizes the mind. This objection highlights the fact that the rationalism-empiricism debate is linked to a broader set of issues beyond questions of development. Rationalists and empiricists do not just differ in how they view psychological development, but also in how they view the *mature* mind. In particular, rationalists often take mature cognitive and behavioural capacities to require complex, sophisticated cognitive mechanisms, where empiricists take such capacities to involve relatively simpler cognitive mechanisms, or no cognitive mechanisms at all (for more on this theme, see Chapter 12). This difference between rationalists and empiricists makes sense given that, other things being equal, the richer and more complex the mature mind is, the harder it is to acquire using only a few, simple, domain-general learning mechanisms. This difference is one of the reasons why the rationalism-empiricism debate has been so fervent and difficult to resolve—and why it so important. In the end, the rationalism-empiricism debate involves very different

competing pictures of the overall structure the mind (Chomsky 1972/2006; Samuels 2002).

The empiricist tendency to see the mind as relatively lacking in complexity is starkly evident in the strands of empiricism that claim there is little more to the mind than perception and action-governing mechanisms. This tendency can be seen in behaviourists like B. F. Skinner and his proposal that a person's verbal behaviour can be accommodated entirely in terms of a socially mediated history of reinforcement (Skinner 1957). But the same impulse is alive and well in more recent work in cognitive science. For example, Rodney Brooks describes the cornerstone of his *subsumption architecture* approach to robotics as the realization "that the so-called central systems of intelligence—or core AI as it has been referred to more recently—was perhaps an unnecessary illusion, and that all the power of intelligence arose from the coupling of perception and actuation systems" (Brooks 1999, p. viii). On this view, perception doesn't interface with a distinct cognitive system that is needed to plan a course of action. Instead, perception directly leads to action.[10] Brooks' team has succeeded in building robots that can move about a room without bumping into obstacles, an achievement that he describes as approximating "simple insect level intelligence" (Brooks 1999, p. 98). Moreover, Brooks' claim isn't that the interest of this work is limited to insect-level cognition. He takes it to provide a model for understanding intelligent behaviour in general, suggesting that "the subsumption architecture (or one like it in spirit) can be expected to scale up to very intelligent applications" (Brooks 1999, p. 175).[11]

We do not see any reason to suppose that this approach *is* likely to "scale up" to explain sophisticated human behaviour. Indeed, we would argue that Brooks has not even begun to capture the complexity of insect behaviour, much less anything that approximates human action. Consider the case of path integration discussed earlier—something that ants can do. Path integration involves combining information about changes in distance and direction. At a minimum, this requires computations that have access to stored representations and hence cannot be accomplished by action-guiding systems that respond merely to what is immediately perceived (Gallistel 1990). Or, to take another example, Brooks is likewise in no position to explain the bee dance system of communication, in which bees consolidate information about the physical movements of fellow bees with other information (e.g., about food quality) and use this to determine whether to leave

[10] See Chapter 22 for discussion of a strand of research in the embodied cognition framework that argues for the related idea that there is a direct coupling between perception and action.

[11] Commenting on the provocative title of one of his papers ("From Earwigs to Humans"), he notes that it alludes to the title of a paper that was a critical discussion of some of Brooks' earlier work (Kirsh's "Today the Earwig, Tomorrow Man?"). Brooks writes that Kirsh's title "was meant in a somewhat contemptuous spirit, arguing that behavior-based approaches, while perhaps adequate for insect-level behavior, could never scale to human-level behavior. The Cog project, and in a little way, this paper, are my response. Or, more precisely 'Yes, exactly!'" (Brooks 1997, p. 301).

the nest and, if so, which direction and distance to fly in, all the while compensating for changes in the sun's position owing to the amount of time spent in the nest (Gould and Gould 1995; Chittka 2023). Moreover, these difficulties just scratch the surface. In a great many cases, rich cognitive explanations are all but inevitable once we begin to look closely at what actual living insects can do.

Take the way that ants select a nest site. Franks et al. (2003) show that ants (of the species *Leptothorax albipennis*) are sensitive to a number of factors, including floor size, ceiling height, entrance size, darkness level, the hygiene of the cavity, and the proximity of hostile ant groups. The ants exhibit consistent ranked preferences for a range of such factors, both in pairwise choices between potential nests, and choices among a larger range of options. Franks et al. were able to rule out several relatively simple domain-general strategies (a satisficing strategy, for example, where the nest with the highest value on the top ranking feature is chosen, with ties decided by the highest value on the second highest ranking feature, and so on). They conclude that ant colonies are using a weighted additive strategy, "one of the most thorough, computationally expensive, and time-consuming, decision-making strategies" (Franks et al. 2003, p. 222).[12] These and other examples strongly suggest that Brooks has greatly underestimated real insect-level intelligence and consequently that approaches like his subsumption architecture *under*intellectualize the mind.

Of course, not all empiricists are as opposed to internal representational processes as Skinner and Brooks. But even less extreme empiricists are prone to overlook behavioural complexities that speak to how intricate the innate structure of the mind may be. Another example, a particularly ironic one for empiricists, is their treatment of learning in terms of strengthened associative bonds. Take conditioned learning. Empiricists have often maintained that the learning that occurs (e.g., learning to press a bar after seeing a light) doesn't require specialized mechanisms. It only requires a general capacity to strengthen an association between seeing the light and pressing the bar (or a corresponding association in the brain). The bond gets stronger when the interval between the events is shorter and as the reward increases in intensity. Although this model has been extremely influential, there is good reason to believe that conditioned learning is not simply a matter of strengthening an association but involves computing the rate at which a contingency occurs.

C. R. Gallistel and John Gibbon (2002) show that empiricists have neglected evidence, often available in the empiricists' own data, that learned associations are in fact independent of each of the standard empiricist factors that are supposed to determine strength of association—the temporal closeness of the pairing, the repetition of the pairing, and the strength of the reinforcement. For

[12] Interestingly, given that ants are not likely to make frequent use of this complex and domain-specific inference procedure, there is good reason to favour a rationalist account of its acquisition.

example, in standard bar press reward experiments, the temporal closeness of the pairing of response and reward is not relevant, provided that the ratio between (a) *the time between trials* and (b) *the time between stimulus and reward* is held constant. On Gallistel and Gibbon's rate estimation theory, this is because the pairing of response and reward is highlighted when the contingency of the one on the other is made salient by fixing this ratio. Empiricists, under the influence of the bond strength model of conditioning, failed to consider this possibility. Much the same applies to the other factors that allegedly determine strength of association. These and related findings strongly suggest that the psychological mechanism that supports conditioned learning is a more specialized mechanism than it had been thought to be, perhaps one with an evolutionary history tied to foraging behaviour.

Seeing the complexity of behaviour—people talking, ants going about their business, etc.—is surprisingly difficult. As Chomsky (1972/2006) has noted:

> One difficulty in the psychological sciences lies in the familiarity of the phenomena with which they deal. A certain intellectual effort is required to see how such phenomena can pose serious problems or call for intricate explanatory theories. One is inclined to take them for granted as necessary or somehow "natural"…we also lose sight of the need for explanation when phenomena are too familiar and "obvious". We tend too easily to assume that explanations must be transparent and close to the surface. (pp. 21–22)[13]

In responding to the charge that rationalists overintellectualize the mind, we have pointed to a few instances where empiricists succumb to the opposite error—they don't posit enough cognitive structure. There are many further examples that we could mention along these lines, but this is not the place for that. For present purposes, our goal is simply to address objections to rationalism that are meant to undermine the very need to undertake a detailed examination of the evidence for rationalism. These few examples, we think, are more than enough to show that an examination of the evidence is essential. Whether rationalists systematically overintellectualize the mind—or, for that matter, whether empiricists systematically underintellectualize the mind—is not a matter that can be settled in advance of empirical inquiry, and is not really an independent question from the

[13] Historically, a further reason why philosophers have had trouble recognizing the complexity of the mind is that they have relied too heavily on introspection. While this is not the case for contemporary philosophers whose work is informed by the cognitive sciences, philosophers who are somewhat at a distance from the scientific study of the mind may not realize that there is no longer any serious scientific dispute about the enormous extent to which unconscious mental processes dominate mental life and consequently about the fact that introspection is often a poor guide to how the mind works. See, e.g., Searle (1992) for an example of a philosopher failing to appreciate this fact. It is no coincidence that Searle is sceptical both about the unconscious and about standard rationalist models of language acquisition.

question of which type of account, rationalist or empiricist for any given psychological trait, is best supported by the available evidence.

Rationalism has a gene shortage problem. Another reason that rationalist accounts have been taken to be not deserving of any real consideration is that they have been thought to be inconsistent with what we have come to learn about the human genome. For example, Paul Ehrlich has argued that rationalist views suffer from a "gene shortage". "Genes cannot incorporate enough instructions into the brain's structure to program an appropriate reaction to every conceivable behavioral situation or even to a very large number of them" (Ehrlich 2000, p. 124). Paul Churchland, who has been a forceful critic of rationalist theorizing, agrees:

> If we put Almighty God and Plato's Heaven aside as nonstarters…[rationalism] confronts the difficulty of how to code for the individual connection-places and connection-strengths of fully 10^{14} synapses—so as to sculpt the target conceptual framework—using the resources of an evolved genome that contains only 20,000 genes, 99 percent of which (all but a paltry 300 of which) we share with *mice*, with whom we parted evolutionary company some fifty million years ago. (Churchland 2012, p. 15)

Tempting as these views may be, they face a number of serious problems. One issue is that it is simply not known how many genes are required to produce a neural system that could subserve a given set of innate psychological structures (that is, a given acquisition base, whether it be a rationalist or an empiricist acquisition base). Churchland tells us that 20,000 genes are not enough. Other critics running much the same argument (but on earlier estimates of the number of genes in the human genome) claim that even twice that number still wouldn't be enough: "We have literally trillions of synaptic connections in our head. There is no way even 40,000 genes could code for that exactly" (Buller and Hardcastle 2000, p. 314).[14]

One reason for these critics' confidence seems to be that they are assuming that rationalists hold that every neuron and every connection between neurons must be individually coded for by our genes. However, there is no reason to suppose that rationalists are committed to this implausible view. As we saw in Chapter 2, the central rationalist claim is about the character of the acquisition base. Rationalists hold that there are a significant number of innate characteristically rationalist psychological structures that play an important role in cognitive development. The innateness claim here is that these psychological structures are part of the acquisition base, that is, their acquisition is not mediated by

[14] For further examples of this type of argument against rationalism, see Bates et al. (1998); Levinson (2003); and Evans (2014); among others.

psychological-level processes. Since this in no way entails that the neural underpinning of these traits (much less the rest of the brain) is genetically specified on a neuron-by-neuron and connection-by-connection basis, there is no reason why rationalists must be committed to this view. And given its implausibility, it is unsurprising that no one—rationalists included—actually supposes that there are genes that specifically and individually code for each neural connection in the human brain.

Nonetheless, the core idea behind the worry about a gene shortage is still one that might be thought to challenge rationalism. Ehrlich, and others, can be read as being sceptical of the idea that even "a very large number" of psychological structures (representations, rationalist learning mechanisms, or resources in the acquisition base that such learning mechanisms trace back to) could be innate given the paucity of genes, a view that is much closer to mainstream rationalist thinking. One might naturally suppose that, other things being equal, the richer the acquisition base, the more difficult it is to explain its development on the basis of a genome containing only 20,000 or so genes. So it is worth examining the argument in more detail.

The first point to note is that the estimate of 20,000 genes concerns what we will refer to as *protein-coding genes*.[15] Before we can discuss the significance of this point, it will be useful to have some background in place. As is well known, cell function and reproduction is governed by genetic material in DNA stored in cells. DNA codes for proteins indirectly. DNA produces messenger RNA (mRNA) through a process known as *transcription*. The structure of the mRNA molecules that are produced mirrors the structure of the portions of DNA it is transcribed from. These mRNA molecules are then used to produce particular proteins in a process known as *translation*. In translation, sets of three of the mirrored units in the mRNA known as *codons* are read off by cellular machinery, with each type of codon corresponding to a particular amino acid building block of proteins (one of the standard twenty amino acids in humans) or to a "stop" signal (which terminates the protein construction process). Different proteins are composed of different long strands of amino acids in different sequences drawn from this very small set.

Proteins are of enormous significance in cells and organisms—in fact, it is almost impossible to overstate their importance. Many key structural components

[15] There are a number of reasons why this remains an estimate. Among other issues, small portions of the genome have been relatively inaccessible to the technology that was used to determine the bulk of the genome. And though there are reliable ways to predict which portions correspond to protein-coding genes, the predictions are not yet definitive. A recent study aimed at comprehensiveness puts the current estimate at 19,969 protein-coding genes (Nurk et al. 2022). In earlier work in molecular biology, the term *gene* was exclusively used to refer to protein-coding genes. Today, however, it is commonly used in a broader way, which also includes stretches of DNA that code for other kinds of RNA (besides mRNA) that are discussed below. For this reason, we use the term *protein-coding genes*—for clarity—to refer specifically to genes that code for proteins.

of cells are built out of proteins, and proteins underpin many specialized functions of different types of cells—for example, the protein haemoglobin in blood allows blood to transport oxygen, the protein rhodopsin in the eyes allows organisms to detect light, and keratin proteins are key structural elements in skin, hair, and nails. Proteins serve as intra- and intercellular signals. Antibodies are specialized proteins. Proteins combine together to form protein complexes that function as tiny machines within cells with "nearly every major process in a cell [being] carried out by [such] assemblies" (Alberts 1998, p. 291). Proteins are the primary catalysts for chemical reactions within cells, being responsible for "almost all the catalytic functions in" cells (Williamson 2012, p. v). And they are extraordinarily effective in this role, enabling as much as a one-hundred-trillion-fold increases in the rate of chemical reactions (Alberts et al. 2015, p. 58).

The core of the gene shortage argument is that 20,000 protein-coding genes is not enough to support a rationalist account of the origins of the human mind. Perhaps the thought is that 20,000 proteins just wouldn't be enough to explain both the development of our bodies and all the psychological structures rationalists take to be innate. However, there are several reasons why it is misleading to focus on the figure of 20,000 here. First, though the one gene / one protein view was once the received view, it isn't any longer. Far more than 20,000 different types of proteins are produced from our 20,000 protein-coding genes. One reason for this is due to the phenomenon known as *alternative splicing*, which is now thought to occur for 95% of human protein-coding genes. In producing an mRNA transcript in the process of transcription, a long portion of DNA is transcribed and then spliced with portions in the mRNA copy being systematically excised after the initial copy is made. Alternative splicing refers to the fact that this editing process can be done in different ways, meaning that the same stretch of transcribed DNA (that is, the same protein-coding gene) can be used to produce different mRNA transcripts that can then be translated into different proteins.

Alternative splicing can in principle produce an enormous number of different proteins from a given protein-coding gene—in some cases, hundreds or even thousands of different proteins (Alberts et al. 2015). Moreover, any given mRNA transcript, including variants produced through alternative splicing, can produce multiple different proteins through processes of *post-translational modification*, where a protein that is generated through the process of translation can be modified after it is produced in a large variety of different ways. All of this yields many further types of proteins. Indeed, although the total number of different types of protein variants in humans is not known, when variation across humans is also taken into account, the total number of protein variants produced by the human genome is estimated to be in the millions (Twyman 2014; Aebersold et al. 2018; Timp and Timp 2020). This strongly suggests that we shouldn't put too much weight on the number of protein-coding genes.

A second, and more important, reason why it is misleading to focus on the estimate of 20,000 genes is the discovery that our 20,000 or so protein-coding genes account for less than 2% of the human genome (The ENCODE Project Consortium 2012, p. 58). For many years, it was thought that the so-called non-coding DNA was useless junk, a harmless but non-functional artefact of our evolutionary history, much like a vestigial tail or appendix. But the monumental 2012 ENCODE study found that at least 80% of the genome is functional (The ENCODE Project Consortium 2012, p. 57).[16] While there is controversy about precisely how much of the genome is functional (in part, turning on different ways of characterizing what constitutes being functional), it is not controversial that protein-coding genes are not the only important elements in DNA and that significant portions of the genome beyond the 20,000 protein-coding genes are enormously important to cell function. In retrospect, this is perhaps unsurprising. After all, the simple nematode worm *Caenorhabditis elegans*, which has only 302 neurons and (excluding gametes) roughly 1000 cells altogether (as compared to our roughly eighty-six billion neurons and roughly thirty-seven *trillion* cells), has a genome with roughly the same number of protein-coding genes as ours (approximately 20,000) (The *C. elegans* Sequencing Consortium 1998). And the genome of a widely cultivated form of rice (*Oryza sativa L. ssp. indica*) has more than double the number in our genome (Yu et al. 2002).

It turns out that, while less than 2% of the human genome is protein coding, as much 75% of the human genome may be capable of being transcribed, suggesting that a large portion of the genome is involved in coding for something other than proteins (Djebali et al. 2012). Much of this DNA seems to be involved in producing regulatory RNA products. Regulatory RNA plays a key role in gene expression, helping to determine where (in which cell types), when, how many copies, and in which combinations transcripts are produced from different segments of DNA (including protein-coding genes) and can control such processes as alternative splicing.[17]

To get a sense of the importance of regulatory RNA, consider how proteins produce their effects. Proteins do not operate in a causal vacuum. Rather, any effect they have depends on interactions with cellular machinery of various kinds. Since a given protein's effects may vary depending on the presence or absence of other proteins, its effects may vary as a function of the synchronous activation

[16] The Encyclopedia of DNA Elements Project (or ENCODE) aimed to "delineate all functional elements within the human genome" (The ENCODE Project Consortium 2012, p. 58). Functionality was operationalized as the participation "in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type") (p. 57). As the ENCODE team notes, the 80% figure is likely an underestimate, since they were not able to examine the activity of the genome in every cell type, and portions of the genome with no assigned function in the examined cell types may well be functional in unexamined cell types (p. 60).

[17] Marcus (2004) emphasized the importance of regulatory genes in an early response to the gene shortage argument.

of a large number of genes. This coordinated activity in the genome is also controlled by the regulatory RNA through selective activation of regulatory RNA (by other regulatory RNA) in a complex hierarchical network of interrelations. This network is only just beginning to be understood. But it is clear that by controlling when, where, and in which combinations protein-coding genes are activated, regulatory RNA can dramatically affect the impact that protein coding genes have on cells and organisms.

In light of the major role that regulatory RNA seems to play, it is perhaps unsurprising that, while the number of protein-coding genes is a poor indicator of the organism complexity, organism complexity *does* correlate with the fraction of the genome devoted to coding for regulatory RNA (Taft et al. 2007). In fact, together with the phenomenon of alternative splicing, regulatory RNA is now widely seen as the key to understanding why the number of protein-coding genes dissociates so dramatically from measures of organismic complexity.[18] Many see the discovery of the major role of RNA produced by non-coding DNA (that is, DNA that codes for RNA but that isn't protein-coding DNA) as completely transforming our understanding of the genome. For example, Nobel Prize winner Thomas Cech and co-author Joan Steitz refer to a "noncoding RNA revolution" in which "every established 'rule' seems destined to be overturned" (Cech and Steitz 2014, p. 77). In a 2014 review of the role of regulatory RNA, Morris and Mattick (2014) likewise suggest that we have to completely rethink even our most basic assumptions about the genetic information:

> in retrospect, it seems that we may have fundamentally misunderstood the nature of the genetic programming in complex organisms because of the assumption that most genetic information is transacted by proteins. This may be true to a large extent in simpler organisms but is turning out not to be the case in more complex organisms, the genomes of which seem to be progressively dominated by regulatory RNAs that orchestrate the epigenetic trajectories of differentiation and development. (p. 431)

Given this relatively recent appreciation of the enormous importance of regulatory RNA, we should expect to find substantial differences between the regulatory RNA of humans and organisms with simpler brains (e.g., mice). And this is exactly what has been found. While the human genome and the mouse genome are very similar in terms of protein-coding DNA, they differ dramatically in terms of regulatory RNA (Yue et al. 2014).

[18] We should note that at least some of the proponents of the gene shortage argument put the argument forward before the importance of alternative splicing and the role of regulatory RNA came to be widely appreciated.

In sum, the gene shortage argument goes wrong in a number of ways. It is misleading to focus on the number of protein-coding genes. The *one gene / one protein* idea is outdated—in principle many different proteins can be produced from a single protein-coding gene because of such things as alternative splicing and post-translational modification. By controlling when, where, how many, and in which combinations proteins are produced, regulatory RNA has an enormous impact on behaviour and ontogeny. Moreover, some of the implicit assumptions connected with the argument are clearly false. For example, rationalism is not committed to individual genes coding for each and every neuron and synaptic connection. In addition—and this may well be the most important point here—since no one knows how many genes or proteins are required to produce any given acquisition base, claims that a given number of protein-coding genes doesn't suffice are groundless. It is crucial to bear in mind that it is very much an open question how what happens at the level of proteins translates to the level of psychological structures. Without a clear sense of how protein numbers constrain varieties of psychological structures that can be produced, it is very much an open question whether limitations in the number of proteins significantly constrain the possible varieties of psychological structures that could be present in the acquisition base. Nature offers many examples where a relatively small number of different kinds of basic elements are capable of combining to generate enormous variety (e.g., as we have noted, a mere twenty amino acids in various combinations compose all of the myriad different proteins in humans, and just one hundred or so basic chemical elements in various combinations compose essentially all visible matter in the universe). So, for all we know, the components that figure in the biological basis for developing the acquisition base are capable of producing essentially any number of psychological structures as part of the acquisition base. Of course, none of this constitutes a positive argument *for* rationalism. But what it does show is that the gene shortage argument fails to establish that rationalism should be rejected out of hand.

Rationalism is anti-developmental. Another objection that is often raised against rationalism is that rationalists fail to recognize, or downplay to their detriment, the fact that meaningful conceptual change takes place in development. This charge—that rationalism is anti-developmental—is sometimes put by saying that when rationalists find any evidence for the early presence of a target cognitive capacity, they immediately jump to the conclusion that the full mature capacity is innate. Haith (1998), for example, criticizes rationalists for using "indications for the earliest fragments of a concept as evidence for virtual mastery of the concept" (p. 168). This hasty inference is said to reflect rationalists' dichotomous thinking, in which cognitive abilities are invariably assumed to be either wholly innate or learned. Rationalists are said to be ignoring the possibility that development is piecemeal. What may look like evidence for an innate cognitive ability to rationalists may just

be an initial stage in development; the full cognitive ability might only be learned over what may be an extended period in childhood. Prinz (2012) also makes this objection, saying that "developmental psychology has forgotten all about development and assumes that knowledge is already in place" (p. 109).

There are a number of problems with this type of objection. First, it is wrong to associate rationalism with the view that psychological development doesn't occur, taking rationalists to suppose that infants are exactly like older toddlers, toddlers exactly like preschoolers, and so on. Rationalists, just like empiricists, suppose that development takes place. It is just that the mechanisms of development that they posit are often different.[19] Consider, for example, a rationalist theory of language acquisition, in which the acquisition base is thought to incorporate the principles of Universal Grammar. Such an account takes language acquisition to be a gradual process and posits numerous distinct stages in the acquisition of phonology, syntax, the lexicon, and much else. An example we encountered earlier, in section 4.1, was the temporary stage in which some English-speaking children adopt a set of linguistic principles that allows for sentences like "What do you think what Cookie Monster eats?", producing a pattern of speech that conforms to grammatical principles in German, Irish, and a number of other languages. Rationalists aren't opposed to developmental changes, developmental stages, or piecemeal acquisition. However, in rationalist theorizing, these changes will often have a distinctively rationalist character.

Rationalists also recognize other forms of development, ones in which learning doesn't take place but changes in a child's cognitive processes occur all the same. One of these is when development is owing to processes of biological maturation. Some cognitive traits may be innate (that is, part of the acquisition base) yet not be present at birth. Such traits may not appear until later in life, in the life stage in which they are needed (e.g., systems for mating or for parental care are of no use to newborns). Another form of development that rationalists can point to in explaining behavioural change relates to the competence-performance distinction that we introduced in Chapter 2. This is development in performance factors that interact with cognitive competences. In such cases, development most definitely occurs, but what develops isn't the basic competence. Rather, features of performance change as other systems (e.g., memory, attention, or motor control) develop. Thus, even when there are dramatic developmental changes, it remains an open question what is driving the development: an empiricist learning process, a rationalist learning process, maturation related to an innate competence, maturation or development related to performance factors, or some combination of these possibilities.

[19] And, while they are different, it's worth noting again that rationalist theories incorporate domain-general learning mechanisms that operate alongside of rationalist domain-specific learning mechanisms (see Chapter 2). So rationalists can and sometime do explain aspects of development in much the same way as empiricists.

What about the claim that rationalists mistake evidence for an early form of a cognitive capacity to be evidence for the full-blown capacity? Here too the objection is misplaced. It is certainly true that in some instances rationalists maintain that the full capacity is present (perhaps somewhat obscured by limiting performance factors), but rationalism is in no way committed to this always being the case. As we have already seen in the examples discussed in Chapter 2 (involving Euclidean geometrical concepts and concepts of natural numbers) and will see in many more cases discussed in later chapters, rationalists often maintain that what is innate simply provides a starting point for learning about a given domain. Such a starting point might involve an innate rationalist learning mechanism, or one or more innate resources involved in such a mechanism. It might also involve innate representations that provide initial access to the domain in question, helping learners to represent and attend to key objects, events, or properties. Moreover, these innate starting points for development may not only be immature or incomplete. They may even be substantially altered or overridden later in development. The value of having innate rationalist structures that provide some form of initial access to a domain isn't that it gives children a fully articulated adult-like response to items within the domain. It's that it gets children past what may be the biggest stumbling block to development—being able to represent the often abstract and peculiar items in the first place (things like other people's mental states, numerical quantities, and instances of physical causation).[20]

Rationalism is committed to primitives whose origins are unexplained or even unexplainable. Another way in which rationalism has been said to marginalize development is by positing developmental primitives that are thought to be inherently problematic. Lewkowicz (2011) writes that "rationalists believe that evolution has endowed the human species with something like primitives, core cognitive capacities, or principles that are directly related to specific domains of knowledge including language, object, number, geometry, space, social relations, morality, and religious belief" (p. 333). According to Lewkowicz, the problem with positing such primitives is that their origin is left entirely unexplained: "even if one assumes that primitives exist, their development must be explained too. Instead of doing so, nativists ask us to take their existence on faith" (p. 356). Likewise, Spencer, Samuelson, et al. (2009) note that "positing innate building blocks does not inform our understanding because, as Moore (2009) correctly points out, building blocks must themselves be built" (p. 104). Indeed, Spencer, Samuelson, et al. go even further, taking rationalists to be committed to primitives that are entirely uncaused and hence inexplicable, writing that "'primitives' *are not developed or derived from anything else*" (p. 79; emphasis in original).

[20] For further discussion of how the need to overcome this stumbling block supports the case for rationalism, see Chapter 12.

Are rationalists really committed to primitives with unexplained origins that must simply be accepted on faith or whose origins are in principle inexplicable? The answer is "no". While it is true that rationalist are committed to psychological structures that are primitive, these primitives are not accepted merely on faith and they are not primitive in the problematic senses of "primitive" that these critics are assuming. The only sense in which rationalists are committed to explanatory primitives is in the sense of there being innate psychological structures in the acquisition base, as was explained in Chapter 2. This means that while these structures are involved in psychological processes that account for the development of further psychological traits, there is no psychological explanation for their origins. But rationalists don't think that there is no explanation of their development at all—that the origins of these structures is completely inexplicable. The primitives that rationalists posit are part of the acquisition base, which means that they are primitive *relative to the psychological level of explanation* (they are not acquired via psychological-level processes). They are not primitive *relative to all processes of acquisition*; they don't materialize out of thin air. Instead, they are acquired by biological processes operating within the developing organism that are not at the same time psychological processes—in particular, they are not learning processes.

Primitives in this sense are essentially mandatory when explaining the development of psychological traits. As we argued in Chapter 2, *any theorist at all* who holds that there are psychological traits—whether they are a rationalist or an empiricist—has no choice but to accept there are psychological traits that are primitive in this sense, which aren't acquired via a psychological processes. Given the importance of this point, it is worth working through the argument for it once more. Consider a given psychological trait, T. Trait T is either acquired via a psychological-level process or not. If it isn't, then T itself is primitive in the relevant sense and we could just stop here—there would be no question about the existence of primitive psychological traits. So let's suppose that it is acquired via a psychological process and that there is a psychological mechanism involved in its acquisition, which we will call LM1 (for learning mechanism 1). Now we have to ask whether LM1 is acquired via a psychological-level process or not. Again, if it isn't, then LM1 is primitive in the relevant sense, and we have arrived at the conclusion that there is at least one psychological primitive. To avoid this, we'd have to say that LM1 is acquired via a psychological process that involves a second learning mechanism, LM2. Clearly this process cannot go on forever, with LM2 being acquired by LM3, LM3 being acquired by LM4, and so on, ad infinitum. This would require that before a mind could possess a single psychological trait, it would already have to possess infinitely many prior learning mechanisms. (The only other alternative—but not a real possibility since it is viciously circular—is a situation where something like this occurs: LM1 is learned by using LM2, LM2 is learned by using LM3, and LM3 is learned by using LM1.)

Empiricists, then, are committed to the existence of innate primitives in exactly the same sense as rationalists and for exactly the same reasons that rationalists are. Of course, empiricists will not posit the *same* set of innate primitives as rationalists. Empiricists will posit primitives appropriate to an empiricist acquisition base, while rationalists posit primitives appropriate to a rationalist acquisition base. But the primitives that each type of theorist posits will be primitive in exactly the same sense. And while the origins of such primitives will not be explicable at the psychological level, this does not make them mysterious or unnatural as their origins will be explicable in terms of biological processes operating within the developing organism (even if the details of such explanations are not presently known).

Rationalism implies that psychological traits and human behaviour are inflexible. Another objection often levelled against rationalism is that it does not sufficiently allow for cognitive and developmental flexibility. This type of objection comes in a variety of forms, some of which are more focused on the flexibility of mental faculties and the psychological traits involved in development and others of which are more focused on the flexibility of human behaviour and social structures.

As applied to innate psychological traits, the main worry is that rationalist accounts imply that innate psychological traits are fixed and unchangeable, especially general traits and aptitudes, such as personality traits, intelligence, or general cognitive abilities, such as a talent for science or music. The objection here is that it is a major mistake to view psychological traits in this way and that rationalists are forced to view them in this way. Perhaps the most prominent versions of this general type of objection, however, focus on the flexibility of human behaviour, and charge that rationalist accounts are incompatible with such flexibility. An especially strong form of this objection claims that because rationalism is committed to an erroneous picture of the mind in which environmental cues activate innate domain-specific mechanisms causing them to issue in specific types of behaviour, this means that once a critical environmental condition is encountered, the corresponding behavioural outcome is all but inevitable. This objection has particularly been raised against rationalist theories associated with evolutionary psychology and evolutionary theorizing about the mind. The perceived inevitability of the resultant behaviour has been described as a pernicious form of determinism and reductionism. As Hilary and Steven Rose (2000) remark, "This new determinism…claims…our biology is our destiny, written in our genes by the shaping forces of human evolution through natural selection and mutation" (p. 4). And Steven Rose (2000) adds that "[evolutionary psychology] offers yet another reductionist account in which presumed biological explanations imperialise and attempt to replace all others" (p. 247).

Related criticisms of rationalism have also been given regarding the societal implications that critics take to be (or to be seen as) a consequence of such inflexible forms of behaviour. One of these associates rationalism with patriarchy, the idea being that if rationalism were true, then there would be no way of avoiding a patriarchal division of labour. For example, Eagly and Wood (2011) remark that the debate between evolutionary psychology and many of its critics comes down to a disagreement about "the potential for change in female and male behavioral patterns" (p. 759). Eagly and Wood take the critics of evolutionary psychology to hold that these patterns are malleable but take evolutionary psychologists to deny this:[21]

> [E]volutionary psychologists…view sex differences, especially male-female inequality, as inevitable consequences of evolutionary adaptations and therefore as largely unresponsive to socioeconomic and political changes in society. Their essentialist explanations of sex differences are rooted in Darwinian sexual selection theory, which has fostered the view that male dominance and female dependence derive from inherited dispositions. (p. 759)

In our view, it is a mistake to think that rationalism—including rationalism as developed in many accounts in evolutionary psychology—entails that human behaviour is inflexible or that patriarchy (or any other similarly objectionable social arrangement) is an inevitable feature of human social life.[22]

[21] Eagly and Wood frame this as a debate between feminists and evolutionary psychologists. Since evolutionary psychologists have at times adopted a condescending and dismissive attitude towards explanations suggested by theorists who have self-identified as feminists, this framing is understandable. But as we see it, framing the debate in this way is also unfortunate both because it may discourage theorists who see themselves as falling into one of these groups from productively engaging with theorists taken to be in the other group and because there is no reason why theorists cannot be *both* evolutionary psychologists and feminists. In fact, there is now arguably substantial overlap of exactly this kind, with many evolutionary psychologists considering themselves to also be feminists. In a report on the number of women working in their field, Frederick et al. (2009) remark: "As feminists and evolutionary social scientists, we agree that studies of women's psychology have lagged behind those of men in the social sciences and that the field of evolutionary social science was historically male-dominated (as is true of most academic fields)…Today, one third of evolution and human behavior and human nature authors are women (January 2001–March 2008), as are one third of current editors and three of the five chief editors. Among the younger generation, 40% of poster presenters (typically graduate students) at the 2007 Human Behavior and Evolution Society were women. The increasing presence of women in evolutionary social science has coincided with an explosion of research concerning women's lives and sexuality" (p. 302).

[22] Even though rationalism doesn't entail inflexibility, the belief that a person's nature is fixed and immutable can be inimical both to personal development and to improvements in relations among conflicting groups. There is research suggesting that introducing the idea that people's psychology is mutable helps to increase people's ability to succeed in areas they would otherwise take to go against their nature (Dweck 2006) and even to increase the extent to which individuals from groups in long-standing conflicts view compromise as a genuine option (Halperin et al. 2011). This is another reason why it is important for rationalists to emphasize the fact that rationalism is a thesis regarding the character of the acquisition base and does not entail fixed or immutable cognitive or behavioural outcomes.

Let's start with the concern about the fixed nature of innate psychological traits. The easiest way to see that this worry about rationalism is misplaced is to return to the characterization of rationalism that we developed in Chapter 2. On this account, rationalism's central claim is that the acquisition base isn't confined to domain-general learning mechanisms and that it also includes many characteristically rationalist psychological structures (such as innate concepts or innate domain-specific resources that rationalist learning mechanisms trace back to). These psychological structures in the acquisition base are understood to be innate in the sense that they aren't acquired on the basis of more fundamental learning mechanisms. Rationalism needn't—and doesn't—hold that innate psychological traits cannot be changed, overridden, or even eliminated in the course of development. And general traits like intelligence and aptitudes for things like mathematics, science, and music are all demonstrably open to a substantial amount of change, all of which is perfectly compatible with rationalism. In later chapters, we will see numerous examples in which rationalist theories of development fully embrace the fact that innate psychological structures can be altered or supplanted in the course of development and the fact that psychological traits can undergo extensive development and can be subject to an enormous amount of cultural variation.

When the objection to rationalism is developed in a way that implies that both the psychological traits acquired via rationalist learning mechanisms and the human behaviour that stems from them are inflexible, it is problematic in other ways as well. In particular, this way of objecting to rationalism implicitly takes rationalism to be committed to the view that behaviour is governed by behavioural programmes that operate like reflexes. However, this is an erroneous view of how rationalists generally explain human behaviour. It is certainly true that there are cases where animals have psychological mechanisms that lock them into a fixed behavioural response. For example, Lettvin et al. (1959) explain how feeding behaviour in frogs is grounded in a simple mechanism that is responsive to a single narrow stimulus condition:

> The frog does not seem to see or, at any rate, is not concerned with the detail of stationary parts of the world around him. He will starve to death surrounded by food if it is not moving. His choice of food is determined only by size and movement. He will leap to capture any object the size of an insect or worm, providing it moves like one. He can be fooled easily not only by a bit of dangled meat but by any moving small object. (p. 1940)

If the psychological traits acquired by innate special-purpose learning mechanisms that rationalists postulate all functioned in this way, then human behaviour would indeed be inflexible. And perhaps it is understandable that some of rationalism's critics who push this type of objection suppose that rationalism's

resources are like behavioural reflexes, as these critics' views are sometimes more informed by the sociobiology of the 1970s than contemporary rationalist psychology. Sociobiology in its heyday did see much speculation about simple mechanisms for controlling behaviour in humans (Laland and Brown 2002). However, contemporary rationalist theories rarely postulate behavioural programmes of this kind. Rather, they take behaviour to depend upon the *interaction* of many cognitive systems. On this approach, how one behaves isn't dictated by the activities of any one of these systems. It emerges from many factors, including which other systems are operative and how they are related to one another in general and at the moment at which the behaviour occurs. These further systems will include, among others, such things as internalized cultural norms, social roles, and stereotypes, all of which are both changeable and exert an enormous influence on behaviour (see the discussion below of women in engineering).

Finally, as we have noted a number of times, rationalism isn't opposed to learning. On the contrary, rationalist theories incorporate the idea that development often involves learning and consequently that a developmental outcome is sensitive to the details of a learner's experience. This means that rationalism allows for, and in fact predicts, a fair amount of behavioural and cognitive variability across populations and individuals. Rationalist learning mechanisms may be focused on a particular domain, but they can be responsive to variations in this domain, so that different life experiences lead to different behavioural proclivities—ones that are more suited to the respective environments in which the learning took place.[23] In addition, it is important to remember that rationalism also embraces general-purpose learning mechanisms; it just holds that these mechanisms don't on their own suffice to explain what the mind can do and that many special-purpose learning mechanisms are needed too (see Chapter 2). So rationalists are free to accept that much environmental sensitivity arises from the operation of general-purpose mechanisms that are responsive to variation in the environment, or from the interaction of rationalist learning mechanisms or the innate characteristically rationalist psychological structures that they trace back to. We can see, then, that

[23] There are a number of ways in which a rationalist learning mechanism might work, and this introduces further respects in which a rationalist cognitive architecture can exhibit environmental sensitivity. For example, one type of rationalist learning mechanism has much of its basic structure already in place, and its learning consists in its sampling the environment to select among a small number of critical innately specified options (e.g., the parameter-setting model in syntax). A rather different type of rationalist learning mechanism might provide a learner with a schematic cognitive or behavioural response to items of a certain type, leaving experience to fill in which items in the environment instantiate this type (see, e.g., the discussion of rationalist mechanisms for responding to dangerous animals and plants in Chapters 14 and 19). A third type of rationalist learning mechanism might provide a learner with initial access to a given domain and the motivation to track items in this domain, but be open to discovering the sorts of information that might be relevant to explaining patterns in this domain (see, e.g., the discussion of how infants learn about physical objects in Chapters 14 and 17). This list isn't meant to be exhaustive, but merely to convey that there are a number of different types of rationalist learning mechanisms and that each of these offers its own type of behavioural flexibility.

rationalists aren't lacking for resources to explain the flexibility of human behaviour.[24]

Let's turn now to the charge that rationalist mechanisms would still restrict human behaviour in ways that would necessitate contentious forms of interpersonal relations—for example, social arrangements like a patriarchal division of labour. We should first note that it is true that some rationalist theorists, and in particular some rationalist evolutionary theorists, have advocated theories of innate sex differences that they have thought are bound to lead men and women to occupy different positions in social life, including the workplace. For example, Eagly and Carli (2007) quote Kingsley Browne, a law professor and evolutionary theorist, as holding that "Across human history, male status has led to greater reproductive success, leading to a predisposition among males to engage in the kinds of status competition that today so often have workplace implications" (Browne 2002, p. 117; Eagly and Carli 2007, pp. 29–30). Elsewhere Browne has alleged that "men's temperament gives them an advantage" in the workplace and that this has the consequence that "women will be forever consigned to lower status" (Browne 1999, p. 57; quoted in Eagly and Carli 2007, p. 30). Others have held that evolutionary considerations show that, in light of innate sex differences, women are unlikely to succeed in specific fields. For example, Ellis (2011) claims that, even in an environment maximally conducive to developing engineering skills in women, "only a minority of engineers will be females because few have brains that are configured for the sort of spatial-mathematic reasoning that will sustain an interest in such an area of study" (p. 711).

We can certainly see how views like these might lead theorists to be hostile towards evolutionary theorizing, and perhaps rationalism in general, taking such theorizing to be more interested in promoting a social and political agenda than in discovering the truth about the origins of social roles. This makes it all the more important to recognize that there is no inherent link between rationalism and views like these. As we have been emphasizing, rationalism is about what is in the acquisition base. But exactly which structures are in the acquisition base, whether they lead to sex differences in cognition, and whether such differences

[24] Another way of interpreting the objection that rationalism entails behavioural inflexibility—one that is closely linked to the charge about biological determinism—is as a worry about rationalism's reliance on law-like casual generalizations in its explanation of behaviour. It is certainly true that many rationalists understand human behaviour in such terms. However, this broad orientation is hardly unique to rationalists; it is just as common for empiricists to understand human behaviour in precisely the same terms. It should also be noted that taking behaviour to be the product of law-like casual generalizations does not entail determinism. Rationalism (like empiricism) can be developed in ways that make behavioural outcomes probabilistically related to stimuli, rather than deterministically related to them (no doubt, this is how many psychologists view the matter). And regardless of whether behavioural outcomes are deterministic or probabilistic, rationalists and empiricists can both avail themselves of the same options regarding the philosophical problem of free will (e.g., both can adopt a compatibilist metaphysics, or, for that matter, a libertarian metaphysics). For more on the different philosophical positions in the free will debate, see Kane (2005).

place significant constraints on human social relations are further issues that are very much up for debate within rationalist circles. The problem with views like Brown's isn't that they are rationalist views, it's that they are wrong.

Many rationalists, including many evolutionary psychologists, claim that there are no innate sex differences that push men and women into different occupational roles or that imply that men are better equipped to achieve higher status in the workplace or in other areas of social life. To illustrate this type of rationalist perspective, we will briefly consider the example of the demographics in the field of engineering. Despite substantial advances in the representation of women in many occupations in recent decades, women continue to be greatly underrepresented in this corner of the workplace. Why is that? Ellis's explanation is that there are innate sex differences in spatial and mathematical abilities and that these put women at a disadvantage in areas that depend heavily on spatial and mathematical reasoning. But a competing explanation—one that many rationalists endorse—is that various social factors work together to account for why there are so few women in this and related fields. We will mention just a few important social factors, drawing on an analysis given by Elizabeth Spelke, a leading figure in developmental psychology and herself a noted rationalist (Edge 2005; Spelke 2005).

One of the key points that Spelke highlights is the fact that in many societies boys and girls are assumed to have different qualities and aptitudes from the moment they are born. For example, in one study, parents who had just learned the sex of their baby described their newborns differently. The boys were described as being stronger, sturdier, and heavier than the girls. But independent assessments of these children found no such differences. The boys and girls were indistinguishable when it came to their true strength, coordination, and weight (Rubin et al. 1974; Karraker et al. 1995). Spelke also points out that adults continue to have similar biased perceptions of children as they grow up and that these biases are commonplace. Although 12-month-old boys and girls have equal motor abilities, parents don't see it this way. When asked to predict whether their child would be able to successfully crawl down a ramp, parents have been found to be more confident that their child can do this if the child is a boy (Mondschein et al. 2000). In another study, parents of sixth graders were examined regarding their views of their child's mathematical abilities. It was found that sons were thought to be more talented at mathematics than daughters, contrary to any objective measure—the boys and girls did equally well on standardized tests, were receiving similar grades in the classroom, had equal interests in mathematics, and so on (Eccles et al. 1990).

Other work has confirmed that parents' and educators' beliefs about children's mathematical abilities directly influence children's self-perceptions of their own abilities (Tiedemann 2000). Given widespread stereotypes about gender and mathematical ability, these self-perceptions can lead to *stereotype threat*, a

cognitive phenomenon where performance is impaired in light of a stereotype that one's group performs poorly in a given domain (Inzlicht and Schmader 2012; Spencer et al. 2016). For example, in one study, women underperformed on a difficult math test when they thought that the test was likely to show sex differences in performance (the usual negative stereotype), but when they were told in advance that men and women do equally well on this type of test, the underperformance vanished (Spencer et al. 1999). The impact of stereotype threat on women's performance in mathematics has been replicated many times in numerous circumstances (Spencer et al. 2016). What this work shows is that women *are* at a disadvantage in many settings that lead up to a career in engineering—not because of a difference in aptitude for mathematics, but because of mistaken expectations that women aren't well suited to this kind of work.

Stereotype threat exists in part because of the prevalence of a negative stereotype. Another social factor that differentially affects men's and women's career prospects is the widespread prevalence of implicit bias, a form of prejudice that largely operates at the unconscious level and that is present even in people who have no conscious explicit prejudices against a given group.[25] Implicit bias can hold women back from awards, jobs, promotions, and other forms of career advancement. In one study, professors were asked to evaluate CVs of candidates for a tenure-track position, where the name on a CV was randomly varied to indicate that the candidate was male or female (Steinpreis et al. 1999). It was found that when comparing average CVs, male and female professors rated a CV more favourably if the candidate was thought to be a man. For example, the same number of publications was considered to indicate a high level of productivity when the candidate's name was "Brian", but was considered insufficient when the candidate's name was "Karen".[26]

There can be little doubt that all of these social factors have a cumulative impact.[27] As Spelke has remarked, "From the moment of birth to the moment of

[25] Implicit bias even operates in people who are members of the group that is the object of bias. For example, female academics (who generally hold no explicit biases against women's academic ability) are subject to the judgement biases we describe below just as much as their male counterparts.

[26] Fortunately, there has been some progress on the question of how to mitigate the harmful effects of stereotype threat and implicit bias. For example, it has shown that strong positive role models are helpful in reducing stereotype threat (McIntyre et al. 2003), and that teaching women about stereotype threat helps to reduce its effects during mathematics examinations (Johns et al. 2005). Policies such as anonymous marking and anonymous reviews of CVs, writing samples, and so on, can also go some way towards eliminating the harmful effects of implicit bias. For work addressing these issues in philosophy, see Antony (2012) and Saul (2013).

[27] An interesting example of the complexity of interacting social factors can be found in work on why there have historically been fewer women than men studying computer science and why women who have studied computer science have often left the field. Margolis and Fisher (2002) have documented many ways in which women were deprived of essential precollege experience with computers and discouraged from pursuing computer science before and during college. Moreover, this discouragement came from all directions—from parents, teachers, and peers—and was exacerbated by the structure of the standard computer science curriculum. These negative experiences often led young women to have unwaranted doubts about their capacity for doing a computer science degree.

tenure, throughout this great developmental progression, there are unintentional but pervasive and important differences in the ways that males and females are perceived and evaluated" (Edge 2005). For present purposes, what we want to emphasize is that this likely alternative explanation for why women aren't as well represented as men in certain occupations is perfectly compatible with rationalism. As we noted earlier, Spelke's own rationalist views includes a commitment to innate representations and innate domain-specific systems for representing and reasoning about physical objects, agency, numerical quantity, geometry, and language (Spelke 2003, 2022). But this commitment doesn't prohibit her from holding that social factors, not innate sex differences, are the key to explaining why there continue to be relatively few female engineers.

We can also see from Spelke's analysis that the goal of achieving the sort of positive social change that is associated with feminist social critiques doesn't turn on the rejection of rationalism and that such concerns shouldn't be at issue when taking a stand on the rationalism-empiricism debate. As the feminist—and rationalist—philosopher Louise Antony has remarked,

> I believe that the fear that any concession to nativism can and will be used against [feminism] largely explains the unfortunate bias of many feminist and progressive theorists toward radically empiricist, social constructivist views of language and the mind, and their correlative hostility toward nativist accounts of the sort proposed by Noam Chomsky in linguistics and by many in cognitive psychology. The result, in my opinion, is a truly unfortunate disjunction between empirically well-grounded work in cognitive science and feminist discussions of language and the mind. (Antony 2000, p. 10)

Spelke's analysis of why women aren't as well represented as men in engineering highlights how a rationalist cognitive architecture doesn't force any particular social outcome. There is nothing about rationalism per se that says that society has to recognize that people are bound to be more suited to certain occupations because of their gender. More generally, rationalism does not entail that innate psychological structures in the acquisition base or the products of rationalist learning mechanisms—or behaviours that are dependent on either of these things—are inflexible. Rationalism is perfectly compatible with both change and variable outcomes. This is a good thing as there is overwhelming reason to suppose that both mundane psychological traits and many of the psychological traits that are the most highly valued and closely associated with being human, from musical or artistic ability to intelligence and abstract reasoning ability, can be dramatically affected by environmental factors, practice, and training (Dweck 2017; Nisbett 2009). Our focus in this book is not with the origins of such general capacities or with how these lead to differences in individuals' aptitudes, but rather with the origins of human concepts. However, the moral here is the same in both

cases. Human behaviour, and the psychological traits that underpin it, are indeed flexible, but this fact in and of itself is just as compatible with rationalist accounts of the mind as it is with empiricist accounts.

Rationalism is undermined by the explanatory failures of adaptationist theorizing. The last objection to rationalism that we will consider in this chapter stems from the association of evolutionary psychology and evolutionary theorizing about adaptive features of the mind with rationalism.[28] This association is seen by some as a major strike against rationalism because the field of evolutionary psychology has been taken by many to be riddled with theoretical confusion.

For those who are opposed to any connection with evolutionary theorizing, it should be noted that rationalism does not require accepting an evolutionary or adaptationist perspective, and in fact some rationalists have shown a great deal of scepticism about adaptationist theorizing about psychological traits (e.g., Fodor 2000). And evolution and adaptationism play no direct role in our characterization of rationalism (or empiricism) (see Chapter 2). So evolutionary approaches and adaptationist theorizing aren't essential to maintaining a rationalist position. Nonetheless, in our view, evolutionary theorizing can be a valuable tool in generating hypotheses about the mind and suggesting productive avenues of research. Our aim here is not to provide a full evaluation of evolutionary psychology or evolutionary theorizing about the mind, but instead just to address some common charges that have been made against this work that have been taken to bear on the status of rationalism. In particular, we will focus on three recurrent charges that claim that, in one way or another, such theorizing is fundamentally flawed and can simply be dismissed without careful consideration of individual claims or hypotheses.

The first is that theories in evolutionary psychology are nothing more than just-so stories—fanciful, purely speculative hypotheses that aren't supported by any evidence. This charge, which we mentioned in Chapter 1 as a worry in connection with the example of geometrical concepts, derives from Gould's critique of sociobiology: "Rudyard Kipling asked how the leopard got its spots, the rhino its wrinkled skin. He called his answers 'Just So stories.' When evolutionists study individual adaptations, when they try to explain form and behavior by reconstructing history and assessing current utility, they also tell just-so stories" (Gould 1978, p. 530). As we noted in Chapter 1, much the same criticism is now routinely made against evolutionary psychology.[29] What this comes to is the charge that it is a trivial matter to devise an account that postulates an innate psychological

[28] For overviews of the field of evolutionary psychology, see Buss (2015, 2019) and Barrett (2015).

[29] Recall from Chapter 1 that Fodor alluded to "the outpouring of just-so stories by which the mainstream of evolutionary cognitive psychology is very largely constituted" (Fodor 2001, p. 627). For further examples of this charge, see S. Rose (2000) and Richardson (2007).

trait and an accompanying "explanation" of its adaptive function, or to come up with multiple accounts with no serious way to choose between them. Often this same point is made by saying that evolutionary psychology's theories are untestable, since they make substantial assumptions about inaccessible features of the distant past regarding the environments in which much of human evolution took place, the *environment of evolutionary adaptiveness* (or EEA). As Gould puts it, "claims about an EEA usually cannot be tested in principle but only subjected to speculation…how can we possibly know in detail what small bands of hunter-gatherers did in Africa two million years ago?" (Gould 2000, p. 100).

Second, evolutionary psychologists are said to see adaptations everywhere, or to be *panadapatationists* who hold that nearly every psychological trait is an adaptation. Critics see this as meaning that evolutionary psychologists thereby ignore the significance of evolutionary processes other than natural selection (e.g., genetic drift) and the possibility that any given trait could turn out to be a byproduct of selection rather than one that was selected for as such. Gould has even suggested that, for all we know about the evolution of the mind, very few psychological traits may turn out to be adaptations: "Natural selection made the human brain big, but most of our mental properties and potentials may be spandrels—that is, nonadaptive side consequences of building a device with such structural complexity" (Gould 2000, p. 104).

Finally, the third charge is that explanations in evolutionary psychology are really pseudo-explanations because they are so flexible that they can accommodate nearly any behavioural outcome. Dupré (2001) voices this objection using an example taken from the study of human mating behaviour. He notes that evolutionary psychologists have attributed to women a mating psychology for long-term pair bonding, and that, at the same time, they have attributed to women an evolved mating psychology for short-term mating outside of a long-term partnership. The result is that, however a woman behaves, there is a mechanism that can be invoked to accommodate the behaviour. Thus "the theory is almost infinitely malleable and consequently empirically empty" (p. 64). Likewise, H. Rose (2000) mentions an example in which evolutionary psychologists have postulated systems that dispose a mother to be protective of her newborn and also systems that promote abandoning an infant in certain extreme conditions. "Used like this selection explains everything and therefore nothing" (p. 123).

What should we make of these criticisms? Let's start with the allegation that evolutionary psychology is little more than a collection of just-so stories. In our view, this objection can be seen to be misplaced once one is clear about the standard methodology that guides much of the research in evolutionary psychology. This research proceeds by using evolutionary thinking to devise hypotheses regarding the innate structure of the mind, hypotheses that are informed by current views about the conditions in which humans lived for the vast majority of the existence of our species. Once a hypothesis is on the table, it is then subject to

experimental testing in exactly the same way that any other psychological theory is tested. As David Buller (a staunch critic of evolutionary psychology) puts it, evolutionary psychology proposes "to discover our universal human nature by analyzing the adaptive problems our ancestors faced, hypothesizing the psychological mechanisms that evolved to solve them and then testing those hypotheses using standard-fare psychological evidence" (2009, p. 76).[30]

Given this methodology, it should be clear that the charge that evolutionary psychology's theories are merely just-so stories doesn't hold up. After all, they are tested in exactly the same way that other psychological theories are tested. Hence they aren't any more speculative or fanciful than psychological theories that aren't grounded in evolutionary thinking (e.g., standard psychological theories of memory, attention, or decision making). Many of evolutionary psychology's hypotheses may well turn out to be false—a likely outcome given that many hypotheses in any area of science will turn out to be false—but they shouldn't be discounted simply because they derive from views about the evolution of our species. In evaluating a hypothesis, the origin of the hypothesis shouldn't matter. What we need to focus on is the evidence for or against the hypothesis, regardless of why the hypothesis was initially proposed.[31]

The contention that we aren't in a position to know anything about the EEA—and that this makes theories in evolutionary psychology untestable—is also unwarranted. Gould and others talk as if we can't make any reasonable conjectures about the physical and social environments of our ancestors, that views about the EEA are nothing but pure speculation. But while it is true that we lack direct evidence regarding many important details, information from different disciplines can be combined to form a reasonable picture of some of these conditions, including information from behavioural ecology, evolutionary biology, genetics, palaeoanthropology, hunter-gatherer archaeology, primatology, and the anthropology of living hunter-gatherers (Tooby and Cosmides 2005). For example, there is little question that our ancestors were omnivores. The fossil record includes evidence of sites where ancestral hominids extracted animal products, and this is consistent with practices found in contemporary

[30] See Buss (2019) for many specific examples of psychological experiments that have been used to test hypotheses about psychological traits that have their origins in evolutionary thinking.

[31] Precisely the same point holds for hypotheses that reflect a theorist's social values and political viewpoint, for example, hypotheses that have been put forward by theorists hoping to promote gender equality. As Eagly and Wood note: "One avenue to defusing nonproductive name-calling is to accept that the source of hypotheses should not be a central issue in science. Whether ideas come from political preference, observations of everyday life, intuition, or prior science, they should be scientifically tested, subjected to methodological critique, and replicated to test their generalizability beyond the initial demonstration. Therefore, the argument should not be about whether feminists and evolutionary psychologists possess values that influence their scientific activity. Surely they do. Scientists' values influence their choice of hypotheses and research methods as well as their interpretations of their findings. Nevertheless, debates should properly focus on the reasoning and research offered by these scientists, regardless of their political persuasions" (Eagly and Wood 2011, p. 760).

hunter-gatherers. Similarly, there is little question that ancestral hominids faced exposure to dangerous toxins and pathogens (including meat-borne bacteria and fungi) and that ancestral hominids, like other animals, had to deal with dangerous predators. Such facts about the ancestral world—and there are many like this—can guide psychological research by helping researchers to form valuable hypotheses about psychological systems that may then be tested for and examined in greater detail.[32]

What about the charge that evolutionary psychologists are panadaptationist in that they hold that nearly every psychological trait is an adaptation? In criticizing evolutionary psychology, it is common to point out that the fact that a hypothesis deriving from evolutionary thinking about the mind is supported by standard psychological tests does not by itself show that the psychological trait in question is an adaptation (Lloyd 1999; Downes 2018). This is true. For our purposes, however, what matters most about evolutionary theorizing about the mind is its role in generating rationalist hypotheses about the contents of the acquisition base, not whether the psychological structures posited through such theorizing are taken to be adaptations. What rationalism cares about is what psychological structures are in the acquisition base, not whether such structures are adaptations or not. Once a proposal that is based on evolutionary theorizing has been tested and has been found to be well supported using the same types of tests that are used to evaluate any other psychological hypothesis, the additional claim that the trait is an adaptation doesn't add anything to the case for a rationalist account of the trait.

If evolutionary psychologists have been prone to advance adaptationist hypotheses, it is not because they have failed to recognize that there are other processes besides natural selection that have had an influence on the innate structure of the mind. For example, an early and influential overview of evolutionary psychology's basic theoretical commitments includes the observation that "in addition to adaptations, the evolutionary process commonly produces two other outcomes visible in the designs of organisms: (1) concomitants or by-products of adaptations (recently nicknamed 'spandrels'; Gould and Lewontin 1979); and (2) random effects" (Tooby and Cosmides 1992, p. 62). And evolutionary psychologists have at times actively favoured alternatives of these types. For example, Pinker (1997) holds that music cognition is a byproduct, Boyer (2003) that concepts of supernatural agents in religious and spiritual thought are byproducts, and Kurzban et al. (2001) that racial thinking is a byproduct.[33] In any case, the

[32] For discussion of some examples, see Chapters 14 and 15.

[33] We will return to Kurzban et al.'s work on racial cognition later (see Chapter 15). For the moment, we will merely note that part of the reason to suppose that racial thinking isn't an adaptation is that ancestral conditions wouldn't have given rise to selection pressure for racial classification. This is because ancestral hunter-gatherers wouldn't have travelled sufficient distances in their lifetimes to encounter individuals with the type of systematically different superficial physical characteristics associated with being considered to be of a different race—another example of a reasonable

key point for us is that whether a hypothesis that posits a particular type of psychological structure as being part of the acquisition base sees this structure as an adaptation is simply irrelevant from the perspective of the rationalism-empiricism debate. All that matters is whether there are good independent grounds to suppose that the structure is in fact part of the acquisition base.

This leaves us with the charge that explanations in evolution psychology are so flexible that they can accommodate nearly any behavioural outcome. This last criticism rests on a misunderstanding of how evolutionary psychologists explain human behaviour. Evolutionary psychologists quite reasonably postulate systems that can produce different reactions, including contrary reactions, depending on contextual factors. These context-sensitive reactions are part of what make the cognitive architecture flexible, reflecting the differing pressures on fitness that ancestral hominids would have faced. A disposition for long-term pair bonding (given the right conditions) along with a disposition to seek sex outside of long-term partnership (given the right conditions) is no less explanatory than, say, a disposition to eat in some circumstances (one is hungry and the meat is fresh) and a disposition to not eat in others (one is hungry but the meat shows clear signs of being rotten, and so is dangerous to eat) (Kurzban 2002).

Finally, we should note that while we think that what really matters for rationalism isn't whether a given psychological trait is an adaptation but just whether it is part of the acquisition base, we think that adaptationist thinking can nevertheless be a valuable research tool, and the claim that a given psychological trait is an adaptation can sometimes be quite compelling. Consider again the case of species-specific food aversions, which came up in connection with the argument from animals. The claim that certain animals possess cognitive adaptations for acquiring a food aversion isn't needed to draw useful (rationalist) conclusions from the food-aversion studies reviewed in section 4.1. However, the overall pattern in the data makes an adaptationist explanation very plausible all the same—one that can be used to guide further research. And this is exactly how the research has played out. Insightful early commentators noticed that both the differential emphasis on taste in rats and the potentially long durations that pass between the "association" of stimulus and punishment make perfect sense from a biological and ecological perspective (see, e.g., Rozin and Kalat 1971). As generalist eaters, rats have to figure out which foods to avoid, and since the toxins in a food might not have an immediate impact, rats have to learn to avoid foods they may have eaten some time ago and not focus on the most immediate stimulus. A food-aversion system that functions in this manner would have considerable adaptive value. But notice that it would be useless for monophagous feeders—animals that eat only one type of food and consequently have little

evidence-based inference about the EEA that informs an evolutionary hypothesis but which plays no role in the testing of the hypothesis.

choice in the matter. This reasoning subsequently led researchers to make predictions about species that should, and species that shouldn't, be able to acquire food aversions, predictions that have held up under experimental testing (Ratcliffe et al. 2003).[34] Later, we will come across a number of similar examples that pertain to human psychology—examples in which evolutionary thinking has had a substantial payoff in vindicating rationalist views of the acquisition base.

Evolutionary psychology and evolutionary theorizing about the mind have been the subject of much acrimonious debate. Our conclusion here is not that all work under these banners is unproblematic—it is not—but only that such work isn't systematically undermined by the sorts of objections that we have been discussing and consequently that theorists should evaluate particular rationalist arguments and claims which involve an element of evolutionary theorizing using the same criteria as apply to rationalist or empiricist hypotheses that are not tied to evolutionary theorizing.

4.3 Conclusion

In this chapter, we have considered a range of fundamental challenges to rationalism as an approach to the origins of different sorts of psychological traits. These challenges are often thought to show that rationalism is so flawed that it can be rejected without having to examine in any detail particular rationalist theories or the evidence that proponents have offered on behalf of such theories. We have argued that none of these challenges systematically undermine rationalist accounts in this way. In fact, many of these arguments involve mistakes and confusions and in the end don't actually raise any substantive difficulties for rationalism at all. We have also seen that a strong preliminary case can be made for rationalism built around the poverty of stimulus argument and the argument from animals. The upshot of these considerations, we would argue, is that rationalism is not only a viable general theoretical framework, but one that is deserving of serious attention.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. DOI: 10.1093/9780191925375.003.0004

[34] As we saw in section 4.1, the monophagous feeders (vampire bats) weren't directly compared to rats in these experiments. They were compared to other species of bats, ones with varied diets. As a result, the adaptationist explanation made two substantive and correct predictions—that the vampire bats wouldn't develop the taste aversion and that the other species of bat would.

5
Abstraction and the Allure of Illusory Explanation

Our aim in Part I is to comprehensively rethink the rationalism-empiricism debate regarding the origins of cognitive traits. So far, in Chapters 2 through 4, we have outlined our positive account of how to understand rationalism and empiricism, we have addressed concerns about the coherence and value of the debate between rationalism and empiricism (particularly with reference to the global debate about the origins of psychological structures in general), and we have addressed a series of concerns regarding the viability of rationalist accounts within this global debate. Our discussion in these chapters has been aimed at establishing a theoretical framework for understanding the rationalism-empiricism debate that can make space for and encourage a productive debate between rationalists and empiricists regarding the origins of cognitive traits while addressing some of the biggest misunderstandings responsible for resistance to such productive engagement with the debate and to rationalist accounts in particular.

This chapter begins our shift toward focusing on the rationalism-empiricism debate as it applies to the origins of concepts in particular. The central aim of this chapter is to highlight and address a type of resistance to rationalist accounts that differs from those that we discussed in the previous chapter in that it tends to operate below the surface. The resistance stems from a cognitive bias that can illicitly lead to empiricist accounts seeming to be obviously correct. When this bias is active, rationalist accounts are often not seriously entertained as competing alternatives. The underlying bias responsible for this dynamic has to do with what we call the *allure of illusory explanations*. This refers to the tendency for explanations that are essentially vacuous to fail to be recognized as such in certain contexts.

Illusory explanations that paper over the complexity of a psychological capacity can often seem to be perfectly satisfactory in the context of the rationalism-empiricism debate, making alternative accounts appear to be needless and extravagant. We will see that historically this tendency has been a major factor in fostering an unwarranted presumption in favour of empiricism. To illustrate this point, we will work through a case study involving one of the most enduring ideas about conceptual development—the idea that much conceptual development is grounded in a process known as *abstraction*. We will show that historically

influential empiricist accounts of the origins of general representations—accounts that rely heavily on the process of abstraction—turn out to be illusory explanations. The chapter then turns to a brief digression from the main line of argument in Part I. Having argued that traditional accounts of abstraction offer only illusory explanations, we sketch the outlines of a new framework for understanding abstraction in which abstraction can offer a substantive account of conceptual development. We also briefly explore the question of what theories in this framework might tell us about the origins of concepts. One of our main conclusions is that the process of abstraction, once it is reimagined in this way, isn't uniquely suited to empiricist theorizing; it turns out to be equally compatible with rationalism.

5.1 Illusory Explanations of Cognitive Capacities

To get an initial sense of the problem of the allure of illusory explanations before we turn to its application in conceptual development, we will begin by looking at the problem in the context of the intellectual climate at the origins of the cognitive revolution and the beginning of contemporary linguistic theory. A persistent theme in much of Chomsky's early theoretical work involved highlighting the inadequacies of empiricist explanations of the origins of knowledge of language that were formulated in terms of linguistic habits being the product of processes seen as involving something like training, instruction, or conditioning.

Chomsky noted that although this type of account was held by many of the most prominent theorists of the time in linguistics (Bloomfield), philosophy (Quine, Wittgenstein), and psychology (Skinner), it was not really an explanatory account grounded in observations so much as a set of "a priori assumptions about what [these theorists] believe must take place" (Chomsky 1966, p. 144). Though this approach to explaining the origins of language was widely seen as being obviously correct, even truistic, Chomsky argued that, at best, the accounts offered were explanatory placeholders which, when subjected to scrutiny, turned out to be either clearly wrong or completely empty. Many of these explanations used quasi-technical terms like "generalization", "analogy", "habit structures", and "dispositions to respond" which gave the appearance of a substantive explanation, but such explanations were manifestly false when they were understood using the official meanings of the terms. Chomsky noted that one could attempt to salvage such claims by reinterpreting them so that they effectively say nothing more than that language is acquired through some process or other involving experience with language. In that case, the claim that language is acquired via training or instruction is essentially vacuous.

In perhaps the most famous discussion of these issues—in his (1959) review of Skinner's book *Verbal Behavior*—Chomsky showed that even the most detailed

and sophisticated attempt to develop this general approach to language faced this type of problem. Skinner's account was framed using the terminology of scientific behaviourism ("stimulus control", "response strength", "stimulus generalization", "history of reinforcement", and so on) which ostensibly picks out purely objective physically describable variables for explaining behaviour. However, Chomsky pointed out that if these explanations were understood in terms of the official account, the theory has no chance at all of explaining even the most basic facts about our knowledge of language. And in practice, the terms were used in so loose a manner that they had no substantive connection with either their technical or their ordinary uses. As a result, the explanations offered by the theory are simply illusory.

By way of illustration, consider Chomsky's discussion of a hypothetical case where a person is shown a painting by a Dutch artist. On Skinner's account, the painting acts as a stimulus that elicits a verbal response, and the utterance that ends up being produced depends on the speaker's prior history of reinforcement. This response is said to be under the control of the stimulus. In this situation, one might respond by commenting on the style of the painting, saying something like *Dutch!* But as Chomsky notes, one might equally instead say any of the following: "*Clashes with the wallpaper*, *I thought you liked abstract work*, *Never saw it before*, *Tilted*, *Hanging too low*, *Beautiful*, *Hideous*, *Remember our camping trip last summer?*" (Chomsky 1959, p. 31) or any of a multitude of other things.

The point is that, in any given situation, there is an enormous range of different things a person might say and consequently there is no substantive sense in which what is said can be understood as being controlled by the stimulus. The obvious way to see the situation is that what is said depends on what the speaker believes (about the painting, the situation, etc.), what the speaker's goals, preferences, and desires are, and of course the speaker's understanding of language. But understanding the situation in these terms is simply abandoning any pretence of understanding what is said as being in accordance with stimulus generalization, response strength, the speaker's history of reinforcement, or the other variables that Skinner's theory officially has at its disposal.

Similarly, Chomsky notes that on the model that one's utterances of the word "chair" are under the stimulus control of chairs, one might equally hold that utterances of "Eisenhower" and "Moscow" are under the stimulus control of Eisenhower and Moscow—despite the fact that most of us make such utterances without ever having any direct contact with either Eisenhower or Moscow. But if an utterance can be under the control of a stimulus that one has never encountered, and if (as we just saw) one might make pretty much any utterance in response to a given stimulus, then the terms "stimulus" and "response" are being used in a way that is completely disconnected from the scientific theory they derive from and that is ultimately supposed to account for linguistic behaviour. So, while these sorts of explanations were not only taken very seriously but, in

broad outlines, were taken to be obviously correct by the leading theorists of the time, they weren't really explanatory at all. They offered only an alluring but ultimately illusory form of explanation.

Although Chomsky's central concern was with the origins of language, similar issues arise for other aspects of cognition, including, as we will see, the origins of concepts. While work in psychology (including contemporary work in cognitive science) has not been immune to the allure of illusory explanations, this allure has been especially problematic in philosophy, particularly in approaches that rely heavily on introspection in developing explanations of psychological phenomena.[1] One consequence of this overreliance on introspection is an inability to recognize the complexity of the very phenomenon at issue and therefore to even attempt to capture this complexity in explanations of the phenomenon. The resulting superficial explanations can create the illusion of fully accounting for those aspects of cognition that are at issue.

An overreliance on introspection is problematic in developing explanations of psychological capacities because introspection provides essentially no access to most of the psychological activity that is involved in such capacities. Introspection is largely blind to the internal working of psychological capacities like vision, memory, language, decision making, reasoning, and categorization, as well as those involved in cognitive and conceptual development.

Consider, for example, the process of segmenting a spoken sentence into its constituent words. Introspection doesn't tell us *anything* about how we manage to do this. In fact, to the extent that it does tell us anything at all, what it tells us is very misleading. A common and natural supposition outside of cognitive science is that, just as there are spaces between words in a typed sentence, so there are spaces between spoken words in the stream of sound when people speak. But normally there are no such spaces. One way to see this is to listen to someone speaking in an unfamiliar language. Their speech will often sound like it is in "high speed" mode with no gaps at all. This is because spoken language isn't like the words on a page; there are no drops in acoustical energy to mark the word boundaries. So how do we manage to pick out each word? From the perspective of an ordinary speaker, relying only on introspection, the most that can be said is "I just do it".

As a statement about the phenomenology of language, this is unobjectionable. But comparable "explanations" have been given by some philosophers who have

[1] Unsurprisingly, lay explanations of cognitive development are also subject to the allure of illusory explanations. For example, a common type of lay explanation takes children to be like sponges "soaking up" information around them. In light of this kind of explanation, rationalist accounts can seem implausibly complex. But without an account of the sponge-like quality of young minds, we aren't actually being told anything more than that children are prodigious learners. The appearance of explanation here is merely illusory. Moreover, for all that such accounts say, it could be that the reason why children are able to soak up information so easily is precisely because they have a rich innate endowment that guides their learning in particular ways in a variety of domains.

taken them to be fully satisfactory, as if nothing more needs to be said. Striking examples of this kind can be found in Wittgenstein, one of the most influential thinkers of the last century. A major recurring theme in his writings is his disdain for efforts at explaining ordinary psychological phenomena by positing inner states and processes. In the opening section of the *Philosophical Investigations*, he famously says "Explanations come to an end somewhere", implying that we shouldn't suppose that psychological abilities (categorizing, reading, attributing thoughts to others, etc.) have explanations in terms of inner states and processes (Wittgenstein 1953, p. 2). He advises us to "try not to think of understanding as a 'mental process' at all" (Wittgenstein 1953, p.61). Later in the *Philosophical Investigations*, Wittgenstein considers and rejects the possibility of there being mental processes underlying memory and recollection. "'There has just taken place in me the mental process of remembering…' means nothing more than: 'I have just remembered…'" (p. 306; ellipses in original). Wittgenstein isn't merely remarking on what it is like to experience remembering something, the feeling that it just happens. In passages like these, he is warning readers not to be tempted by the desire for a deeper account—really, any account—as there is nothing further to say about the matter.[2] But, of course, if the explanation of cognitive development amounts to no more than "you just do it", accounts that posit a rich rationalist acquisition base and rationalist learning mechanisms (whose existence and operations are invisible to introspection) will seem excessive and extremely implausible.

One might have thought that these sorts of objections to scientific psychological explanations would no longer have much sway. However, contemporary philosophers in a number of influential philosophical traditions continue to be attracted to remarkably superficial explanations of psychological abilities. Consider John McDowell's remarks on cognitive development, which arise in the context of his influential and highly regarded views on concepts (McDowell 1994).[3] McDowell maintains that prelinguistic children do not have genuine concepts and thoughts and cannot engage in reasoning, since, in his view, these

[2] To the extent that Wittgenstein has an argument for this view, it's that the absence of conscious access to an internal process shows that such processes aren't real (e.g., at another point in the *Philosophical Investigations*, he writes: "I said that when one reads the spoken words come 'in a special way': but in what way? Isn't this a fiction?…Read the letter A.—Now, how did the sound come?—We have no idea what to say about it" (1953, p. 67)). However, it can hardly count as evidence against postulated unconscious mental states and processes that we have no conscious experience of them. If unconscious mental states and processes exist, then by their very nature, we *shouldn't* have conscious access to them. Wittgenstein is essentially in the same position as someone who rejects the existence of microscopic organisms on the grounds that they can't be seen by the naked eye. Given that the theories that postulate their existence are themselves telling us that microscopic organisms aren't visible to the naked eye, the fact that they can't be seen in this way is hardly a reason for concluding they don't exist.

[3] When McDowell makes these claims about concepts, he is adopting a different philosophical account of concepts than the one that we are using. However, this doesn't affect our criticism here. We discuss his alternative view of concepts (which takes concepts to be the meanings of natural language

require natural language.[4] So how then do infants make the transition from their prelinguistic condition to become rational adults with genuine concepts and thoughts? According to McDowell:

> This transformation risks looking mysterious. But we can take it in our stride if, in our conception of the *Bildung* that is a central element in the normal maturation of human beings, we give pride of place to the learning of language. In being initiated into a language, a human being is introduced into something that already embodies putatively rational linkages between concepts, putatively constitutive of the space of reasons, before she comes on the scene. This is a picture of initiation into the space of reasons as an already going concern; there is no problem about how something describable in those terms could emancipate a human individual from a merely animal mode of living into being a full-fledged subject, open to the world. (p. 125)

In short, McDowell's answer is that the rational relations between concepts that need to be acquired appear in language itself.

But this view is highly problematic. The explanation that McDowell's reply offers is essentially an illusory one as he fails to recognize that it immediately raises the question of how children are able to be "initiated into a language" in the first place (Laurence and Margolis 2012a). McDowell's account requires that children come to grasp the rational linkages in language despite not having any concepts or the ability to reason. But how could a child come to appreciate these rational linkages without engaging in some form of reasoning? The mere fact that language itself "embodies putatively rational linkages between concepts" cannot explain this accomplishment, since pet hamsters and potted plants—even ones whose human caretakers talk to them on a daily basis—do not learn language. If the noises, marks, and gestures in language convey reasons or exhibit aspects of rationality, they do so only for beings who are able to understand and appreciate these things.

Although McDowell claims otherwise, his account does nothing at all to remove the mystery of how children can learn language when it is assumed that they come to this task without genuine concepts, thought, and reasoning abilities. What's missing is any recognition that something has to be said about learners' minds and how they are able to appropriately process the purported rational linkages in language.[5] Any substantive explanation of how this is accomplished—or of how reading, memory, or virtually any cognitive capacity works—has to go

words and not mental representations) in Chapter 6 and show how the psychological issues about conceptual development can be reformulated given an alternative view of concepts of this type.

[4] McDowell is not alone in this view. See also Davidson (1975); Dummett (1993); and Brandom (2000).

[5] Similar issues in related philosophical work have been highlighted by Rey (2001).

deeper by providing an explanation of how these processes actually work. And since these workings are not accessible to conscious introspection, this means that they will invoke unconscious states and mental processes.

As it turns out, there is overwhelming reason to accept that much of the mind is inaccessible to conscious introspection. In cognitive science, unconscious mental states and processes function as theoretical posits that play crucial roles in detailed explanations of a wide range of everyday psychological abilities. It is no exaggeration to say that without unconscious states and processes, cognitive science as we know it would not be possible. Virtually every substantive explanation of a cognitive ability relies extensively and ineliminably on the supposition of unconscious mental activity.[6] As Dehaene (2014) notes regarding the unconscious processes underlying visual experience:

> what we experience as a conscious visual scene is a highly processed image, quite different from the raw input that we receive from the eyes. We never see the world as our retina sees it. In fact, it would be a pretty horrible sight: a highly distorted set of light and dark pixels, blown up toward the center of the retina, masked by blood vessels, with a massive hole at the location of the "blind spot" where cables leave for the brain; the image would constantly blur and change as our gaze moved around. What we see, instead, is a three-dimensional scene, corrected for retinal defects, mended at the blind spot, stabilized for our eye and head movements, and massively reinterpreted based on our previous experience of similar visual scenes. All of these operations unfold unconsciously—although many of them are so complicated that they resist computer modelling…At a glance, our brain unconsciously infers the sources of lights and deduces the shape, opacity, reflectance, and luminance of the objects [we see]. (p. 60)

With the discovery and acceptance of the cognitive unconscious, the study of the mind was revolutionized in much the same way that astronomy and geology were revolutionized by the discovery and acceptance of deep space (the fact that the universe is vastly larger than prescientific thinkers supposed) and deep time (the fact that the Earth and the universe are vastly older than prescientific thinkers supposed).

In some ways it is surprising how long it took the unconscious to enter into the intellectual mainstream and why it has received so much resistance in philosophy. Philosophers are known for their theoretical imagination and their willingness to entertain highly speculative conjectures in an effort to explain perplexing phenomena. The history of philosophy is replete with extraordinary theories—for example, that physical objects are really mental (Berkeley 1713/1975), that souls

[6] See, for example, standard textbooks discussing vision, language processing, learning, memory, categorization, and so on, such as, Frisby and Stone (2010); Gleitman et al. (2010); Gilovich et al. (2015); Schacter et al. (2017); Gazzaniga et al. (2019); and Baddeley et al. (2020).

and bodies function independently but in parallel in a "preestablished harmony" (Leibniz 1714/1965), or that there is a distinct non-physical realm where the eternal and unchangeable ideals of beauty and goodness exist as abstract entities (Plato 360 BCE/1992).

And yet despite this openness to unusual new ideas and to theories that are hardly commonsensical, the postulation of unconscious phenomena has often been deemed beyond the pale. For example, Locke (1690/1975) clearly thought that inaccessible psychological states verge on incoherence, a factor that played a major role in his argument against innate ideas:[7]

> it seeming to me near a Contradiction, to say, that there are Truths imprinted on the Soul, which it perceives or understands not; imprinting, if it signifies any thing, being nothing else, but the making certain Truths to be perceived. For to imprint any thing on the Mind without the Mind's perceiving it, seems to me hardly intelligible. If therefore Children and *Ideots* have Souls, have Minds, with those Impressions upon them, they must unavoidably perceive them, and necessarily know and assent to these Truths, which, since they do not, it is evident that there are no such Impressions… To say a Notion is imprinted on the Mind, and yet at the same time to say that the mind is ignorant of it, and never yet took notice of it, is to make this Impression nothing. No Proposition can be said to be in the Mind, which it never yet knew, which it was never yet conscious of. (I.ii.5)

Likewise, consider Berkeley's dismissive response to Descartes' proposals regarding depth perception. In the *Optics*, Descartes made the prescient suggestion that perceived distance in vision is based on a form of geometrical reasoning "quite similar to that used by surveyors when they measure inaccessible places by means of two different vantage points" (1637/1985, p. 170). In this case, the two vantage points are given by the positions of the eyes: "When our two eyes A and B are turned towards point X, the length of the line AB and the size of the two angles XAB and XBA enable us to know where the point X is" (1637/1985, p. 170). Berkeley would have none of this. The problem, as he saw it, was that no one is aware of the processes of geometrical reasoning that Descartes claimed to be involved in visual distance perception:

> I appeal to any one's experience, whether, upon sight of an object, he compute its distance by the bigness of the angle made by the meeting of the two optic axes? Or whether he ever think of the greater or lesser divergence of the rays, which arrive from any point to his pupil?… Every one is himself the best judge of what he perceives, and what not. In vain shall all the mathematicians in the

[7] See De Rosa (2004) for an argument that Locke's case against innate ideas is unsuccessful precisely because he fails to take seriously the possibility of unconscious mental states.

> world tell me, that I perceive certain lines and angles which introduce into my mind the various ideas of distance; so long as I myself am conscious of no such thing. (Berkeley 1709/1975, p. 10)

In hindsight, however, there is no question that Berkeley got this wrong and that the source of his error was the fact that he readily inferred that these representations couldn't be involved in computing distance if they weren't consciously accessible in introspection. There is now overwhelming evidence that depth perception is resolved, in part, by the sorts of factors that Descartes had identified. As Carey (2009) has remarked in discussing this exchange between Berkeley and Descartes, "there is hardly any classical debate from the history of philosophy of mind that has been more conclusively settled" (p. 31).

Scepticism about unconscious states and process among philosophers never fully dissipated. In some ways, it only intensified in the twentieth century in reaction to Chomsky's rationalist proposals regarding unconscious rules of grammar.[8] And while most philosophers of mind and cognitive science today are happy to acknowledge the existence of unconscious states and processes, there remains continued resistance to theories that postulate them among some very influential contemporary philosophers.

Much of our discussion in this section has focused on philosophy and on historical examples of the allure of illusory explanation. Such examples can be particularly useful ones to consider because it can be easier to recognize this bias in hindsight with the benefit of some additional theoretical distance and because some of these examples offer especially clear illustrations of the bias. This is not to say, however, that contemporary scientific accounts in cognitive science are immune to this bias. They aren't.

One place where we take these kinds of issues to be in play is in connection with the way that the technical concept of an *affordance* sometimes gets used. An affordance is a potential way that an organism can interact with an object owing to both the features of the object and the organism's body (Gibson 1979). For

[8] For example, in a discussion that strongly echoes Locke, Goodman (1967) offered a scathing critique in which he proclaimed that the very idea of mental states that can't be brought to conscious awareness is unintelligible:

> A: . . . Now I gather that the theory here proposed is that certain ideas are implanted in the mind as original equipment.
> J: Roughly that.
> A: And being ideas, they are in consciousness?
> J: No, not necessarily; not even usually.
> A: Then they are in the subconscious mind, operating upon cognitive processes, and capable of being brought into full consciousness?
> J: Not even that. I may have no direct access to them at all. My only way of discovering them in my own mind may be by the same methods that someone else might use to infer that I have them, or I to infer that he does.
> A: Then I am puzzled. You seem to be saying that these innate ideas are neither innate nor ideas. (pp. 27–28)

example, a chair has the affordance of *sitting-on* (being *sitting-on-able*) for most adults, given the shape, size, and sturdiness of typical chairs and the normal range of motion and postures that people can comfortably undertake. There is nothing wrong with this notion of an affordance in itself. However, there is a danger in relying on affordances in explanations of cognition, given how easy it is to slip from the affordance being present—in the sense that this type of organism-object interactions is biomechanically possible—to a much more loaded sense in which the presence of the affordance means that such interactions are recognized for what they are simply by virtue of the objects themselves being perceived.

A young child's shoelace may have the affordance of being tie-able, and the child may want to secure her shoe and have the biomechanical potential to draw the lace ends together to form a knot. But that doesn't mean that she can mentally represent the tie-ability of the laces or that she knows how to tie them. Likewise, a juvenile chimpanzee may want the termites that are encased in a log and have the biomechanical potential to retrieve them using a twig. But that doesn't mean that the termite-extraction affordance of a twig is part of its psychology, since juvenile chimps haven't yet figured out how to use them in this way. Affordances of these types are often seen as explaining how an agent comes to possess a cognitive capacity, but on their own, they don't explain anything at all. The illusory explanation here is fundamentally of the same type as that involved in McDowell's illusory explanation of language learning. An agent's ability to recognize an affordance that an object offers cannot be substantively explained merely by the fact that the object has that affordance. The real explanation only begins when an account is given of how such affordances come to be recognized and appreciated.[9]

To be clear, we are not saying that the problem of illusory explanations is a problem for all empiricist accounts. It isn't. And we will be discussing many substantive empiricist explanations later in the book. However, the allure of illusory explanations has historically been a persistent issue affecting empiricist accounts, and has been a significant factor when it comes to understanding why rationalist explanations of many cognitive phenomena haven't been given serious consideration.[10] Given how easy it has been for even the most gifted philosophers to fall prey to the allure of illusory explanations, it is important to consider an example

[9] In later chapters, we will see further examples of seemingly explanatory accounts which in reality only provide illusory explanations in connection with a number of current large-scale empirical research programmes in cognitive science (see Chapters 12 and 21).

[10] It's worth noting there are also contexts outside of the rationalism-empiricism debate where illusory explanations have proven to be tempting. One of these is in the context of explanations that incorporate neuroscientific details. When people are given a choice between two explanations for a psychological phenomenon only one of which includes neuroscientific details, the explanation that includes them is often found to be more satisfying—even when the neuroscientific details are completely irrelevant and so don't contribute substantively to the explanation (Weisberg et al. 2008). A similar pattern has been found regarding people's assessment of other types of scientific explanations, including ones in biology and chemistry. More reductionistic explanations are often preferred even when the reductive information does not contribute substantively to the underlying logic of the explanation (Hopkins et al. 2016).

that is more directly connected with the origins of concepts, the main question in this book. The example we will discuss is of both historical and contemporary interest—offering an especially clear case in the history of empiricist thought, while at the same time, involving a type of psychological process that can be reimagined in a way that is directly relevant to ongoing research in cognitive science.

5.2 Abstraction as a Theory of the Origin of General Representation

Our central case study of an illusory explanation of conceptual development involves what is perhaps the single most important type of positive empiricist account of conceptual development in the history of philosophy. It has dominated philosophical thinking about conceptual development for centuries, with variants having been proposed by virtual every major empiricist philosopher, including John Locke (1690/1975); Bishop George Berkeley (Berkeley 1710/1975); David Hume (1739/1978); Thomas Reid (1785/2002); John Stuart Mill (1882); William James (1890); and Bertrand Russell (1912), among others. The fact that the key features of this account have persisted through generation after generation of leading philosophers, and the fact that it was absolutely central to their empiricism, highlights just how difficult it can be to recognize an illusory explanation for what it is and how strong the allure of such explanations can be.

The account in question is that certain core aspects of conceptual development are made possible through a process known as *abstraction*. What is abstraction? Roughly speaking, it is a form of perceptual learning that is supposed to explain how general representations can be learned through observation. For example, it might be thought that a colour concept like RED is acquired via abstracting the concept from perceptual experiences of a number of red objects.

Arguably the most important discussion of abstraction is in Locke's (1690/1975) *An Essay Concerning Human Understanding*. As we will see, Locke's account is deeply flawed, but his discussion is nonetheless singularly illuminating and highlights issues of continuing contemporary relevance. Distinguishing general ideas (representations that denote *types* of objects or events) from particular ideas (representations that denote specific individuals), Locke famously claims that abstraction doesn't just explain where *some* general representations come from. It is meant to be the source of *all* general representations.[11]

[11] In this section, we will often use the term *general representation* in place of Locke's term *general idea*. This term is intended to capture both conceptual and nonconceptual general representations (and likewise, we will extend our use of the small caps notation to cover all general representations for this section). We discuss the distinction between conceptual and nonconceptual representations in the next chapter. The conceptual/nonconceptual distinction is not one that Locke made, and

According to Locke, abstraction is the power of mind that involves "separating [Ideas] from all other *Ideas* that accompany them in their real existence; this is called *Abstraction*. And thus all its General *Ideas* are made" (1690/1975, II.xii.1). Locke gives several examples that are meant to illustrate the workings of abstraction. Regarding the origins of the general representation WHITE, he writes:

> the same Colour being observed to day in Chalk or Snow, which the Mind yesterday received from Milk, it considers that Appearance alone, makes it representative of all of that kind; and having given it the name *Whiteness*, it by that sound signifies the same quality wheresoever to be imagin'd or met with; and thus Universals, whether *Ideas* or Terms, are made. (II.xi.9)

The claim is that a general representation for a simple quality is formed by (in some sense) leaving out specific details about where and when it originated, as well as other ideas that may have initially accompanied it. Later, Locke discusses a different kind of example—the formation of a complex idea. He suggests that children may acquire MAN by first attending to particular individuals, such as their nurse or mother, and later observing that other things resemble them. This leads children to:

> frame an *Idea*, which they find those many Particulars do partake in; and to that they give, with others, the name *Man*, for Example. And *thus they come to have a general Name*, and a general *Idea*. Wherein they make nothing new, but only leave out of the complex *Idea* they had of *Peter* and *James*, *Mary* and *Jane*, that which is peculiar to each, and retain only what is common to them all. (III.iii.7)

There is debate as to how best to interpret Locke's remarks about the nature of abstraction and even whether he has a single account. This is understandable, since there is some unclarity about whether Lockean general ideas are formed by retaining the full representations associated with the particular objects that an agent perceives. To some readers, it sounds like the full representations *are* retained and that abstraction involves attending to certain features as opposed to others. However, to other readers, there is the suggestion that an abstract idea may involve the construction of a new representation, one that takes some features from the representations of particular objects while omitting others.[12]

unsurprisingly his use of the notion of a general idea blurs the distinction as it is variously understood by contemporary theorists. For present purposes, the key point is that Locke sought to explain the origins of *all* general representations via abstraction. For Locke, as for many other empiricists, *generality*—that is, the ability to represent what particular things are taken to have in common—is not something that is built into the mind. It must be learned.

[12] See Mackie's (1976) description of abstraction as *selective attention* and Dancy's (1987) contrasting claim that *abstraction is subtraction*.

Regardless of what the right story is about Locke, it is clear that he views abstraction as a process that is grounded in perception and that operations on the representations resulting from contact with particulars are the source of the ability to represent far more than the items that were originally perceived—not just this white paper but all white objects, not just this man but all human beings, and so on.

But how exactly can abstraction be the source of literally all general ideas? To see the force of this question, we need to step back and consider more carefully what input gets the process going. If abstraction is to explain the origins of all general representations, what kinds of representations can it draw upon and how do they depict the particular objects that an agent perceives? We will argue that there are four models of the representational input that are available to Locke, but that when these models are fleshed out, it is clear that none of them can provide a satisfactory account of the origins of all general representations—and that reflection on these models shows how accounts like Locke's, despite their enormous influence, turn out to provide nothing more than the illusion of explanation.[13]

To simplify the discussion, we will suppose that the general representation whose acquisition we are trying to understand is WHITE and that the experience from which it is abstracted is the visual perception of a snowball (or a number of snowballs). We can now rephrase the issue as identifying how the snowball is initially represented so that WHITE can be acquired via abstraction from the experience. We will consider the four potential models in turn.

1. Individual-representations and feature-representations. The first model takes as input a combination of individual-representations (i.e., representations which function like names or demonstratives and represent individuals *as* individuals) and representations for each of the salient features of the experienced particular. Thus the snowball might initially be represented with such representations as THAT, COLD, SPHERICAL, and SOLID.

This model faces a number of problems, but the most serious is that it simply presupposes that the process of abstraction takes as input *general representations*. This clearly won't do if the goal is for abstraction to explain the acquisition of *all* general representations, as the supposition that there are prior general representations will lead to a regress. Moreover, *colour* will undoubtedly be among the

[13] As we will see, these considerations also suggest that not only is it the case that abstraction cannot plausibly be the source of all general representations, but that it is highly unlikely that *any* learning process could be the source of all general representations. If an organism has any general representations at all, then, in all likelihood, some of these must be innate. We should emphasize that our argument for this claim is intended as an inference to the best explanation, not a proof. We do not claim that it is logically impossible for all general representations to be acquired without there being some innate general representations. Rather, our point is that exceedingly austere empiricist models incur prohibitive explanatory costs.

salient general features of the snowball that would comprise the input to the acquisition process, and it would presumably be the perception of its colour that would support the acquisition of WHITE. But then the process of acquiring WHITE would depend upon prior representations that include, among others, the representation WHITE. The model is plainly circular. It ends up saying that WHITE is the product of a process that takes WHITE as its input.

2. Individual-representations only. In order to address the problem with the previous model, one might suppose instead that particulars are initially represented only by individual-representations without any general representations coming into it until abstraction has taken place.

We don't know of any empiricists who have proposed a model of this kind, however, and for good reason. The problem is that individual-representations alone don't provide enough information to get the process of abstraction going. If particular objects are represented simply as individuals, without representing any of their features, then the input just isn't rich enough. After all, with the paradigmatic individual-representations—demonstratives—the whole idea is that they represent their referents directly, conveying no information about what the represented objects are like. But if all the mind has to go on in representing two white objects is THIS and THAT, it would have no basis for cognitively grouping the two together and certainly no basis for bringing them under a specific general representation such as WHITE. By only representing the individual objects as such, the initial representations would effectively leave the agent representationally cut off from all the features of the objects.

Suppose, however, that we overlook the question of *why* different individual-representations get grouped together cognitively and simply allow that they are. Then a number of individual-representations could be combined, yielding a representation like THIS AND THIS AND THIS (with each 'THIS' referring to one of three different white snowballs). Still, the resulting representation wouldn't do for two reasons.

First, it lacks the representational breadth of WHITE. WHITE covers the full scope of white items and has open-ended application (including application to the many white objects that the agent hasn't and won't ever encounter), whereas the conjoined individual representations only pick out the particular objects that have been encountered. Second, it fails to single out the relevant feature that these objects have in common (whiteness, as opposed to, for example, sphericality, coldness, snowballness, etc.). It's one thing to represent whiteness (or to represent white things in general) and quite another to represent a number of perceived objects that happen to be white. But it's the general representation WHITE that we are after, not a representation of several things that, as it turns out, happen to be white. No finite conjunction of individual-representations of white things would constitute a general representation of whiteness. And, of course, if

abstraction requires an infinite conjunction of individual-representations, then it is not the type of process that finite creatures like ourselves could ever accomplish—no general representation would ever actually be acquired.

3. Tropes. The third possibility for what the input might be to a process of abstraction that is taken to be the source of all general representations is that this input consists of representations of what in contemporary philosophy are called *tropes* (Daly 1994).[14] Tropes are somewhat counterintuitive for many people, but we can get at the basic idea in the following way. There are two aspects to what makes something a property (e.g., the property *greenness*, which a given leaf might have). One aspect is that properties constitute features of the object that have them (the leaf's greenness). The other aspect is that properties are the kind of thing that are in principle shareable. Other objects (another leaf, a grape, a blade of grass, etc.) might also have the same property. Tropes have the first aspect of properties in that they constitute features of particular objects, but they lack the second in that, by their very nature, they cannot be shared or possessed by different objects. On the trope view, where the greenness of a particular leaf is a trope, this trope constitutes the feature of the leaf's being green, but the trope that is the leaf's greenness is a feature that cannot—even in principle—be something that any other object has. This is not merely because no other particulars happen to have that feature (that particular shade of green), but because by its metaphysical nature a trope just is the kind of thing that can only be possessed by a single individual—tropes aren't multiply instantiable. So, on the trope view, when there are two green leaves, each possesses a trope which makes it green, and the trope that each possesses is utterly unique. It's not something that is (or could be) instantiated by any other leaf.

Returning to the snowball example, the proposal is that the input to the process of abstraction includes a representation of the snowball's whiteness, where this is taken to be a trope that is inherent to the snowball; no other object can share this very whiteness. This model might be thought to combine the best elements of the previous two models without having any of their drawbacks. This is because this model restricts the input to the process of abstraction to representations of individuals and contains no representations of general features of objects (tropes being abstract *individuals*). At the same time, however, it offers the hope that the agent is no longer cut off from representing the features of the particulars she perceives, since it does in a way represent the features of objects. It is just that these features are not general features in that they can be possessed only by the single individual that has them (tropes being *property-like* entities). In this way, the proposal aims to ground abstraction in the representation of features while

[14] Historically, perhaps the most famous advocate of abstraction as grounded in the representation of tropes was Reid (1785/2002).

at the same time avoiding any general representations being illicitly smuggled into the foundations of the acquisition process.

Unfortunately, promising as this suggestion may initially appear, tropes don't help. The problem is that the whiteness of a given snowball is constituted by its possession of a trope that constitutes the whiteness of this snowball and no other. A second snowball's whiteness is constituted by something else entirely—its possession of a *different* trope constituting *its* whiteness (even if the two snowballs are the very same shade of white). So, to represent the whiteness of two white objects, an agent would have to deploy two distinct representations, WHITE$_1$ and WHITE$_2$, to represent these two whiteness tropes as such. Because these representations are essentially of individuals (namely, the two tropes), this gives rise to exactly the same sorts of difficulties that arose for the previous model. In particular, there is a question about why these individuals (the two whiteness tropes) are to be grouped together and how representing them together yields a fully general representation as opposed to one that merely picks out the individuals that have been encountered thus far.[15] And, just as with representations of conjunctions of particulars in the previous model, any representation of the form WHITE$_1$ AND WHITE$_2$ (for any finite number of conjuncts of this sort) will always fall short of representing the open-ended character of WHITE. Once again, it looks as if we need a richer source of input if we are going to explain how general representations are acquired.

4. Generality without discrete representations. Thus far, we have considered three general approaches to the question of what representations might ground the process of abstraction: (model 1) approaches that take a combination of representations of individuals as such and representations of features as such as input, (model 2) approaches that take only representations of individuals as such as input, and (model 3) approaches that take as input representations of particularized properties (tropes) as such. These come close to exhausting the options that ought to be considered. However, one further possibility is that more complex metaphysical entities than individuals and features

[15] It may be tempting to think that some headway can be made on the question of why tropes are grouped together by holding that the agent also represents the similarity among these tropes. However, this approach flounders as soon as one considers the question of how this similarity would be represented. If it is represented through employing a general representation of the feature of similarity, this would make the approach circular (as it would be taking general representations to be part of the input to the abstraction process, as in the first model we considered). On the other hand, if it instead employed further trope representations—for example, a trope representing the similarity between the trope constituting the whiteness of the first snowball and the trope constituting the whiteness of the second snowball—then we would again fall short of representing anything approaching the open-ended character of WHITE, instead merely representing the particular feature of the similarity of these two tropes. And, further iterations of this type of strategy (e.g., attempting to represent the similarity of this similarity trope to another similarity trope) would only lead to a pernicious regress of trope representations (Laurence and Margolis 2012b).

figure in the input and are represented as such—something akin to events or states of affairs.

The initial representations that form the input to the abstraction process in this case might be taken to be unstructured representations that manage to pick out these more complex entities without any components representing any objects, properties, or tropes that are present in the event. For example, a snowball might be represented as being cold, spherical, *and white* but without separate representations corresponding to each of these features. The snowball's being cold, spherical, and white would be represented by a single unstructured representation (THIS-IS-COLD-SPHERICAL-WHITE), not by a structured representation composed of distinct representations capable of independently representing the object and these several features (THIS, COLD, SPHERICAL, and WHITE). In this way, WHITE wouldn't have to be a precursor to abstraction, nor would there have to be prior access to any other general representations corresponding to a particular's features.[16]

Although this model isn't obviously circular or inherently problematic for relying on input that is manifestly too austere, it won't do either. One problem with the model stems from the *productivity* of human cognition—the fact that our minds can represent an indefinite number of distinct combinations of features. The best explanation of the productivity of human thought is that discrete mental representations are combined and recombined in accordance with a compositional semantics, where the meanings of complex representations are a function of the meanings of representations that they are composed of and their manner of combination.[17] However, the model under consideration (generality without discrete representations) is built on the assumption that the representational system doesn't have the compositional structure that this explanation requires. Instead, for each new combination of features attributed to an object there would have to be a corresponding new and unique primitive (that is, unstructured) representation. Taking such representations to be foundational, however, is singularly implausible given the sheer number of such representations that would be required to represent even a very small sample of what we can represent. Since for any n features there are 2^n possible combinations of these features, this means that in order to represent a single object and just one hundred basic features and their combinations which it might possess—a grossly simplifying assumption—there would have to be 2^{100} distinct representations. That's roughly 1,250,000,000,000,000,000,000,000,000,000 distinct representations—about 2.5 trillion times more representations than there have been seconds in the

[16] An account involving something like this kind of model seems to have been suggested in James (1890).

[17] See Chapter 6 for further discussion of productivity and compositional semantics in relation to theories of concepts.

history of the universe.[18] The truly staggering number of primitive representations at play is enough to undermine any model that relies wholly on unstructured representations.

But the problem with this model isn't just the sheer number of primitive representations that it would require. The real problem is with how it could account for our ability to acquire WHITE from such representations as THIS-IS-COLD-SPHERICAL-WHITE without a representational basis for homing in on just the whiteness in the experience. To mentally focus on whiteness itself would seem to require the prior ability to represent whiteness as such, but this amounts to helping ourselves to the general representation WHITE. Once again, the account in question turns out to be circular. It cannot explain how the system could derive a representation of WHITE from the input without presupposing that the system already has the ability to represent whiteness.

Locke took it to be simply obvious that abstraction explains conceptual development and that it is the source of all general representations. "That this is the *way*, *whereby Men first formed general* Ideas, *and general Names to them*, I think, is so evident, that there needs no other proof of it, but the considering of a Man's self, or others, and the ordinary proceedings of their Minds in Knowledge" (1690/1975, III.iii.9). But by attending to the details of how such a process might work, we can see that it is anything but obvious how abstraction could play such a role. While abstraction may seem to offer an explanation for the origin of all general representations, it turns out that all four of the options for how abstraction might actually work abysmally fail to explain how it could do this. The failure is so profound that while it initially appears as if abstraction can provide such an explanation, this appearance turns out to simply be illusory. In each case, the account either presupposes elements that it claims to explain the origin of, or it simply does not have the resources to even begin to provide a substantive explanation of the explanatory target.

While our discussion has focused on Locke's account, it is important to note that Locke was not alone in failing to appreciate the sorts of difficulties that we have been pointing to in which the explanation involving abstraction turns out to be illusory. It is just a particularly illuminating example to consider. Famously, Locke's account of abstraction was rejected by Berkeley, and by Hume as well (largely based on Berkeley's vigorous criticism of the account). However, Berkeley and Hume did not reject Locke's account for the sorts of reasons that we have been pointing to. In fact, despite their spirited critique of Lockean abstraction, the alternatives to abstraction embraced by Berkeley and Hume turn out to face

[18] This comparison is based on the supposition that the universe is around fourteen billion years old (which is roughly 450,000,000,000,000,000 seconds).

much the same sorts of problems as Locke's account regarding the input to the process of abstraction and thereby the vacuity of the resulting account.[19]

Consider Berkeley's own theory of the origins of general representations. According to Berkeley, a general representation arises as an image becomes used to represent a range of particulars that are similar to the one that the image initially picks out. In this way, a representation that is initially particular can become general. Berkeley gives as an analogy a drawing of a line in a geometrical proof. Although the line may be 1 inch long, it comes to represent all lines, not just 1 inch lines, because the proof doesn't turn on its particular length:

> And, as that particular line becomes general by being made a sign, so the name *line*, which taken absolutely is particular, by being a sign, is made general. And as the former owes its generality, not to its being the sign of an abstract or general line, but of all particular right lines that may possibly exist, so the latter must be thought to derive its generality from the same cause, namely, the various particular lines which it indifferently denotes. (Berkeley 1710/1975, introduction, §12)

Hume described Berkeley's treatment of general representation as "one of the greatest and most valuable discoveries that have been made of late years in the republic of letters" (1739/1978, I.i.7). But despite this high praise, it's hard to see how Berkeley's account is any improvement at all on Locke's. Basically, we are told that an image achieves generality because it is *used* as a general representation. An agent starts out with an image of a particular and goes on to enlist it to reason about other things by ignoring irrelevant aspects of the image and focusing on just the relevant ones. The problem with this account becomes apparent when we ask *how* the mind manages to achieve this feat.

[19] Though it is not relevant to the main point that we are making here, it is worth pointing out that from a contemporary perspective, Berkeley's criticisms don't cut very deep in any case, since an advocate of abstraction can simply drop the assumptions that these criticisms turn on. Berkeley attacks Locke's construal of ideas as mental images and the view that these images can only represent what they resemble (Berkeley 1710/1975). Among other things, Berkeley points out that images are determinate in ways that bar them from achieving the generality that Locke requires. For example, you can't have an image of a generic man that represents men in general. To be recognizable as an image of a man, it would have to include specific details (e.g., size, shape, and colour) that might be true of some men but not of others. However, a contemporary advocate of abstraction needn't be committed to the view that concepts or ideas are mental images or to the view that resemblance explains representation, not even for the representations that subserve perceptual processes. We mention this in part because in the next section we will argue that there is a way of reconceiving abstraction, or at least an account that we think is still deserving of the name of abstraction, which can provide an important type of mechanism for the acquisition of at least some concepts—though, as we will see, this type of account would necessarily have to abandon Locke's explanatory project and it would no longer be distinctively empiricist. Nonetheless, given the view we will defend in the next section, it is useful to see that Berkeley's criticisms of abstraction lose their force when the assumptions they turn on are rejected.

Suppose the image is of a specific snowball that a child has just seen and that she ignores the depicted shape and texture, among other things, in the service of thinking about white things in general. To do this, she needs to selectively attend to the colour in the image. Yet Berkeley tells us nothing about how he proposes to account for the ability to selectively attend to certain aspects of an image while ignoring others. In order to psychologically focus one's attention on whiteness, one must, in effect, represent whiteness. But in order to do this, the options are essentially those we considered above for the Lockean account. Representing only particulars, whether concrete particulars or tropes, doesn't allow one to attend to whiteness as such. Employing a general representation of whiteness would of course allow one to attend to whiteness, but that would require prior possession of the general representation WHITE and hence reintroduce the problem of circularity. And general representations aren't really an option for Berkeley anyway, since the whole point of his treatment of generality is that it is supposed to do away with fully abstract general ideas.

The situation for Berkeley isn't relevantly different than the situation for Locke.[20] What's remarkable about all of this is that such explanations have been

[20] Is there *any* type of account that could provide a substantive psychological-level account of the origins of all general representation? Our view is that such an account is unlikely and that a large part of the problem with the accounts of abstraction that we have been discussing—and the fundamental reason why they fail to provide substantive accounts of the origins of all general representations—is that at least some general representations have to be available to get any such acquisition process going. In other words, some general representations should be supposed to be innate.

Of course, there is always the possibility that there might be some further model of how abstraction gets started that we have not considered, one that can (somehow) account for the origins of all general representations. One possibility, for example, is an Aristotelean model where sensible forms are taken to be literally transmitted from an object through a perceiver's sense organs into the mind. Adams (1975) succinctly describes such a view as follows: "Perception was interpreted as a transaction in which a form (the sensible form) is transmitted from the perceived object to the perceiver…There is something (the sensible form) which literally comes into the mind from the object. This theory of perception is the basis for the Aristotelian empiricist answer to the question, how we get our ideas" (p. 73). For contemporary theorists, this approach isn't at all attractive as it treats perception as a nearly magical process. And in addition, it also faces a number of potent objections (Annas 1992). For example, if the shape of an object is literally transmitted from the object to the mind, why do objects exhibit perspectival differences in appearance (as when a coin has a circular appearance when viewed from above but an elliptical appearance when viewed from the side)? Why don't we see the true colours of objects in dim lighting? And why do we sometimes mistakenly take an object of one type to be of a different type?

There is also, of course, the option of abandoning the idea that the acquisition process is representational, taking all general representations to be acquired via a wholly non-psychological process. If we do that, then the input needn't include any representations at all, much less general ones. All that is required are causal interactions with property instances. We will see later that a proposal along something like these lines has been made by Fodor (2008) as a general account of the origins of concepts. As theories of the origin of all general concepts, we think that such non-psychological accounts are decidedly unattractive. They effectively postulate mysterious neurological processes that inexplicably yield content-appropriate general representations simply on the basis of causal contact with the world. To explain the origin of all general concepts in this way is extremely implausible (see Chapter 26 for discussion of Fodor's proposal and the difficulties it faces). On the other hand, it seems entirely reasonable that there should be a non-psychological account of the origins of at least some general concepts—indeed this is effectively what it means to say that some general concepts are innate on the

put forward as though they were providing substantive developmental accounts of the origins of general representation and have been endorsed—and lauded—by generation after generation of empiricist philosophers who simply failed to notice that in fact such accounts only create the illusion of an explanation. And while the example of abstraction as an explanation of the origin of all general representations allows for a particularly nice illustration of this phenomena, we saw in the previous section that illusory explanations are not restricted to this case study. The targets of Chomsky's critique were just as eminent and just as susceptible to illusory developmental explanations as the advocates of abstraction we have discussed here.

5.3 A New Framework for Theories of Abstraction

While the accounts of abstraction that we have just been looking at offered nothing more than the appearance of explanation, we don't think that this means that the idea of explaining aspects of conceptual development by a process something like that of abstraction is completely misguided or that it should just be abandoned altogether. In fact, we think that abstraction can actually be turned into a powerful account of conceptual development that can explain the origins of an important class of concepts. However, doing this—developing an account of abstraction that provides a substantive developmental explanation—requires shedding many of the details that empiricist advocates of abstraction have found most attractive about abstraction. In particular, we need to abandon some of the empiricist aspirations that have long been associated with abstraction, along with the claim that it can provide an explanation of the origin of all general representations.

In this section, we will present a new way of understanding abstraction that does just this. We believe that accounts based on this new understanding are still worthy of the name "abstraction" and that they can provide an important part of the story of how concepts are acquired. Spelling out this new understanding of abstraction and exploring some of its consequences will involve a small digression from our main line of argument in Part I. But the digression is of direct relevance to contemporary theories of conceptual development and is valuable both for the light it sheds on the space of options available in the rationalism-empiricism debate and for its instructive contrast with the illusory explanations provided by traditional accounts of abstraction. Rather than develop any particular version of abstraction based on this new understanding, our aim instead will be to sketch the broad outlines of a framework in which many specific accounts can be developed.

account we gave in Chapter 2, as such representations would be part of the acquisition base in virtue of not being acquired via a psychological-level acquisition process.

The core idea behind our framework is that abstraction-based accounts involve a developmental process of moving from relatively specific—but nonetheless general—representations as input (e.g., a representation for a given shade of colour or a narrowly circumscribed type of shape) and delivering representations with a greater degree of generality as output (e.g., broader colour or shape representations such as RED or TRIANGULAR). The relative specificity of the input representations is able to capture the particularity of the represented qualities in experience—what is often called the fine-grainedness of perceptual experience. While the output representations can be seen as "abstracting" away from the particularities of the individually experienced colours, shapes, and so on, through being comparatively *more general* representations. This general approach to how the process of abstraction works is based on a suggestion that we take from W. V. O. Quine's treatment of learning in his paper "Natural Kinds" (Quine 1969b). Though we reject many elements of Quine's account of learning, we think that it contains an important kernel that can be adapted and expanded in various ways to provide a promising basis for understanding abstraction. For this reason we will refer to the framework we develop as a *neo-Quinean framework* for understanding abstraction.[21] This framework makes it possible to explain how abstraction can account for the origins of many concepts without falling prey to the difficulties associated with the accounts of abstraction discussed in the previous section.

Let's start by looking at Quine's account, which is couched in terms of an account of word learning rather than as an account of concept learning. His account has three main components. First, Quine assumes that the learner can innately discriminate a range of fine-grained properties in the learning domain, for example, different shades of colour in learning colour words like "white" and "green". These fine-grained discriminatory capacities provide the input to the process of abstraction. By building generality (albeit fine-grained generality) in from the outset in the form of general capacities for discriminating shades of colour, Quine dramatically limits the scope of his account in comparison with the empiricist philosophers we discussed in the previous section. He doesn't take abstraction to explain the origin of *all* general discriminatory capacities. Nonetheless, for Quine abstraction can explain how a general word like "white" could be learned on the basis of the fine-grained discriminatory capacities associated with particular shades of colour.

The second component of Quine's account is a similarity metric. Quine assumes that the fine-grained discriminatory capacities are innately ordered in terms of similarity (an innate "spacing of qualities"), which he interprets behaviouristically. "A standard of similarity is in some sense innate. This point is not

[21] We should note, however, that Quine doesn't describe himself as offering a theory of abstraction.

against empiricism; it is a commonplace of behavioral psychology" (1969b, p. 123). Quine's innate similarity metric incorporates a further element of innate generality, but it also facilitates learning, allowing the account to avoid the difficulties that earlier empiricist accounts of abstraction had in capturing the similarity in the input without any innate general representations.

The third and final component of Quine's account is a selection process. Quine assumes that learners engage in hypothesis testing, where overt behaviours (e.g., calling a colour sample "white") are selected through positive and negative feedback in accordance with the principles of conditioning. The selection process operates in tandem with the innate quality space to isolate a region within that space corresponding to a conventional term (e.g., the white region within the innate similarity space). In this way, the innate similarity space can come to be partitioned in culture-specific ways.[22]

The structural features of Quine's basic account—a set of innate fine-grained general discriminatory capacities, an innate similarity space, and a selection process to isolate regions within that similarity space—provide the foundation to develop a workable theory of abstraction. However, the details of Quine's account are problematic in various ways. The most serious difficulties stem from his behaviourism. Consider his explanation of the innate similarity space. Quine's account of what it is to have an innate similarity space is essentially that we are innately disposed to respond to certain stimuli in a similar manner. "A response to a red circle, if it is rewarded, will be elicited again by a pink ellipse more readily than by a blue triangle" (1969b, p. 123). This explanation is little more than a restatement of the phenomenon to be explained. It is no better than saying that we tend to respond to certain stimuli similarly (thing to be explained) because we are innately disposed to respond to those stimuli similarly (proposed explanation). True enough, but what we need to know is *why* people have the same response to the stimuli. This requires at least the outlines of an underlying *psychological* mechanism.

For this reason, a better account would be one that explains the innate sense of similarity in terms of an innate computational process operating over an innate class of fine-grained mental representations, where features of the representations and the computational process produce the similarity effects. Many computational-representational systems are possible, and this is not the place to try to adjudicate between such accounts. We will assume that some such account of similarity is the right way to proceed, as an account that sticks purely to behavioural dispositions isn't substantive. This is the first step in developing the neo-Quinean framework for understanding abstraction. And once a computational-representational system is used to explain the

[22] Quine also envisions more radical changes to the similarity space through further language learning, formal education, and the impact of science.

similarity space, it is only natural to adopt representational versions of the other components of Quine's account—the fine-grained discriminatory capacities and the selection process. So, our neo-Quinean framework will also include innate fine-grained general representations and a selection process that is a computational process—one that operates over a quality space of representational states, not a field of behavioural dispositions.[23]

The most important aspect of Quine's account that needs to be addressed, however, is the character of the selection process. Quine narrowly focuses on a single type of selection process (hypothesis testing guided by conditioning). This is really only the tip of an iceberg of possibilities, however. Though a representational version of this suggestion can account for the acquisition of general perceptual representations, accounts based on this idea constitute only one of the many possibilities for how a selection process might function, and the neo-Quinean framework should be taken to encompass the full range of such possibilities. And while there are many ways the selection process may work that do not involve hypothesis testing, even among those that do, there will be differences in the assumptions they make.

Here we will just briefly mention some of the variables in terms of which such accounts might differ. (We will consider some examples of these different approaches in the next section.) At one end of the spectrum, the way that a region in the innate quality space is isolated might involve a relatively unconstrained process (e.g., it might involve a simple summation of positive instances, a minimal subspace including all positive instances, or simple regularly shaped regions containing all positive instances). At the other end of the spectrum, the way that a region in the innate quality space is isolated might involve a highly constrained process (e.g., it might involve selection from a prespecified and highly circumscribed hypothesis space, or a hypothesis space that evolves in an innately specified manner). In delimiting regions in the quality space, the abstraction process might also begin with a set of default regions that is broadened, narrowed, or otherwise altered through the selection process. The important point for present purposes is that a wide variety of options are available for the selection process, each of which isolates a region of the innate quality space in response to the fine-grained general representations that are taken as input in its own way.

[23] Without a representational account of the selection process, we would need an explanation of why reinforcement has its effects on overt behaviour and would face difficulties arising from the fact that the principles of conditioning don't apply to many instances of learning, including word learning (Chomsky 1959). Citing only external factors (the impingement of stimuli, the imposition of rewards, etc.) is inadequate, since these clearly don't have the same effects on every physical system. There has to be something about the intrinsic character of the learning system that explains why conditioning shapes its responses. The best account that psychology has to offer is that, in many cases, the mechanism is deeply cognitive. It's because of the way that the contingencies of rewards and punishments are represented that the principles of conditioning have any purchase on changes in behavioural regularities (Gallistel 1990; Gallistel and Gibbon 2002).

There are also a number of other important sources of potential variation in accounts within the neo-Quinean framework that Quine himself does not discuss but which ought to be included in the neo-Quinean framework. For example, the fine-grained representations that form the basis of abstraction needn't always be innate. In some cases, they might be learned. Likewise, the innate quality space might not be developmentally fixed. The size of this space or the number and types of similarity dimensions might be altered. There is also no reason to have only one quality space in play. There could be multiple distinct quality spaces, and quality spaces that stand in different relations of psychological accessibility to one another. And, of course, another variable is the class of concepts that might be acquired by such a process. This is likely to include standard perceptual concepts (e.g., concepts for colours, textures, and odours). But it might also include concepts for bodily sensations (pleasure, pain, heat, etc.) as well as concepts for amodal categories such as shape concepts and concepts for spatial relations, among others. Taken together, these and the previously mentioned sources of variation introduce considerable flexibility within the neo-Quinean framework. Since our aim is simply to sketch the general outlines of a framework for learning by abstraction, we won't attempt to systematically explore all these different possibilities.[24]

In sum, the neo-Quinean framework that we are proposing takes the following form. Abstraction is a computational-representational learning process that operates over one or more quality spaces of fine-grained general representations that are ordered by one or more similarity metrics. These similarity metrics needn't be simple. In fact, they might be quite complex and multifaceted. Abstraction involves a selection process which delimits or carves out regions of a quality space. This selection process can take many different forms. Despite the variation across these accounts, all accounts in the neo-Quinean framework have in common the fact that they build in sufficient structure as input to the process—some general representations organized in terms of a suitable similarity metric—to avoid the criticisms that were so damaging to the theories of abstraction discussed in the previous section.

If we return to the example of the general concept WHITE, there are numerous alternative models for how such a representation might be acquired in the neo-Quinean framework. Just to get the feel of the framework, one possibility is a model much like the computational-representational analogue of Quine's own account of colour words. In this case, a learner comes equipped for the task with general representations for different shades of white (among other colours), as

[24] However, see the next section for discussion of a small selection of possible models for the acquisition of colour concepts within the neo-Quinean framework and for their implications regarding the rationalism-empiricism debate.

well as an innate similarity metric that organizes her colour space. Then upon encountering different instances of white things (snowballs, paper, milk, etc.), she would represent the particular shades of those encountered objects, and through a process of positive and negative feedback, develop a representation that incorporates all of the shades that received a positive signal and none of the shades that received a negative signal.

This is just one example, but notice that such a model avoids the difficulties we raised in the previous section for Locke and others, specifically by abandoning the Lockean ambition of trying to explain the origins of all general representations via abstraction. Instead, the model works by supposing that some general representations are innate (e.g., the fine-grained but still general representations of particular shades of white). Abstraction, according to the neo-Quinean framework, can't account for all general representations, but that doesn't matter, since no framework can account for the ability to learn all general representations. What this new framework does do, however, is very much in the spirit of earlier theories of abstraction in that it explains how certain concepts can be learned on the basis of fine-grained perceptual experience.

5.4 Why Our Framework for Understanding Abstraction Is Compatible with Rationalism as Well as Empiricism

The neo-Quinean framework has profound implications for the rationalism-empiricism debate. The first of these, which will be the topic of this section, is that abstraction as it is understood on the neo-Quinean framework is not a distinctively empiricist account of conceptual development. Although abstraction has historically been seen as an account of development that vindicates empiricism, as it is understood in the neo-Quinean framework, there is nothing about abstraction per se that limits it to an empiricist psychology. Abstraction is equally compatible with rationalist views of the mind. Of course, abstraction is a learning process, but as we have emphasized in earlier chapters, rationalists and empiricists do not disagree about whether learning is critical to development—they both agree that it is. Instead, their disagreement is about the character of the acquisition base (or what they take to be innate) and how learning takes place. In this section, we will consider some sample rationalist and empiricist accounts of abstraction for the domain of colour concepts.

Let's begin with a sample account towards the empiricist side of the spectrum. Consider the following description provided by Regier and Kay (2009) of a view that should sound familiar:

> Debi Roberson and colleagues…[concluded] that "color categories are formed from boundary demarcation based predominantly on language"…subject to the constraint of "grouping by similarity": namely, that categories must form contiguous regions of color space. The implication is that apart from that rather loose constraint, category boundaries are determined exclusively by local linguistic convention. (Regier and Kay 2009, p. 442)

Put in these terms, Roberson et al.'s position bears a striking resemblance to Quine's (minus the behaviourism).

In support of their view, Roberson et al. point to cross-cultural evidence demonstrating significant variation in colour concepts. For example, an important study reports that the Berinmo of Papua New Guinea use five basic colour terms that cross-cut the basic colour terms in English (Davidoff et al. 1999). One Berinmo term covers both yellow (i.e., what's called *yellow* in English) and numerous shades that English speakers think of as green. On Roberson et al.'s account, colour concepts are learned by identifying different culturally salient regions within a common initial similarity space. Since there are only weak internal constraints on the learning process, colour concepts end up varying significantly across cultures.

This model is certainly more empiricist than many. While it does posit an innate set of fine-grained representations, an innate similarity metric, and an innate constraint on the selection process (which selects only continuous regions in the similarity space), the model employs a domain-general selection mechanism, namely conditioning.

In contrast, what might a broadly rationalist account of concept learning via abstraction look like for colour concepts? One possibility would be a model that is based on the idea that certain portions of the innate colour space are innately privileged, creating an innate domain-specific bias in the selection process. One way that a view like this might be developed is suggested by a landmark cross-linguistic and cross-cultural study examining colour naming in over one hundred non-industrialized societies around the world (Regier et al. 2005). This study showed that the best examples of colours picked out by colour terms across all of the languages in these societies tended to cluster around the best examples of colours picked out by the English terms "black", "white", "red", "yellow", "green", and "blue". A *best example* of a colour for a given colour term—also known as a *focal colour*—is a shade that is taken to be paradigmatic for the broader colour category denoted by the term. For example, a shade like fire engine red is generally considered to be the best example for the English term "red". It is highly unlikely that the best examples of colours should cluster in this way across such diverse societies on an empiricist model. If there are no built-in ways to group colours, why should people from such different cultures wind up with highly similar best examples of colour terms, especially when their colour terms pick out

different regions in colour space? Importantly, this study also compared the best examples of colour terms to the centre points of the extensions of these terms (a *centre point* being the shade that represents the mean of all of the points in the extension of a colour term). What was found was that the best examples of colour terms across these many languages were more closely clustered than the centre points across these languages. This suggests that the best examples of colour categories are not simply derived from the colour fields associated with the terms by taking the best examples to be the centres of the colour category extensions. Instead, it suggests that it is the best examples of colour categories that are primary and that they function as privileged elements in the colour space—and so colour concepts may be generated by colour fields forming around focal colours, albeit in different ways in different cultures.

The upshot of these findings is that colour concepts may be learned via an abstraction process that takes the selection process to be influenced by *innate representations of the best examples* of colour categories, making the selection process more domain specific. Note that while this type of model is clearly more rationalist, this does not mean that it is insensitive to cultural input. A rationalist model along these lines could explain the strong universalist tendency Regier et al. found, while at the same time allowing for cases where there is substantial cross-cultural variation, as in the Davidoff et al. study. This point can be seen even more clearly if we consider another type of rationalist account of abstraction for colour concepts.

On this second sort of rationalist account, the selection process involved in learning colour concepts might be taken to be influenced not by focal colour representations but by a preliminary set of innate colour concepts. Mature colour concepts might then develop through a process that adjusts the borders on the colour categories represented by this initial innate set of colour concepts in a way that is sensitive to cultural factors, resulting in different sets of colour concepts for different cultures. While our aim is not to defend any of the rationalist accounts we are offering as illustrations of rationalist accounts of abstraction, it is useful to develop a fuller sense of this sort of model to consider some of the evidence that might be offered in favour of it.

One source of evidence for a model like this comes from studies of colour categorization in infants. In an important early study, Bornstein et al. (1976) showed 4-month-old infants different samples of the same shade of a primary colour until the infants began to lose interest in the new samples, and then showed them novel shades of the same colour as well as equally novel shades of a *new* colour (one that was equally different from the original shade but that crossed the colour boundary).[25] For example, infants were repeatedly shown a shade of blue

[25] The general experimental method being used here is called the *habituation method*. See Chapters 8 and 9 for more on this and related ways of investigating infants' representational abilities.

(480 nm light, i.e., 480 nanometre light) and then subsequently shown a novel shade of blue (450 nm light) and an equally novel shade of green (510 nm light). The result was that the infants looked significantly longer at the novel shade of the new colour (green) but not at the novel shade of the old colour (blue). Since the infants responded differently to the two shades—shades that were objectively equally similar to the shade of the original samples they were shown—this suggests that they represented them differently. In particular, it suggests that they represented the shade in which they showed renewed interest as a new colour relative to the original samples and that they represented the shade to which they showed reduced interest as the same colour as the original. That is, they didn't merely represent the shades as being particular fine-grained colours, but also represented them in terms of more general colour categories.[26]

Subsequent work with infants has provided a more comprehensive picture of infants' colour categorization. The most important study of infant categorization to date systematically explored infants' colour categorization using evenly spaced shades across the full colour spectrum (Skelton et al. 2017). This study used essentially the same methodology as Bornstein et al.'s study, but by sampling the full colour spectrum in this way, they were able to get much more detailed information about infant colour categorization. What they found was that infants partition the colour spectrum into five general colour categories: *red*, *yellow*, *green*, *blue*, and *purple*. Infants sharply distinguish minimal colour differences at the boundaries of these categories while treating different colours within them as equivalent. This is so even though it was also shown that infants are perfectly capable of discriminating colour differences within these colour categories, too.[27]

How might this work bear on rationalist accounts of abstraction? If infants possess innate colour concepts, these initial concepts would not be learned, and so would not be learned via abstraction. But a process of abstraction might still explain the origins of *adult* colour concepts, which are known to vary across cultures. While Regier et al.'s massive survey of colour terms across non-industrialized societies shows that such variation is subject to some constraints, we also know that in at least some cases the variation can be quite dramatic, as in Davidoff et al.'s study of Berinmo colour terms. A rationalist model of abstraction could explain the results from both of these studies, as well as the studies on infant colour categorization. On the sort of model we have in mind, rather than starting with an innate similarity space that lacks category boundaries, the abstraction process would start with a similarity space that comes with its own innately bounded regions that are modified in light of later experience. The

[26] The representation of broader regions of colour space as general colour categories (as opposed to just sensitivity to the different fine-grained colours) isn't unique to humans. For related findings with birds, see Caves et al. (2018) and Zipple et al. (2019).

[27] For related neurological evidence supporting infant colour categories, see Clifford et al. (2009) and Yang et al. (2016).

selection process would involve *adjusting* these boundaries in light of evidence regarding the colour categories that are represented by words in the learner's linguistic community, leading to the innate colour concepts being replaced by new ones during the course of development.

A variation on this rationalist model might take both initial colour concepts *and* focal colours to be innate. A model of this sort would mean that there is an initial innate partitioning of the colour space into colour categories, which might subsequently be modified or overridden. The fact that there are also innate focal colours might mean in addition that some types of subsequent changes in development would be more likely than others. For example, adjustments to colour boundaries might be biased towards ones where the new boundaries continued to include a focal colour over ones which exclude focal colours. Rationalist models of abstraction of any of these types—focusing on innate focal colours, innate bounded regions in the similarity space, or both—would be able to explain the findings on infant colour categorization and would also be fully compatible with the cross-cultural variation in adult colour concepts that has motivated broadly empiricist models like the proposal by Davidoff et al. (1999).

The compatibility of rationalism with variation in colour categories is useful in highlighting a crucial point that we emphasized earlier, particularly in Chapter 4, namely that *rationalist accounts of the origins of concepts don't entail that innate psychological structures can't be changed or overridden in development.* A psychological structure's being part of the acquisition base is perfectly compatible with its being altered or eliminated in subsequent development. This point is easy to miss, particularly as it is obscured within the debates that are often used to frame the sorts of experimental results that we have just mentioned. Work such as the study by Davidoff et al. or the study by Regier et al. is typically presented in the context of a debate between universalist accounts and accounts involving linguistic relativity. This debate is often characterized in terms of a question of whether elements of colour concepts or colour cognition are either universal or vary with language.[28] It is now common to reject this whole debate, noting that colour concepts and colour cognition are influenced both by linguistic and cultural factors and by universal aspects of human psychology or biology.

While we agree that both of these types of influences are important, we think that it is a mistake to simply dismiss the debate on these grounds. We will be discussing arguments based on universality in Chapter 11, but for now the important point is that the debate about whether colour concepts or colour cognition is universal or influenced by cultural factors is essentially a proxy for

[28] Many different factors have been explored to explain apparent constraints on the pattern of variation (Lindsey and Brown 2021). One factor that has played an important role in recent discussions in this debate has been the role of communicative needs when making use of categories in a social setting (see, e.g., Gibson et al. 2017; Zaslavsky et al. 2019).

the rationalism-empiricism debate regarding the origins of these traits. So the situation here is analogous to the one in which rationalism-empiricism debates are rejected on the grounds that we should reject the nature-nurture debate—since both nature (i.e., genetic factors) and nurture (i.e., environmental factors) contribute to development. As we emphasized in Chapter 3, while the fact that genetic factors and environmental factors both contribute to development may undermine the nature-nurture debate, it leaves the rationalism-empiricism debate perfectly intact. The moral regarding the debate between universality and linguistic relativism is similar. The larger issue that the controversy about universality and linguistic relativity regarding colour categories was getting at is how to think about the origins and development of colour concepts vis-à-vis the acquisition base—in effect, a local rationalism-empiricism debate focusing on colour concepts. This is not resolved by broad agreement that the development of colour concepts is subject to both linguistic and cultural factors and universal aspects of human psychology or biology; there is still a question of whether the best account of how colour categories develop is rationalist or empiricist.

The rationalist models of abstraction for learning new colour concepts that we have sketched so far have largely focused on making the selection process more rationalist. But it's also worth noting that there are ways in which the similarity space might be made richer and more rationalist as well. One way to see this is to consider the phenomenon of colour constancy, in which the colour of an object or different parts of an object are experienced to be the same across different lighting conditions. While it may be tempting to suppose that applying a colour concept is a simple matter of detecting the different wavelengths in the light corresponding to a given region, the phenomenon of colour constancy shows that the situation is actually rather more complicated than that.

Consider, for example, the two squares labelled "A" and "B" in the left-hand side image of a checkerboard in Figure 5.1. These are readily categorized by our visual systems as being of contrasting colours (square A being blackish or dark grey, and square B being whitish or light grey), while at the same time, B looks like it is the same colour as the light square on the bottom right of the checkerboard. Incredibly, however, in terms of luminance, not only is B not the same colour as the bottom right corner square, it is actually exactly the same colour as square A! It only appears otherwise because our visual system is making assumptions about the levels of illumination over the checkerboard, taking into account the shadow cast by the cylinder.[29]

How does colour constancy bear on the issues we have been discussing? If colour concepts apply to a colour space that incorporates colour constancy, as

[29] Given the importance of colour as diagnostic of the value of real-world objects, it is unsurprising that colour constancy isn't unique to humans. For evidence of colour constancy in non-human animals, see Neumeyer (1998); Chittka et al. (2014); and Olsson et al. (2016).

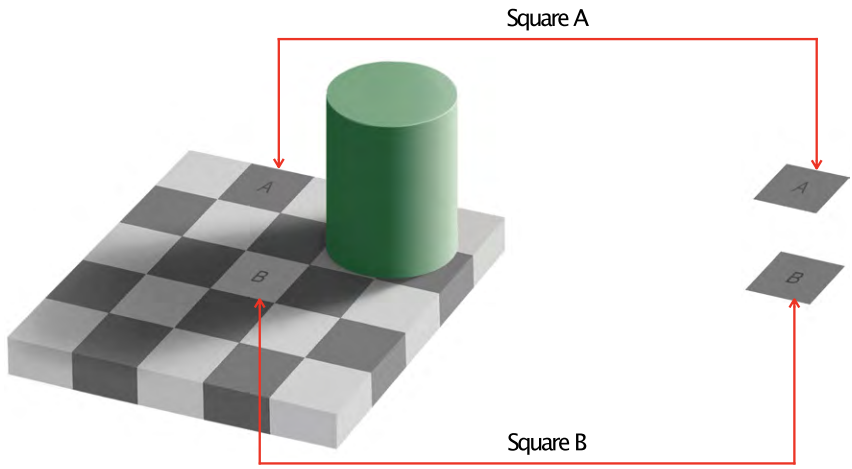

Figure 5.1 Checkerboard image illustrating colour constancy. Colour constancy allows us to see a coloured object as being the same colour in different illuminations. In the image on the left, the two squares A and B are seen to have contrasting colours due to colour constancy, which causes B to be seen as considerably lighter than A, compensating for the fact that B is in the shadow of the cylinder. In fact, these two squares are precisely the same shade of grey, as can be seen when the squares are viewed independent of context; see squares A and B on the right, where the context is removed. (Figure credit: image by Edward H Adelson, CC-BY, https://creativecommons.org/licenses/by/4.0/. The checkerboard image on the left is the original except that the image credit has been removed and put in the caption; the image of the two squares on the right is the same as the one on the left but with everything except the two squares masked by white overlay; the labelled red arrows have also been added for clarity.)

seems to be the case, then they don't just carve out regions of a representational space that is organized around the different amounts of light of different wavelengths that are reflected off of a surface. Instead, it seems that they apply to a colour space that involves richer and more sophisticated representations of colours that factor in different assumed levels and types of illumination, based on such things as representations of the relations among surfaces, orientations of light sources, and the presence of shadows. This means that the fine-grained general representations and their similarity relations (all of which constitute the input to the selection process in abstraction) are likely to be far richer and more abstract than they would be on a simpler account which doesn't factor in colour constancy.

And, if, as is widely assumed, the processes underpinning colour constancy are themselves innate (and specific to the domain of colour), then this means that *all* of the models of abstraction that we have briefly sketched in this section are likely to be more rationalist than has been noted so far. Even an account as seemingly

empiricist-friendly as the one we attributed to Roberson et al. would need to involve sophisticated innate domain-specific machinery that plays a key role in concept learning. Their account would still be more empiricist than the other models we have mentioned, but the initial description of the model we gave omits a significant innate domain-specific element that results in their model being considerably more towards the rationalist side of the rationalism-empiricism spectrum than it would otherwise be.

Other models could be considered as well, but the models we have mentioned are sufficient to illustrate how abstraction is compatible with both empiricist and rationalist accounts of the origin of concepts. We should emphasize that our aim has not been to argue for a rationalist account of the origins of colour concepts, but only to show that there is nothing that ties our framework for understanding abstraction to an empiricist psychology, even though abstraction is usually associated with empiricist accounts of conceptual development. Rationalists and empiricists can *both* help themselves to the process of abstraction.

5.5 Abstraction, Conceptual Structure, and the ABC Model of Conceptual Development

In this section, we consider a second important implication for the rationalism-empiricism debate stemming from the neo-Quinean framework. This implication concerns a less familiar issue than the question of whether abstraction is compatible with rationalist accounts of conceptual development, and so it requires us to back up a little bit to introduce the issue involved.

One of the most widely held assumptions concerning conceptual development is that semantically primitive concepts—concepts that are not themselves composed of other representations—cannot be learned and therefore must be innate. This assumption about primitive concepts is widely accepted by both rationalists and empiricists. In fact, it is closely tied to a standard way of thinking about conceptual development, which we will call the *Acquisition by Composition model* (or the ABC model) of conceptual development.[30] According to this way of thinking about conceptual development, concept learning requires a complex concept to be formed from a compositional process drawing on its semantic constituents. These semantic constituents might themselves be learned in a similar way, but eventually development has to bottom out in the semantic primitives that are the basis for all complex concepts. For this

[30] In early work, we referred to this model as the *building blocks model of conceptual development*. We now prefer "Acquisition by Composition", which does more to convey the nature of the process. In calling this book *The Building Blocks of Thought*, we aren't endorsing the ABC model of conceptual development. Rather, we have adjusted our terminology to co-opt the image of a building block as a picturesque way of talking about concepts in general.

reason, the ABC model takes them to constitute a fixed stock of innate representations and maintains that they set a fixed limit on the expressive power of the conceptual system.

Steven Pinker conveys much of the spirit of the ABC model in a discussion of the rationalism-empiricism debate:

> On the nurture side, empiricists tend to make do with the abstemious inventory of sensori-motor features, invoking only the process of association to build more complex ones. On the nature side, nativists argue that a larger and more abstract set of concepts, such as "cause," "number," "living thing," "exchange," "kin," and "danger," come ready-made rather than being assembled onsite.
>
> Both sides, if pressed, have to agree that the simple building blocks of cognition—like the keys on a piano, the alphabet in a typewriter, or the crayons in a box—must themselves be innate. Type on a standard typewriter all you want; though you can bang out any number of English words and sentences, you'll never see a single character of Hebrew or Tamil or Japanese. (Pinker 2007, p. 93)

In other words, rationalists and empiricists agree about one thing: primitive concepts are the fundamental semantic units that learning draws upon, so a theory of how they in turn are learned is not just improbable—it is downright impossible.

While a full evaluation of the ABC model will have to wait until later, we will argue here that the neo-Quinean framework for understanding abstraction provides a possible model for how some primitive concepts might be learned. In this way, the neo-Quinean framework casts doubt on the ABC model and the widespread assumptions that only complex concepts can be learned and that primitive concepts must be innate.[31]

To see how the neo-Quinean framework allows for learned primitive concepts, consider again colour concepts like WHITE, which are often taken to be primitive representations. Given the neo-Quinean framework, we can take the input to the process of abstraction to be a set of representations of various specific shades within a similarity space (particular shades of white, each corresponding to the colour of an experienced white object). A selection process operating on this input results in the demarcation of a field within this similarity space (a region in the colour space corresponding to whiteness is delimited). Let's suppose that this process also generates a new concept, WHITE, which is linked to each of the representations in the selected field such that the activation of any element in the field brings about the activation of this new higher-level representation. We can now

[31] We will return to this issue when we take up the question of why rationalists should reject Fodor's case for his claim that concepts can't be learned (see Part IV).

ask whether this new concept should be understood as primitive or complex. (Our discussion in the remainder of this section will draw on some ideas about mental representations and their meaning or content which will be more systematically introduced and explained in the next chapter. Readers who are less familiar with philosophical thinking about concepts and theories of mental content may want to return to this section after reading Chapter 6.)

Complex concepts are concepts that are composed of simpler concepts according to the principles of a compositional semantics. For example, BROWN COW is composed of BROWN and COW, and WHITE HAIR is composed WHITE and HAIR, where, in each case, the complex concept's content is a function of the contents of the concepts that it is composed of. There has been much debate concerning the question of whether lexical concepts (concepts associated with individual words) might also have a compositional structure that isn't manifest in language.[32] At the same time, colour concepts like WHITE and BROWN are examples of concepts that have seemed unlikely to possess such an internal structure, even if many other lexical concepts do.

Supposing that colour concepts are primitive, how should we think about their content? What could make it the case that they represent what they do? This is a difficult question and one for which no one has a fully satisfying answer. One option, which we will make reference to for illustrative purposes, is that the content of the representation is determined by the environmental conditions that it is causally dependent on and that it has the function of responding to (Dretske 1995). On this type of account, the concept WHITE represents whiteness because there is a systematic causal dependence between occurrences of the concept WHITE and instances of whiteness, and the concept has the function of responding to whiteness.[33] Then the role of the representations for specific shades of white would simply be that they serve to mediate the mind-world link between external conditions (whiteness) and the concept WHITE. Elsewhere, we have called such mediating factors *sustaining mechanisms* (Margolis 1998; Laurence and Margolis 2002). A sustaining mechanism doesn't directly determine a concept's content—and, in particular, doesn't underwrite a compositional semantics for the content of the concept. Rather, it makes its contribution indirectly, by establishing the mind-world relation that directly determines the concept's content. On such an account, the products of the process of abstraction—concepts like WHITE, CIRCULAR, SMOOTH, etc.—would not have their content determined compositionally but rather by the mind-world relations established by sustaining

[32] See Laurence and Margolis (1999) and Murphy (2002) for discussion.

[33] What makes the concept WHITE have this as its function, on this type of account, is ultimately the fact that the concept is the product of an evolutionary selection process, where the selection for the concept depended on the concept's having been responsive to whiteness. This general approach to functions is due to Wright (1973).

mechanisms.[34] Hence, in this case, abstraction would provide a mechanism whereby new primitive concepts could be learned.

So, there is at least some reason to suppose that abstraction, as it is understood in the neo-Quinean framework, provides a direct challenge to one of the most widely held views regarding the origins of concepts, a view embodied by the ABC model that is endorsed by most rationalists and empiricists alike and that implies that primitive concepts cannot be learned and must be innate. The reason that we are being tentative about our claim here—saying only that it provides *some* reason to call into question the ABC model and its associated claim that primitive concepts cannot be learned—is that the model of content determination for colour concepts that we have just outlined isn't mandatory. There are other approaches that are consistent with the neo-Quinean framework that would treat the output of the process of abstraction as a complex representation, not as a primitive one.

How might an account like this go? How might we understand the output—in this case the concept WHITE—as a complex representation? As before, we can take the input to abstraction on the neo-Quinean framework to be a set of representations of various specific shades within a similarity space, and a selection process will result in the demarcation of a field within the similarity space. This time, though, we will suppose that this process also generates a new concept that is a highly disjunctive representation whose many disjuncts are just the representations that appear in the demarcated field—that is, a representation of the form SHADE_1 OR SHADE_2 OR … SHADE_n, where each of these disjuncts represents a different shade of white. On this model, the semantics of the abstracted concept is plainly compositional. The content of WHITE is a function of the contents of its constituents and their compositional structure.

Both the compositional model and the sustaining mechanism model are compatible with the neo-Quinean framework. Abstraction could produce complex concepts that *incorporate* the fine-grained representations that are the input to the process through a compositional semantics, or it could produce simple concepts that are *activated by sustaining mechanisms* that incorporate the fine-grained representations. Nonetheless, a number of considerations suggest that the sustaining mechanism model may be preferable in many cases. We will briefly mention some of these in closing.

[34] Unlike BROWN and COW in BROWN COW, the representations of the various fine-grained shades of white aren't *constituents* of the concept WHITE. Theorists who opt for sustaining mechanisms rather than constituency relations often do so specifically because it weakens the relationship between the representations in the sustaining mechanism and the concept whose content is indirectly established, thus allowing for the possession of a given concept despite a great deal of perceptual and cognitive variability across agents in how the mind-world link is established and maintained (see Dretske 1981; Fodor 1987; Laurence and Margolis 1999). Different agents can possess very different sustaining mechanisms for the very same concept, provided that each agent's sustaining mechanism supports the same general content determining relations.

One consideration in favour of the sustaining mechanism model concerns the computational load for processes that occur at the level of the abstracted concept. If these processes have to operate on a highly complex structured representation and deal with each of its numerous constituents, this is likely to place a heavy processing burden on the system. On the other hand, if the processes can stick to an unstructured concept and ignore all of the structure that is inherent in its sustaining mechanism, the computational load would be considerably eased.

Presumably this type of consideration is a large part of the reason why we possess so many different general concepts—it is significantly easier on processing systems to work with the thought ANIMALS NEED ENERGY TO SURVIVE than to work with the thought AARDVARKS, ALLIGATORS, ANTEATERS, . . . AND ZEBRAS NEED ENERGY TO SURVIVE. There may also be advantages in the informational loss that is inherent to the employment of an unstructured concept. If what matters in applying a learned rule is the more general category *white*, then a representation that focuses attention on just that category (and not on various particular shades) puts the emphasis just where it should be. When it doesn't matter which precise shade is at issue, it is important not to fixate too strongly on any particular shade.

For these reasons, we think it isn't merely possible that abstraction produces primitive concepts. It seems like this should be the preferred account in this case. At the same time, it is only natural to suppose that the bar should be high when it comes to overturning something as deeply entrenched as the ABC model and its associated claim that primitive concepts cannot be learned. However, as we will see in Part IV, there are other learning mechanisms in addition to our reconceived process of abstraction which can also explain how some types of primitive concepts can be learned. After looking at those in detail and how they complement the considerations we have just given, we will see that there is ample reason to call the ABC model into question.

5.6 Conclusion

This chapter had two primary aims—one was to highlight the phenomenon of illusory explanations and how they can illicitly impede serious consideration of rationalist theories, and the other was to explore how we might move beyond illusory explanations in the case of a particular type of psychological process that has been especially influential in the history of empiricism (namely, the process of abstraction). In the first part of the chapter, we showed that illusory explanations can be remarkably hard to see for what they are and argued that this has led to an unwarranted presumption in favour of empiricism. And we have highlighted how this tendency has been fostered, in certain philosophical circles, by the assumption that the inner workings of the mind must be accessible to consciousness. In the second part of the chapter, we presented our new neo-Quinean

framework in which abstraction is reconceived in a way that drops many of the details that empiricists have associated with abstraction but retains the idea that abstraction provides a mechanism for learning new general concepts on the basis of fine-grained perceptual experience. These changes allow particular theories based on abstraction in the neo-Quinean framework to provide substantive accounts of conceptual development. Finally, we ended the chapter by highlighting two important implications that this new framework for understanding abstraction has for the rationalism-empiricism debate concerning the origins of concepts. The first is that this framework isn't inherently empiricist. Abstraction within this framework is perfectly compatible with rationalism as well; concept learning mechanisms based on abstraction provide a useful tool for understanding conceptual development for empiricists and rationalists alike. The second implication is that the neo-Quinean framework arguably provides an account not only of how new concepts can be learned, but of how new *primitive* concepts can be learned, thereby calling into question the ABC model of conceptual development and its associated claim that primitive concepts must be innate.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0005

6

Concepts, Innateness, and Why Concept Nativism Is about More Than Just Innate Concepts

The central aim of this book is to offer a systematic defence of the rationalist framework for understanding the origins of concepts and to make the case for our own version of concept nativism. Our main focus in Part I, however, has been on the closely associated secondary aim of comprehensively rethinking what the rationalism-empiricism debate is about and introducing and clarifying the theoretical notions that we take to be central to understanding this debate. This chapter serves as a bridge between Part I and the remainder of the book. In it, we outline how the account that we developed for understanding the rationalism-empiricism debate applies specifically to the origins of concepts. We begin, in section 6.1 and section 6.2, by exploring in greater detail two fundamental questions: the question of what innateness is and the question of what concepts are. Although we have argued that the rationalism-empiricism debate can be framed without relying on the notion of innateness, we think the notion still has a role to play in the debate. To make clear why, we explain our own understanding of what innateness consists in and why we think much recent scepticism about this notion is misguided. Our discussion of what concepts are is meant to help readers navigate a complex set of questions and issues that are naturally entwined with the question of how concepts are acquired. But more importantly, it will help us address a common but deeply mistaken view about rationalist accounts of the origins of concepts—the view that these just claim that there are innate concepts. We argue that this is simply wrong. Building on our explanation of the rationalism-empiricism debate in Chapter 2, which highlighted the diversity of views within the rationalist framework as it applies to any type of psychological trait, we show that there is a similar diversity of rationalist accounts of the origins of concepts, and that the extent to which an account is rationalist depends on a number of factors and is not simply a matter of how many innate concepts it posits. As we turn to our positive case for concept nativism in Part II, this clarification of what concept nativism does, and does not, entail will be essential for fully appreciating what counts as a successful argument for concept nativism.

6.1 What Is Innateness?

In this section, we begin by taking a closer look at the question of what innateness is. Rather than provide a comprehensive survey of the many different accounts on offer, we will focus on the account that we favour, briefly explaining some of its advantages and comparing it to two close alternatives.[1] These comparisons are meant to highlight several key features of our account, to help clarify the sorts of constraints on the notion of innateness that we think matter most, and to illustrate the advantages of our general approach. In addition, we will address the increasingly common charge that the notion of innateness should be abandoned altogether on the grounds that there are so many different conceptions of what innateness is and that this is bound to foster confusion among theorists who have very different ideas about what it means to say that a trait is innate. Our own view is that this concern about the continued use of the notion of innateness is overstated and that there is nothing wrong with—and a lot to be gained by—its continued use.

So, what is it for a psychological trait to be innate? In Chapter 2, we suggested that a psychological trait's being innate essentially comes down to whether the trait is part of the acquisition base. Our own account of innateness is based on the core idea behind an account in the literature known as *primitivism*. The most carefully worked out version of primitivism is due to Richard Samuels (2002; see also Cowie 1999). Samuels' primitivist account takes the form of a set of necessary and sufficient conditions for innateness according to which a psychological trait is innate just in case the following two conditions hold.

(i) The psychological trait is *psychologically primitive* in the sense that it is not acquired by a psychological process.
(ii) The psychological trait is acquired in the normal course of development.[2]

In our view, condition (i)—the core of the account—is basically the right way to understand innateness. This is because, on our understanding of the rationalism-empiricism debate, what is at stake in the debate comes down to the composition of the acquisition base, and this in turn is a matter of which traits are psychologically primitive in Samuels' sense (i.e., which psychological traits aren't learned or otherwise acquired via psychological processes).

[1] For other accounts of innateness that we won't be able to discuss here, see Stich (1975); Ariew (1996); Sober (1998); Quartz (2003); Khalidi (2007); and O'Neill (2015).

[2] This formulation is based on Mallon and Weinberg's (2006) friendly modification of Samuels' characterization of primitivism. On Samuels' original account, clause (i) makes reference to explanatory considerations ("there is no correct scientific psychological theory that explains the acquisition" of the trait (Samuels 2002, p. 246)), whereas Mallon and Weinberg's formulation is helpfully framed directly in terms of the facts regarding the acquisition of the trait.

However, while we endorse what we take to be the core claim of primitivism, we can't simply adopt Samuels' account as it stands. In particular, we take exception to his condition (ii) and have doubts about the driving motivation for adding a second clause in the first place. This will take some explaining.

As we read his proposal, Samuels begins with the core claim of primitivism, according to which a psychological trait's status as an innate trait has to do with its being a psychological primitive. But then he points out that this condition taken by itself is too broad. It classifies quite a few traits as innate that clearly are not innate—Samuels calls this the *overgeneralization problem*. He mentions the hypothetical case of acquiring the ability to speak and understand Latin by taking a futuristic pill that produces knowledge of Latin without having to go through the usual mental exertions (Fodor 1975). Since it is stipulated in the example that the pill produces knowledge of Latin without mediating psychological processes, this knowledge would be psychologically primitive. Nonetheless, it seems wrong to say that this knowledge would be innate. Likewise, Samuels discusses a case in which infection by the Ross River virus leads its victims to hallucinate that buildings are crashing down around them. Here too the psychological trait at issue—the tendency to hallucinate in this way—is presumed to be psychologically primitive, but it seems wrong to say that it is innate. Samuels' condition (ii) is brought in to address cases along these lines. It allows for a trait to be psychologically primitive without being innate only so long as the trait isn't acquired in the normal course of development.

While Samuels' condition (ii) may help him to deal with these particular examples, we don't think this is the best way to respond to the threat of overgeneralization. For starters, it isn't sufficiently clear what counts as "the normal course of development". Consider, for example, the case of *Toxoplasma gondii* infection. *T. gondii* is a protozoan parasite with a complex life cycle. Felines, including domestic cats, are its primary host, and *T. gondii* can only reproduce in the feline gut. But prior to reaching this life stage, it finds its way into cat faeces (i.e., the faeces of previously infected felines) and then the soil, where it is ingested by other animals, including mice, rats, and livestock. Humans, too, can be infected by the parasite by eating uncooked meat from an infected animal (e.g., raw beef) or, like other animals, through contact with cat faeces or soil. The impact of the parasite on intermediary host animals includes a number of peculiar psychological effects. Mice and rats lose their normal fear of cats and aversion to the smell of cat urine (Berdoy et al. 2000), perhaps even becoming sexually aroused by the scent (House et al. 2011).[3] There seem to be psychological effects in humans, too, with differential consequences for men and women. As Flegr (2007) reports, in psychological tests, infected men "were more likely to

[3] This is presumably adaptive for the parasite, leading to more mice and rats being eaten by cats, and so to more cats infected with *T. gondii*.

disregard rules and were more expedient, suspicious, jealous, and dogmatic [relative to uninfected controls]…[infected women] were more warm hearted, outgoing, conscientious, persistent, and moralistic [relative to uninfected controls]. Both men and women had significantly higher apprehension…compared with the uninfected controls" (p. 757).

Supposing that these various acquired psychological traits in mice, rats, and humans are psychologically primitive, they provide another apparent counterexample to primitivism, as it is intuitively implausible to say that they are innate. Unlike Latin pills and Ross River Fever, however, *T. gondii* infection is not at all uncommon. In many countries, it is estimated that more than 50% of the population is infected and estimates of global human infections range from 20% to 60% (Tenter et al. 2000; Lindová et al. 2006). The problem for Samuels, then, is that this looks exactly like the sort of case that he would deem a potential counterexample to (i) taken on its own, but also it isn't ruled out by (ii), since this parasite *is* part of the normal course of human development in many communities if not globally.[4]

However, we don't want to rest our case against adopting Samuels' normalcy condition with the charge that his primitivism still overgeneralizes. Rather, we'd suggest that this whole way of approaching the issue of explaining innateness—with the aim of providing a set of necessary and sufficient conditions that are immune to all potential counterexamples—is misguided. This is because it makes Samuels' account of innateness hostage to cases that have no bearing on what is actually at stake between rationalists and empiricists. Notice that the various "counterexamples" we have just been discussing—Fodor's Latin pill, Ross River virus, *T. gondii* infection—have in common that *they aren't the least bit germane to the contemporary rationalism-empiricism debate*. No rationalist or empiricist theory turns on the details of such cases. And no rationalist or empiricist takes any of these (or similar cases) to provide an important test case for deciding between rationalism and empiricism. For this reason, we think it is a serious mistake to let one's account of innateness be guided by such cases and by our intuitions about them. There is absolutely nothing to be gained by arguing about how to exclude them in crafting a definition of "innate". What we need isn't a formula that precludes all possible overgeneralizations—including overgeneralizations into territory that is irrelevant to the debate between rationalists and empiricists—but rather a serviceable characterization of innateness that illuminates the competing claims of rationalists and empiricists regarding what mental traits are innate.

[4] Even if it weren't, there are *possible* situations in which it would unquestionably be part of the normal course of human development. Much the same is true of the Latin pill example, assuming that such a pill were possible. One can imagine its active ingredient being administered to the water supply, as fluoride is in some countries, with the population as a whole coming to acquire knowledge of Latin without any psychological processes mediating its acquisition.

Our view of innateness, then, is that we should take the insight at the heart of primitivism—that innate traits are psychologically primitive—and stop there. There is no need and no benefit to grappling with Samuels' overgeneralization problem where it has no bearing on the origins of traits that rationalists and empiricists dispute.[5] This stripped-down account is all that is needed to make sense of the rationalism-empiricism debate and the role that innateness plays in rationalist and empiricist theories of psychological development. If we focus on this role, we can even dispense with using the term "innate" altogether—as we largely did in Chapter 2—and instead simply characterize the disagreement between rationalists and empiricists directly in terms of their competing views of the acquisition base. Rationalists and empiricists agree that many psychological traits are acquired via psychologically-mediated processes. On pain of an infinite regress, they also agree that not all psychological traits are acquired in this way. So both rationalists and empiricists must take there to be psychological traits that are part of the acquisition base—traits that are not themselves explained by more fundamental psychological traits and processes. Still, while it is possible to avoid using the term "innate" in this way, we see no drawback to retaining the term, and much to gain, as it emphasizes the similarity in outlook across different fields and different research traditions that have used this terminology and appropriately links current debates to the traditional philosophical debate about innate ideas.

In order to highlight the virtues of our version of the primitivist account of innateness, we want to briefly look at two related competing accounts of innateness. The first of these incorporates and emphasizes an evolutionary constraint in characterizing innateness.[6] The simplest way to do this would be to adopt an account that says that innate psychological traits are psychologically primitive, as ours does, but that has a further condition that also requires them to be biological adaptations—products of natural selection.[7] This type of adaptationist version of primitivism might be seen as a friendly variation on both our account and Samuels'.

Is an account of this sort an improvement over our own stripped-down version of primitivism? It could certainly handle the examples of overgeneralization that Samuels mentions, as well as the changes in human psychology that are

[5] Moreover, adding an additional constraint like Samuels' has the unfortunate consequence of prohibiting innate traits that are part of the acquisition base but that are not acquired in the normal course of development—ruling out the possibility of statistically uncommon traits that are part of the acquisition base for only a minority of individuals being innate traits (e.g., psychological structures involved in or contributing to such traits as perfect pitch, synaesthesia, tetrachromacy, and the like).

[6] See, e.g., the characterization of innate cognitive mechanisms in Tooby and Cosmides (1992), in which they are understood to be "universal *evolved* psychological mechanisms" (p. 37; emphasis added).

[7] For those who are committed to the project of providing necessary and sufficient conditions for innateness, substituting an adaptation condition for Samuels' normalcy condition might be thought to be a better way of handling Samuels' overgeneralization problem. Alternatively, another possibility would be to keep Samuels' normalcy condition and add the adaptation requirement as a third condition.

brought about by *T. gondii* infection, since none of these would count as human adaptations. Moreover, many traits that are taken to be innate *are* widely thought to be adaptations. Even empiricists who credit the acquisition base with little more than domain-general statistical learning mechanisms are likely to take such mechanisms to be adaptations for learning that have been shaped by natural selection. Nonetheless, *requiring* innate traits to be adaptations may well be too strong in that it can lead to an inability to distinguish between theories that clearly fall on different sides of the rationalism-empiricism continuum.

Consider the status of language. The typical rationalist view about language is that language acquisition is grounded in an innate acquisition system that is specific to language. But within rationalist circles, there is debate about whether the language acquisition system is an adaptation, with Chomsky and some other prominent rationalists having claimed that it is a byproduct of selection for other abilities (see, e.g., Chomsky 1972/2006, 1988). In this case, the sharp contrast with empiricist theories of language is still there; it just doesn't come down to the question of whether language is ultimately an adaptation. Likewise, as we noted in Chapter 4, rationalists can and do point to other domain-specific systems in the acquisition base that they take to be byproducts, or perhaps the result of genetic drift. An account of the mind that postulated many systems like this would still be considered rationalist by rationalists and empiricists alike. So if we want to respect the way that the terms *rationalism* and *empiricism* are used in the rationalism-empiricism debate, it is better to opt for an account of innateness that is neutral about natural selection, as our stripped-down version of primitivism is.

The second alternative theory of innateness we will consider is due to Mallon and Weinberg (2006). Their account rejects condition (i) of Samuels' primitivist account but retains his condition (ii), and so constitutes a more radical departure from our version of primitivism. On Mallon and Weinberg's account,

> a trait *t* is innate in an organism *O* to the extent that
>
> (i) *O* would develop *t* across the range of normal environments (*invariance condition*); and
>
> (ii) The proximal cause of *O's* development of *t* is by a closed process or processes (*closed process condition*). (2006, pp. 339–340)

Condition (i) of their account, the *invariance condition*, is largely equivalent to Samuels' normalcy condition. We will focus here, however, on condition (ii) of their account, the *closed process condition*, which holds that the proximal cause of the development of an innate trait in a given organism is governed by a closed process (or set of processes).[8]

[8] See Quartz (2003) for a related account.

For Mallon and Weinberg, a process is *closed* to the extent that it generates relatively few different types of traits in response to environmental variation; conversely a process is *open* to the extent that it generates a relatively large number of different types of traits in response to environmental variation. For psychological traits, the core idea of a closed process is similar to that of the process of acquiring a psychological trait via a special-purpose learning mechanism, while the core idea of an open process is similar to that of the process of acquiring a psychological trait via a general-purpose learning mechanism. We will discuss how their account deals with some sample traits in a moment, but before we get to that, we want to briefly highlight some key features of their account.

By design, Mallon and Weinberg's account applies not only to psychological traits but also to non-psychological traits, such as eye colour or height, and it doesn't require an innate trait to be acquired by a non-psychological process—it only requires that the process be sufficiently closed. This feature of their account distinguishes it from both our account and Samuels', which by design are exclusively directed to psychological traits and aren't meant to cover non-psychological traits. One interesting and important consequence of this feature of Mallon and Weinberg's account, which we will return to shortly, is that their account allows for the possibility that psychologically primitive traits can fail to be innate. (This would happen whenever a psychologically primitive trait is acquired by relatively open non-psychological processes.) One motivation for having an account of innateness that applies to both psychological and non-psychological traits is that it wouldn't require that there be a principled distinction between psychological and non-psychological phenomena. This motivation dovetails with one of Mallon and Weinberg's main criticisms of primitivism, which is that psychological traits and processes are too heterogenous and don't constitute a well-defined domain—and consequently that there is no principled distinction between the psychological and the non-psychological. Another potentially attractive feature of their account is that it also paves the way for a graded notion of innateness, since acquisition processes can be closed (or open) to varying degrees. Some theorists might see this as an advantage over approaches—like ours—that are limited to saying that a trait is either innate or not.

What should we make of Mallon and Weinberg's account of innateness? Let's first consider some of the general features of their account that distinguish it from our own. It is an open question how heterogenous the psychological domain is, but we do not think that such heterogeneity is problematic for our account and for others that are only directed at providing an account of innateness for psychological traits.[9] Nor is it problematic that there will be phenomena at or near the border of any reasonable dividing line between the psychological and the

[9] See Chapter 25 for further discussion of the issue of what distinguishes psychological phenomena from non-psychological phenomena.

non-psychological, phenomena for which it isn't particularly clear whether they are psychological or not. Much the same issues arise for other perfectly respectable scientific domains—for example, even the physical sciences, including disciplines like biology and chemistry, admit a certain amount of vagueness concerning the phenomena they encompass. If primitivism inherits some vagueness because of a little unclarity regarding what exactly counts as a *psychological* primitive, so be it. A certain amount of unclarity around the edges is perfectly acceptable. Not only are there many cases where there is simply no question that the traits or processes in question are psychological, but as we argued above (and will illustrate throughout the book), the core of the contemporary debate between rationalists and empiricists concerns just such cases where rationalists and empiricists disagree about the character of the psychological processes involved in the acquisition of paradigmatic psychological traits.

What about the fact that Mallon and Weinberg's account provides a graded notion of innateness? Some theorists may welcome being able to make comparisons regarding the extent to which different traits are innate. But even on an account that has a categorical conception of innateness, there are ways of achieving much the same effect. Our own account of the rationalism-empiricism debate takes innateness to be an all-or-nothing notion, but as discussed in Chapter 2, it is graded in terms of the extent to which different theories of development are rationalist (or empiricist). For example, it says that one theory of development is *more rationalist* than another to the extent that it postulates a greater quantity of characteristically rationalist psychological structures, characteristically rationalist psychological structures across a greater diversity of content domains, characteristically rationalist psychological structures of greater complexity, or more richly articulated characteristically rationalist psychological structures than competing theories (see Box 7 in Chapter 2).[10] So while our account has no place for a graded notion of innateness, it offers a natural and straightforward way to capture the idea that the difference between rationalism and empiricism isn't all or nothing all the same.

Let's turn now to the core of their account, the *closed process condition.* It turns out that this condition is problematic. To begin with, it is not at all clear which processes should be seen as open (or relatively open) and which should be seen as closed (or relatively closed). One of the examples that Mallon and Weinberg give to illustrate this distinction is the Chomskyan view that language acquisition depends on a domain-specific system that incorporates the principles of Universal Grammar. Mallon and Weinberg refer to this Chomskyan view as an example of an open process, since variation in the environment (whether children hear

[10] As explained in Chapter 2, however, it is also necessary to take into account the possibility of trade-offs among these and related factors in assessing the extent to which one account is more rationalist than another.

French, English, Japanese, etc.) leads to considerable variation in the psychological outcome (speaking French, English, Japanese, etc.). But this is rather awkward given that the Chomskyan view of language acquisition is a paradigmatic example of a *rationalist* account of cognitive development. So if the core processes involved in this case are illustrative of open processes, this would suggest that paradigmatically rationalist accounts of cognitive development imply that the trait being acquired is not particularly innate.

Then again, although Mallon and Weinberg do take the Chomskyan account of language acquisition to be illustrative of an open process of development, it's not clear that they should. In fact, there is a good case to be made that this sort of account is actually a better illustration of a *closed* process. After all, while Mallon and Weinberg are right that Universal Grammar allows for the acquisition of a wide range of different languages, the range of psychological traits that the envisioned developmental processes can produce is highly restricted. They can't produce the ability to see the world in three dimensions, recognize faces, reorient oneself in an environment following disorientation, and so on. In fact, they can only produce one type of cognitive ability—the ability to speak a natural language. Moreover, the languages they are capable of acquiring are limited to those that are compatible with Universal Grammar. Taken together, these facts strongly suggest that language acquisition, as it is envisioned in the example, is actually a highly *closed* process. They are also the very reasons why Chomskyan accounts of language acquisition are usually taken to involve innate domain-specific learning mechanisms. In this respect, if it turns out that this example is taken to involve a closed process, contrary to Mallon and Weinberg's own claim that it involves an open process, this could help them somewhat regarding the overall case for their account of innateness.[11] Nevertheless, what this example shows is that it isn't particularly clear what makes a developmental process open or closed or how to apply the open/closed distinction even in what ought to be a fairly straightforward case.

A bigger problem for Mallon and Weinberg's account springs from the fact that their account allows that the psychologically primitive traits that comprise the acquisition base needn't be innate. Mallon and Weinberg's account of innateness notably drops the condition that links being psychologically primitive with being innate—a link that is central to both our account and to Samuels'. Since their

[11] On the other hand, if the Chomskyan account of language acquisition is taken to involve a closed developmental process, this leads to a different sort of awkward result for their account. It suggests that a language like French is both learned and innate. It is learned because it is acquired by a language-specific acquisition system that is sensitive to the features of the child's linguistic environment; at the same time, on their account it is innate, or substantially innate, because the developmental process is taken to be closed. In contrast, our account of innateness wouldn't say that French is both learned and innate on this sort of account of language acquisition. Instead, it simply says that French is learned and not innate. What's innate are the elements of the domain-specific acquisition system for learning a natural language that are part of the acquisition base. Our account captures the graded rationalism here not by saying that French is innate to some given extent, but rather by saying that French is acquired via a rationalist learning mechanism that is rationalist to some given extent.

account of innateness turns on how open or closed the developmental process is that's responsible for a trait, the question of whether the primitive psychological traits in the acquisition base are innate comes down to whether the biological processes that are responsible for their acquisition are themselves open or closed. To see what this means, consider again the sort of Chomskyan account of language acquisition in which language acquisition depends on a psychologically primitive domain-specific system that incorporates the principles of Universal Grammar. Typically on such accounts, Universal Grammar is taken to be part of the innate foundation for the acquisition of natural language and hence part of the acquisition base. Now on Mallon and Weinberg's account, Universal Grammar would still be part of the acquisition base (given the assumption that the language acquisition system is psychologically primitive), but it is not at all clear whether it would be counted as innate. It's possible that the biological processes that lead to the development of Universal Grammar, or to other elements of the acquisition base, are closed processes. In that case, these traits would be innate. But equally it's possible that the biological processes that lead to the development of Universal Grammar, or to other elements of the acquisition base, are open processes. In that case, these traits wouldn't be innate. As things stand, Mallon and Weinberg's account is completely neutral about whether such traits are innate, and it may well be that many psychologically primitive traits wouldn't count as innate on their account of what innateness is.

This gives rise to an even bigger concern for their account. If the process of development leading to the acquisition of Universal Grammar turns out to be relatively open while the process of acquiring a natural language on the basis of Universal Grammar is relatively closed, Universal Grammar could turn out to be *less innate* than the natural language one acquires on the basis of Universal Grammar (e.g., less innate than the ability to speak French). More generally, it seems entirely possible that the non-psychological processes that lead to the development of a psychological acquisition mechanism, M, could be more open than the acquisition process mediated by M. Hypothetically, the psychological acquisition mechanism M might allow only two possible outcomes, A and B, which are triggered by specific environmental stimuli, and yet the non-psychological processes involved in the development of mechanism M could be considerably more open, allowing for far more than two possible outcomes in response to environmental variation. In that case M would be less innate than its product A, even though A is acquired on the basis of M. For all we know, such cases could in fact be common.

Much the same situation could also occur sticking just to the psychological level. That is, a relatively open psychological acquisition process could be the proximal cause of a subsequent acquisition mechanism that implements a relatively closed psychological acquisition process for acquiring a further psychological trait. Consider, for instance, the suggestion made by Fiona Cowie (1999) that

language learning might depend on there being a domain-specific "helping hand" for language learning to proceed, but that this domain-specific learning mechanism is itself learned via domain-general processes. This is clearly an empiricist model, since language is ultimately acquired through an empiricist learning mechanism—a domain-general mechanism—even if this empiricist mechanism does its job by first creating a domain-specific acquisition system that is involved in language acquisition. But given Mallon and Weinberg's account of innateness, the aspects of our knowledge of language that this domain-specific learning mechanism produce would turn out to be substantially innate on this empiricist account, since the proximal cause involves a closed acquisition process.

This general pattern could turn out to be quite a common one—an empiricist view proposes an initial domain-general process that produces a domain-specific mechanism, which in turn is involved in the acquisition of further psychological traits. Another example along these lines is Karmiloff-Smith's (1992) claim that many psychological modules (which for present purposes we can think of simply as domain-specific processing mechanisms) are not innate and are instead acquired through a relatively domain-general psychological process, which she refers to as a process of *representational redescription*. On this approach, although the psychological process of module building is supposed to be quite flexible, the further development that an acquired module supports may be relatively closed. But then Mallon and Weinberg's account would end up classifying Karmiloff-Smith's view about products of this further development as one in which they are substantially innate, despite the fact that the whole point of tracing the developmental process back to an underlying process of representational redescription for Karmiloff-Smith is to provide a *non-rationalist* theoretical framework (see Figure 6.1). The difficulty here goes right to the heart of Mallon and Weinberg's proposal. In the rationalism-empiricism debate, it isn't the proximal cause of a trait that matters. What matters is the full psychological-level cause, going all the way back to the acquisition base. If this full psychological account traces back to solely domain-general learning mechanisms in the acquisition base, it is simply not plausible to take that trait to be substantially innate, regardless of what the more proximal psychological causes of the trait are like.[12]

[12] Another issue for Mallon and Weinberg is how the two conditions in their account (the closed process condition and the invariance condition) should be weighted. A trait might be invariant to a very low degree (e.g., the trait might only be triggered by a single rare type of environmental stimulus) but be acquired via a highly closed acquisition process (one that has just one or a few outcomes). Such traits would typically be considered the result of rationalist acquisition mechanisms, at least when the acquisition process for such traits is part of the acquisition base. In contrast, a trait may be highly invariant (e.g., nearly universal), but be acquired via a very open acquisition process (one with numerous potential outcomes). Such traits would typically be considered the result of empiricist acquisition mechanisms. On Mallon and Weinberg's account, it is unclear which of these should be considered more innate or why.

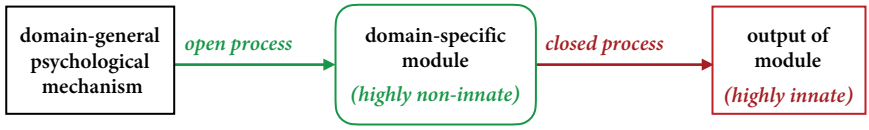

Figure 6.1 A problem for the closed process invariance account of innateness. On Mallon and Weinberg's account, an open process implemented by the operation of one psychological mechanism can produce another psychological mechanism that implements a closed process. For example, on Karmiloff-Smith's model for acquiring learned modules, a general learning process (known as *representational redescription*) is responsible for creating new domain-specific modules of many different types. This process is an open process, so the domain-specific modules that it produces count as highly non-innate. However, these modules implement further, highly constrained learning processes, each of which produce only a very narrow range of psychological traits, making them closed processes whose products are highly innate. Accordingly, Mallon and Weinberg's account of innateness would imply that highly innate traits can be acquired on the basis of highly non-innate traits, which in turn trace back to an empiricist acquisition base.

These arguments highlight how Mallon and Weinberg's account struggles precisely because it abandons the core primitivist condition that we take as the basis for our own account of innateness. Because they abandon this condition, they open themselves up to the very real possibility that unlearned psychological traits that are part of the acquisition base might not come out as innate. At the same time, learned psychological traits may turn out to be more innate than the learning mechanisms that are involved in acquiring them, and learned psychological traits that are acquired via domain-specific learning mechanisms may turn out to be innate even if their acquisition ultimately traces back to wholly domain-general learning mechanisms. By contrast, our stripped-down version of primitivism has none of these problems. It is also worth noting that our account can also maintain the association of relatively closed processes with rationalism, and relatively open processes with empiricism, when a process has one of these properties in light of the domain specificity or domain generality of psychological structures in the acquisition base.

Before closing this section, we should say something about the widespread view that the existence of a large variety of different accounts of the nature of innateness leads to confusion and argues for abandoning innateness as a theoretical notion. It has become increasingly popular to argue for eliminativism regarding the notion of innateness on just these grounds. On this view, the theoretical construct *innateness* only leads to confusion because different theorists have very different ideas about what it means for something to be innate. Further, it is claimed that the term "innate" in ordinary language will not help as a guide, since this term similarly conflates a number of distinct ideas (present at birth, universal, genetically determined, etc.). Given all of this variability, the argument

continues, we should expect that different theorists may be talking at cross-purposes or that individual theorists, not fully recognizing the many different ideas that have been packed into the same term may formulate arguments that aren't sound because they equivocate between these different meanings. As a result, we would be better off abandoning the notion of innateness altogether (Bateson 2000; Griffiths 2002; Mameli and Bateson 2006; Mameli 2008; Cowie 2009; Shea 2012).

One list of the different meanings of "innate"—a fairly typical one—teases apart no less than sixteen distinct meanings, many of which have variants that also can be teased apart, giving us as many as twenty-five distinct accounts of what it means to say that a trait is innate (Cowie 2009, pp. 82–83):

1 T is innate if T is present at birth.
2 T is innate if T emerges in the normal course of development.
3 T is innate if T's external or experiential causes are inadequate to explain its existence or properties.
4 T is innate if T is acquired as a result of the operation of a highly specialized, or domain-specific, mechanism of learning.
4a T is innate if T was acquired by means of a developmental mechanism designed by natural selection reliably to produce T at the appropriate point in the organism's development.
5 T is innate if the processes responsible for T's development are inexplicable given the explanatory apparatus and concepts of psychology ('Primitivism').
5a T is innate (by the lights of some science, S) if the processes responsible for T's development are inexplicable given the explanatory apparatus and concepts of S.
6 T is innate if T is genetically determined.
6a T is innate if T is caused by the genes alone.
6b T is innate if T is genetically influenced.
6c T is innate if T is appropriately caused by the genes.
6d T is innate if (i) T has been selected for and (ii) the architecture of the part(s) of the brain responsible for the acquisition of T developed under the control of genetic factors together with merely 'permissive' environmental factors.
7 T is innate if T is genetically encoded.
7a T is innate if all the information required for T's development is encoded in the genes.
8 T is innate if T is generatively entrenched.
9 T is innate if T is highly canalized.
10 T is innate if T is highly heritable.
11 T is innate if T is species typical.

12 T is innate if T is an adaptation.
13 T is innate if T is "produced by internal causes".
14 T is innate if T is not caused by the environment.
15 T is innate if T is not environmentally alterable (i.e., if changes in the environment cannot produce alternative phenotypes T').
15a T is innate if T is not alterable in normal environments (i.e., if alternative phenotypes T' cannot occur in normal environments).
16 T is innate if T is unlearned.
16a T is innate if T does not result from mechanisms designed to produce plasticity.

Put this way, as a list of so many possibilities, it may seem that the very idea of an innate trait is deeply problematic. However, in identifying the different possibilities, we need to take into account that the number of distinctions one makes in a list like this is in part a function of one's tendencies to see differences as opposed to similarities among different views. "Splitters" will naturally come up with a larger number than "lumpers". In this case, a lumper could easily argue that items 2, 8, 9, 11, and 15 are all more or less the same, as they are directed at some type of invariance idea.[13] Likewise, 3, 6, 7, 13, and 14 are all more or less the same, as they are directed at the idea that innateness involves an explanation that references a type of cause that is, in some sense, internal to an organism. Collapsing these minor differences drastically reduces the number of claimed alternatives to be considered. In addition, some can be eliminated for being components of accounts of innateness rather than accounts in their own right (4 and 12). And others obviously are not serious candidates at all (1 and 2);[14] they may inform how "innate" is used in ordinary speech and outside of a scientific context, but the rationalism-empiricism debate isn't about ordinary language or common-sense views of the mind. Factoring in all of these considerations leaves us with just a few genuinely different theories to contend with, rather than a daunting sixteen to twenty-five.

The situation for innateness turns out to be not that different than what one finds with any other philosophically interesting term. Consider "knowledge". A splitter might easily generate a long list of apparently competing accounts and caution people from continuing to use the term:

1 S knows *p* if S has a true belief that *p*.
2 S knows *p* if S has a justified true belief that *p*.

[13] For invariantist accounts of innateness, see, for example, Ariew (1996) and Sober (1998).
[14] Virtually every list of potential accounts of innateness given by advocates of the "confused construct" argument is padded in this way with a number of obvious non-starters—accounts that no contemporary theorist who accepts the notion of innateness actually adopts as their own understanding of what innateness amounts to.

3 S knows *p* if S has a true belief that *p*, and *p* is part of a maximally coherent set of beliefs.
4 S knows *p* if S has a true belief that *p*, and S's belief that *p* is grounded in foundational beliefs that are non-inferentially known.
5 S knows *p* if *p* is a justified true belief, and if S has ruled out all relevant alternatives to *p*.
6 S knows *p* if S has a true belief that *p* was caused by the fact that *p*.
7 S knows *p* if *p* is a justified true belief, and S's belief that *p* is not inferred from any false belief.
8 S knows *p* if *p* is a justified true belief, and S would not believe *p* if *p* were false.
9 S knows *p* if *p* is a justified true belief, and if were S to believe *p*, *p* wouldn't be false.
9a S knows *p* if *p* is a justified true belief, and in all nearby worlds where S believes that *p*, *p* is not false.
10 S knows *p* if *p* is a justified true belief, and S's belief that *p* is not true merely by luck.
11 S knows *p* if S has a true belief that *p* and S's reasons for believing that *p* are a reliable indication that *p* is true.
11a S knows *p* if S has a true belief that *p* and S's reasons for believing that *p* necessitate *p*.
11b S knows *p* if S has a true belief that *p* and S's reasons for believing that *p* make the probability that *p* is true very high.
12 S knows *p* if S has a true belief that *p* that was produced by a reliable cognitive process.
12a S knows *p* if S has a true belief that *p* that was produced by a cognitive process that is reliable in normal worlds.
12b S knows *p* if S has a true belief that *p* that was produced by a reliable cognitive process and S has no reason to believe that *p* wasn't reliably caused.
12c S knows *p* if S has a true belief that *p* that was produced by a reliable cognitive process where the relevant cognitive faculties are functioning properly in an appropriate environment.
13 S knows *p* if S has a true belief that *p*, which is true in a way that manifests S's skill in believing.
14 S knows *p* if S's total evidence includes the proposition that *p*.
15 S knows *p* if *p* is true and no epistemic weakness vis-à-vis *p* prevents S from properly using *p* as a reason for action.
16 S knows *p* if *p* is true and S is justified in believing *p*, relative to standards of justification appropriate to the context in which S believes that *p*.[15]
…

[15] See Goldman and Beddor (2015) and Ichikawa and Steup (2017) for an overview of some different proposals regarding the analysis of knowledge.

The same could be said about "morally right":

1a T is morally right if T maximizes net pleasure (versus pain).
1b T is morally right if T maximizes desire satisfaction or preference fulfilment.
1c T is morally right if T maximizes a plurality of values.
1d T is morally right if the total set of consequences of T is better than the total set of consequences of not doing T.
1e T is morally right if T maximizes respect for rights.
…
2 T is morally right if T maximizes foreseeable good consequences.
3 T is morally right if T maximizes likely good consequences.
4 T is morally right if T is in accord with rules whose acceptance would result in the greatest happiness for the greatest number.
5 T is morally right if T produces consequences with the highest average utility.
…
6 T is morally right if God commands us to do T.
…
7a T is morally right if T is in accord with a principle that could be willed to be a universal law.
7b T is morally right if doing T involves acting in such a way that you treat humanity not merely as a means to an end, but at the same time as an end.
…
8a T is morally right if T is the virtuous (honest, charitable,…) action.
8b T is morally right if T is the action that a virtuous person would do.[16]
…

In our view, worries about there being too many alternative meanings associated with the term "innate", or there being widespread confusion over the idea of innateness, are overblown. It is true that fallacious inferences are occasionally drawn in the literature on rationalism and empiricism because of different ideas about innateness. But as we saw earlier, there are equally fallacious inferences drawn in this literature by people who explicitly disavow the idea of innateness (see Chapter 3). We do have some sympathy with the suggestion that is often associated with this type of eliminativism—that to avoid confusion, rather than talking in terms of "innateness", theorists should talk directly in terms of whatever account of innateness they are using. It is for this reason that we framed our characterization of the rationalism-empiricism debate in Chapter 2 directly in terms of the acquisition base. However, as we noted there, we also think that there is nothing wrong with, and much to be gained by, continuing to use the

[16] See Hursthouse (2013), Alexander and Moore (2015), and Sinnott-Armstrong (2015) for an overview of some different proposals regarding the analysis of moral rightness.

term "innate", as this links the contemporary debate about the character and contents of the acquisition base with the traditional philosophical debate about innate ideas. We just need to recognize that the understanding of this term should be grounded in what is at stake in the contemporary rationalism-empiricism debate.

6.2 What Is a Concept?

In this section, we shift our focus from the nature of innateness to the question of what concepts are. As we briefly noted in Chapter 1, there are a number of debates surrounding the nature of concepts and no widely agreed upon account of what concepts are. Our aim here is not to discuss all of the issues at stake in these various debates, much less to attempt to resolve them all.[17] Rather, our aim is simply to set out some essential background regarding a number of these issues so that we can consider how disputes about the nature of concepts interact with the debate between rationalist and empiricist accounts of the origins of concepts and what implications different theories of concepts might have for our case for concept nativism. To some readers, it may seem obvious that the first order of business in a book about concept nativism should be for the authors to state their own theory of concepts. We have not done this because we don't want our case for concept nativism to be hostage to any particular account of the nature of concepts. We want our case for concept nativism to be robust in the face of different theories of concepts. If our case turned on the particular account of concepts that we ourselves favour, it would only matter to other theorists who held the same account of concepts as we do. But if our case for concept nativism is consistent with any of a broad range of approaches, as we will argue it is, then it should be of interest to theorists with very different views about the nature of concepts, as well those who may not be committed to any particular theory of concepts. In light of this, we will proceed first, in this section, by providing a brief overview of theories of concepts in order to convey some of the key theoretical options. Then, in the next section, we will show why our case for concept nativism doesn't depend on adopting any particular account of the nature of concepts.[18]

[17] For a range of different views on concepts and discussion of some of the many debates surrounding the nature of concepts, see Margolis and Laurence (2019) and the papers in Margolis and Laurence (1999, 2015).

[18] Some of the issues that we discuss in this section are ones that we have already touched on to some extent in earlier chapters. But we want to say a bit more about them here, particularly for readers who may be new to debates about concepts and mental representations. For those interested in reading more about our own views on many of these issues regarding the nature of concepts, see particularly Laurence and Margolis (1999); Margolis and Laurence (2007a); and Laurence and Margolis (2012a). But again, we should note that we will largely be putting our own views on these issues to the side in this book, since our defence of concept nativism is meant to be consistent with any of a broad range of different theories of concepts.

To begin, concepts are often understood to be mental representations that are the constituents, or building blocks, of thoughts—the components that make up thoughts and that are shared among different thoughts that have overlapping meanings or contents. Take, for example, the thought that *polar bears are large carnivorous animals*. This thought is made up of the concepts POLAR BEAR, LARGE, CARNIVOROUS, and so on. These same concepts can also be part of other thoughts. For example, the concept POLAR BEAR is also a component of the thought that *polar bears are powerful swimmers*, and the concepts LARGE and ANIMAL are also components of the thought that *blue whales are large animals*. The similarity of these thoughts to the thought that *polar bears are large carnivorous animals* is due to the fact that they share conceptual components—the concept POLAR BEAR, the concept LARGE, and the concept ANIMAL.

In characterizing concepts as mental representations that are the constituents of thoughts, we are relying on an intuitive understanding of what a thought is. This notion can be spelled out further by noting that paradigmatic thoughts are complex representations that are involved in the exercise of psychological capacities that are often described as "high level" cognitive capacities—things like categorization, recalling facts, forming explanations, analogical reasoning, problem solving, and planning a course of action. For example, in deciding what to do about that animal that you see in the distance, it may matter that you first categorize it *as a coyote*, allowing you to infer that it may be dangerous and that you should leave it alone. In contrast, if you were to categorize it as your pet dog, then you'd have reason to act in a completely different manner—to call its name or to walk right up to it. The difference here has to do with whether your thoughts involve the conceptual component COYOTE or MY DOG. Representations of this sort feed into psychological processes involved in remembered events in our lives, our general understanding of related situations, inferences we are prepared to draw about these entities, decision making, and ultimately action. And the complex representations involved in all of these types of processes—from memories of walking your dog to decisions about what to do when you encounter a coyote—all count as thoughts.[19]

Just as thoughts are composed of concepts, some concepts are themselves composed of other concepts. The concept FIVE-LAYER CHOCOLATE SPONGE CAKE WITH CHOCOLATE BUTTERCREAM FROSTING TOPPED WITH DARK CHOCOLATE FLAKES incorporates the concepts FIVE-LAYER CHOCOLATE SPONGE CAKE and TOPPED WITH DARK CHOCOLATE FLAKES, among others, and these are composed of simpler concepts (FIVE-LAYER and DARK CHOCOLATE FLAKES), which in turn are composed of even simpler concepts (LAYER and CHOCOLATE), and so on. There comes a point, however, when we arrive at concepts that are not

[19] For overviews of some of the core psychological phenomena involving concepts, see Murphy (2004); Medin and Rips (2005); and Goldstone et al. (2018).

themselves composed of other concepts or representations. These concepts are known as *primitive concepts*.[20] There has been a great deal of controversy about which concepts are primitive, including whether everyday concepts that aren't obviously composed of other concepts (such as CAKE and TOPPED) have some form of deeper, hidden internal structure. Based on a variety of different methods—in psychology, linguistics, and philosophy—researchers have proposed that certain types of conceptual structure may be fairly common among such concepts (e.g., definitional structure or prototype structure).[21] We won't enter into this debate here, but just want to note that whether a given concept is primitive or not is not something that can be settled a priori or by consulting one's intuitions. It is a broadly empirical question, and identifying the conceptual primitives that ultimately form the basis from which all other concepts are composed is no easy task.

We have said that thoughts and complex concepts are composed of simpler concepts, but what is the rationale for supposing this? One of the most important reasons has to do with what is known as the *productivity of thought*. This refers to the fact that there is no upper bound to the number of distinct thoughts that human beings are capable of forming.[22] One way to appreciate the fact that human thought is productive is by noting a related fact that we briefly mentioned earlier—that nearly every *sentence* that we hear or read is new to us. Given that the process of understanding a sentence commonly involves recovering the thought that it expresses, this means that we are continually entertaining new thoughts. This isn't to say that some sentences aren't used repeatedly ("How are you doing?", "Have a good night!"). But consider the sentences in a novel or the sentences that appear in the latest episode of your favourite television programme. The themes and general ideas may be similar to other sentences that

[20] It is important to note that the use of the term "primitive" here is completely distinct from the use of "primitive" in connection with accounts of innateness and the acquisition base. In debates about the structure of concepts, "primitive" (or "semantically primitive") simply means *semantically unstructured*. This use is unrelated to the use of the term "primitive" in the debate about theories of innateness, where a psychological primitive is understood to be a psychological structure that is not acquired via a psychological-level process—that is, a psychological structure that is in the acquisition base. It is unfortunate that the same term is used with these two very different meanings, but context should make clear which is intended.

[21] Definitional structure and prototype structure are two broad types of conceptual structure. There are a number of different ways of understanding what conceptual structure is and the functions it serves (Laurence and Margolis 1999). Among other functions, conceptual structure is sometimes postulated to play a role in one or more of the following: explaining psychological processes such as perceptual categorization (i.e., applying the concept C to a perceived item), embodying key information that is considered essential to possessing the concept, determining which things the concept refers to or is true of, and explaining how the meaning of a complex concept is a function of the meanings of the concepts it is composed of (Laurence and Margolis 1999). Definitional structure involves constituent representations that collectively specify a definition of the concept. Prototype structure involves constituent representations that collectively specify an abstractly represented best example or central tendency.

[22] Of course, human beings are finite creatures with a limited memory and attention span. The productivity of thought concerns competence (rather than performance), abstracting away from such factors.

you've encountered, but for the vast majority of these sentences, the precise sentences involved will be ones you have never encountered before.

The productivity of thought isn't just a reflection of our facility with language, however. Its effects can be seen by looking more directly at certain types of thoughts. Suppose, for instance, that it is nearing lunch time but you aren't particularly hungry. You might think to yourself I COULD HAVE A SANDWICH OR I COULD GO FOR A WALK. But, of course, these aren't your only options. You might allow yourself to start to think about many other things you could do: I COULD HAVE A SANDWICH OR I COULD GO FOR A WALK OR I COULD TAKE A NAP OR I COULD CALL MY MOTHER. At some point you would of course stop, but notice that there is no limit to the number of possibilities you might consider. In principle, you could keep packing more and more into this thought, simply by making further use of the concept OR to incorporate each additional represented possibility.

Here is another example. Suppose Anne is thinking about some of her friends and acquaintances and trying to plan the menu for a party. She might begin by thinking BETH DOESN'T EAT NUTS, and this might lead her to think of her friend Cathy and think CATHY WAS SURPRISED THAT BETH DOESN'T EAT NUTS (remembering that time when Cathy asked Beth if she likes granola), which might lead her to think DAVE WAS SURPRISED THAT CATHY WAS SURPRISED THAT BETH DOESN'T EAT NUTS (remembering when Dave and Cathy were arguing about what snack to bring for Beth). There is no reason why the same sort of cognitive operation couldn't be repeated further, forming a new thought about someone else's thinking, which itself is about someone else's thinking, and so on.

What these and similar examples suggest is that we are able to entertain an unbounded number of thoughts because the mind is equipped with a finite stock of primitive concepts and with rules for combining and recombining them into more complex concepts and ultimately into full thoughts. Just as a language like English or Spanish can generate an infinite number of sentences by combining a limited number of words in different ways, the mind can generate an infinite number of thoughts by combining its concepts in different ways. The key to this unbounded potential, in either case, is that the combinatorial rules for the system include at least some rules that are *recursive* in that the same rule that generates a given structure can be applied to the structure it generates. For example, the combinatorial rule for disjunction can take as inputs the contents of complete thoughts, *x* and *y*, and generates a new disjunctive thought, *x* OR *y*. Since any thought content can be plugged in for the variables *x* and *y* in the combinatorial rule for disjunction, products of applying this rule can be used as inputs as well. So the rule can form new disjunctions based on previously formed disjunctions (*x* OR *y* can be combined with *z* to form the thought *x* OR *y* OR *z*). Suppose that we also include the further principle that from any thought *x* we can generate a new thought NOT-*x*. Using just these two simple recursive rules, even with just a few initial thoughts to begin with (*x*, *y*, *z*), an infinite number of new thoughts can be

generated—*x*, NOT-*y*, (*x* OR NOT-*y*), NOT-(*x* OR NOT-*y*), (*z* OR NOT-(*x* OR NOT-*y*)), ((*x* OR NOT-*y*) OR (*z* OR NOT-(*x* OR NOT-*y*))), and so on indefinitely. Precisely the same is true if we take *x*, *y*, *z* here to be concepts instead of complete thoughts.

One of the most important aspects of concepts and the thoughts that they compose is that they possess *representational* or *semantic* features. By representational or semantic features, we just mean any features/properties something possesses that are directly related to having some sort of meaning, broadly construed. Traditionally, the term "semantic" was restricted to natural language and used to capture the properties that linguistic items (e.g., English words, phrases, and sentences) were taken to possess in having meaning. But it has since been extended so that it is now also used to refer to meaning-related properties of concepts and thoughts. Just as words and sentences have meanings and representational features, so too do concepts and thoughts. Philosophers also use the terms *possessing intentionality* or *being about* something or *having content* for this same broad property of having representational or semantic features. These terms have different histories and connotations, but they all refer to the same basic thing, and we will use them interchangeably.[23]

Crucially, it should be noted that the term *intentionality* in this context doesn't refer to the intent or purpose of an action or to having an intention to do something. Such states do have intentionality, but intentionality isn't restricted to these types of states. In the philosophy of mind, the term has been used in a deliberately broad way to apply to anything that might be thought to represent or be about something else—not only linguistic and psychological entities (words, sentences, concepts, thoughts, etc.) but any type of sign or symbol. Philosophers also distinguish between *intrinsic intentionality* from *derived intentionality* and distinguish both of these from *as-if* intentionality. Intrinsic intentionality and derived intentionality are both understood in a realistic manner—both involve something genuinely possessing representational properties. These contrast with as-if intentionality, which is understood to be merely metaphorical, as when someone says of her stalled car that isn't restarting that "it just wants to make me late for my appointment". The distinction between intrinsic and derived intentionality, in contrast, has to do with the source of the intentionality. In the case of derived intentionality, the representational properties in question are taken to be real but to be inherited from the representational properties of something else. With intrinsic intentionality, they aren't taken to derive from something else—the representational properties originate with the very entity that possesses them. Arguably, the physical objects that constitute words or sentences—ink on a page, sound waves produced by speaking—have representational properties, but only

[23] In other words, we will use *all* of the following as stylistic variants: *possessing intentionality*, *being about something*, *having content*, *possessing representational features*, and *possessing semantic properties*.

because they are related to concepts and thoughts. If this is the case, then words and sentences (and other public symbols, such as traffic signs) can be said to have derived intentionality, whereas concepts and thoughts would have intrinsic intentionality.

The term that we will make the most use of in this family of terms is *content*. The expression *having content* in this context is a relatively new term, and it is used more for concepts and thoughts than natural language words and sentences, but it refers to the same basic property of having intentionality, having aboutness, or having representational features. Thoughts and concepts are said not only to have content, but to have particular *contents*, which is to say that a given thought or concept has a specific semantic property (or set of properties). At the same time, precisely which sorts of representational properties constitute the contents of concepts and thoughts is a matter of debate. On some accounts, concepts have only referential or truth conditional representational properties. So, for example, the only semantic property that the concept ALAN TURING would have is that it refers to Alan Turing (or, putting the same claim in different terms, it would have as its content the person Alan Turing). Likewise, the only semantic property that APPLES would have is that it has in its extension or truly applies to all and only apples (it has as its content the property of being an apple or perhaps the set of all apples). A different position holds that in addition to referential properties, concepts can have meanings, where the term "meaning" in this particular context is used in a narrower technical sense than we have been using it so far. This technical sense is intended to capture a semantic property that is supposed to explain how two concepts can be distinct from one another even if they have the same referent or extension. For example, the concepts MARK TWAIN and SAMUEL CLEMENS both refer to the same person, but some would argue they differ in content and hence have a different meaning.[24] For present purposes, we will abstract away from the details of different theories of content and their differing views of precisely which semantic properties concepts should be said to have.

Earlier we highlighted the fact that simpler concepts can be composed to form more complex concepts and thoughts and that this is what explains the productivity of thought. This explanation turns on the supposition that the contents of complex concepts and thoughts depend on, or are a function of, the contents of the simpler concepts that compose them and how they are arranged—or, in other words, that there is a *compositional semantics* for concepts and thoughts. The productivity of thought gives good reason to suppose that thought has a

[24] A classic test for such differences in content has to do with whether it is possible for someone to possess a belief vis-à-vis one concept but not the other—for instance, believing that Superman flies but not believing (or outright disbelieving) that Clark Kent flies, despite the fact that Superman is Clark Kent. In a moment, we will return to this type of test, where it figures in an approach to concepts that claims that concepts aren't mental entities.

compositional semantics. And if it does have a compositional semantics, then the content of our thoughts can be seen to ultimately stem from the contents of our concepts.

This still leaves us with the question of where the content of these simpler concepts comes from and ultimately the content of all thought. What is it about a given concept that gives it the specific conceptual content that it has? There are a number of competing approaches to explaining the basis of conceptual content, but two that stand out are conceptual role theories and causal theories. *Conceptual role theories* hold that the content of a concept is largely determined by its distinctive role in cognitive and perceptual processes. Take the concept AND. It contributes to a pattern of inference in which, if someone has the belief *P* and the belief *Q*, then, other things being equal, they are more disposed to form the belief *P* AND *Q* than certain other beliefs as a consequence; likewise, if they have the belief *P* AND *Q*, then they should be similarly disposed to hold the belief *P* and the belief *Q*. According to the conceptual role framework for explaining conceptual content, the concept AND expresses *conjunction* precisely because of the way that it facilitates these patterns. In contrast, *casual theories of content* say that the content of a concept isn't, in the first instance, a matter of its role in the mind. Instead, it is determined by how a concept is causally related to things in the world. For example, the core idea of one such account is that the content of a concept like ZEBRA comes down to the fact that instances of zebras (the animals) reliably cause thoughts involving ZEBRA (with certain added qualifications that we needn't go into here).[25]

On the face of it, conceptual role theories and causal theories take opposite approaches to explaining conceptual content. One is inward looking (focusing on internal psychological processes), while the other is outward looking (focusing on mind-world causal relations). At the same time, it's possible to maintain that the two should be combined in such a way that a concept's content is partly determined by its conceptual role and partly by its mind-world causal relations.[26] Another possibility is to maintain that these two approaches should be kept separate but that each is right for certain types of concepts and that no single approach to conceptual content is appropriate for every type of concept. This second proposal is quite plausible when one considers the sheer variety of concepts in human cognition. This includes concepts that are grounded in bodily experiences (TINGLE, HEADACHE), indexical concepts (THIS, NOW), and concepts for such diverse types of things as individuals (DESCARTES, CHICAGO), properties and relations (WET, ABOVE), actions (JUMP, DECORATE), natural kinds (NITROGEN, TARANTULA), artefacts (KNIFE, CONTRACT), processes (EVAPORATE, DECOMPOSE),

[25] For more on these two general approaches to explaining mental content, see Rey (1997).

[26] There are various ways to do this; see, e.g., Block (1986) and Carey (2009) on two-factor theories of content.

institutions (MARRIAGE, THE DEPARTMENT OF STATE), theoretical constructs (MOLECULE, GRAVITATIONAL WAVE), numerical quantities (FIVE, π), logical relations and quantification (NOT, SOME), imaginary entities (UNICORN, SANTA), modality (MUST, COULD), pluralities (BEES, STARS), time (YESTERDAY, WILL), things we could never experience (THE BIG BANG, TRAVELLING FASTER THAN THE SPEED OF LIGHT), and much else. It is entirely possible that these many different types of concepts' contents will not all be explained in the same way. In any case, since researchers working on theories of content and theories of concepts sometimes focus on an overly narrow range of concepts, it is good to remind ourselves from time to time of some of the breadth of the types of concepts that exists.[27]

So far, much of what we have said about concepts is broadly agreed upon by a diverse group of theorists. We now want to discuss two important areas of disagreement about concepts in order to examine the question of how different approaches to the question of what concepts are might affect the rationalism-empiricism debate regarding the origins of concepts—and ultimately whether such disagreements affect our case for concept nativism. As we will see, the impact on the rationalism-empiricism debate is smaller than one might initially expect.

The first area of disagreement is one that may be surprising to some readers, as it concerns the question of what of kinds of things concepts even are—in particular, whether concepts are psychological entities that occupy a place in the mind. In Chapter 1, we noted the connection between contemporary scientific research on the origins of concepts and historical philosophical discussions of the origins of ideas. The notion of an idea that was in play in some of this earlier philosophical work was that ideas are representational units that enter into psychological processes—for example, that the idea of a dog is a consciously accessible mental image that resembles the appearance of a typical dog. In cognitive science, much of this "idea" idea has been retained. The presupposition that ideas must be consciously accessible and that they resemble what they are about may have been dropped, but in other respects, the "ideas" of philosophers like Hume and Locke are quite similar to cognitive science's "concepts". Both are understood to be mental representations or mental symbols and hence to be psychological entities. We have adopted this view as a background assumption in this section (and in the book) so far.

However, this assumption isn't mandatory. An alternative way of thinking about concepts understands concepts in a way in which they aren't in the mind at all.

[27] The focus on a narrow range of types of concepts in developing theories of content stems in part from the methodological principle of starting with simple cases first and trying to develop some sort of workable theory before confronting the full complexity of phenomena to be explained. This principle is not at all unreasonable given that developing a workable theory of content for even the simplest cases has proven to be enormously difficult. Nonetheless, it is important to guard against distortions that can stem from considering an overly narrow range of examples.

Instead, they are taken to be abstract objects (and so not physical entities located in space and time), not unlike mathematical entities such as numbers and sets. (While numerals and number words—which are signs for numbers—*are* in space and time, numbers themselves on many views are not; you can't bump into or destroy the number 3.) On this approach, concepts are components of the abstract objects that are the meanings associated with sentences and with beliefs, desires, and related types of mental states (see, e.g., Peacocke 1992). Take the sentence "Our galaxy contains billions of stars". This approach holds that this sentence has an abstract object that is its meaning which is made up of meanings for "galaxy", "star", and so on, and that these component meanings are the concepts—as opposed to the mental representations that may be activated in a speaker's mind when using this sentence.

A motivation often connected with taking meanings to be abstract objects is the perceived need for meanings that can function as intermediary semantic entities that stand between a linguistic or mental representation and its referent and embody a particular perspective on the referent (Frege 1892). These intermediaries serve as *modes of presentation*—entities that represent a referent in a particular way, as in the difference between representing Venus *as the morning star* and representing that same entity, Venus, *as the evening star*. Part of the reason to suppose that there are meaning intermediaries of this kind is that this would help to explain the differing cognitive significance of different expressions that have the same referent. For example, someone might hold that "Mark Twain wrote *Huckleberry Finn*" is true but that "Samuel Clemens wrote *Huckleberry Finn*" is false, even though Mark Twain and Samuel Clemens are in fact the same person. The explanation for why this can happen would be that there is more to the semantics of "Mark Twain" and "Samuel Clemens" than their referents. They have modes of presentation too, and their having *different* modes of presentation captures the fact that they involve different ways of construing one and the same referent.[28]

[28] Proponents of the mental representations views of concepts needn't deny that there are abstract objects that are meanings, or that such meanings can play the role of modes-of-presentations. Even though they wouldn't say that concepts *are* meanings, they are free to say that concepts *have* meanings. Then they could adopt much the same explanation for the differing cognitive significance of thoughts involving the concept MARK TWAIN and thoughts involving SAMUEL CLEMENS. They could just say that thoughts of the first type involve a concept (understood as a mental representation) that has a Mark-Twain-mode-of-presentation as its meaning, while thoughts of the second type involve a concept (again understood as a mental representation) that has a Samuel-Clemens-mode-of-presentation as its meaning. At the same time, it's worth noting that some proponents of the mental representations view of concepts hold that there is no need for taking meanings to involve abstract objects that act as intermediaries between representations and their referents. Such a view might hold that the only semantic properties that are really needed are ones that pertain to reference and truth (see, e.g., Fodor 1998; Fodor and Pylyshyn 2015). A view like this might nevertheless hold that there are modes of presentation by taking mental representations themselves to be modes of presentation, where different mental representations with the same referent effectively constitute modes of presentation by representing that referent in different ways. For further discussion of the different options

While we favour the mental representations view of concepts and adopt this view as a background assumption as we develop our case for concept nativism throughout the book, this commitment is not required in order to take part in the rationalism-empiricism debate about concepts, nor is it essential to our case for concept nativism. To draw out these points, we will briefly explain how the psychological and developmental issues that we are couching within a mental representations framework can be recast in the abstract objects framework.

To begin, we should point out that the terminology here can be confusing since the mental representations framework and the abstract objects framework both make use of the terms "thought" and "concept" but use these terms in systematically different ways. Within the mental representations framework, thoughts are generally the sorts of mental representations that have full propositional content, and concepts are the subpropositional mental representations that make up a thought. On the abstract objects framework, thoughts are understood to be the meanings of sentences and of corresponding psychological states, and concepts are the meaning components (modes of presentation) that make up a thought. Despite these systematic differences, though, the abstract objects framework can draw much the same distinctions as the mental representations framework and offer broadly similar explanations for many phenomena. For example, the abstract objects framework can draw a distinction between primitive and complex concepts that is parallel to the distinction between primitive and complex concepts on the mental representations view. It's just that in one case this distinction is applied to mental representations and in the other it is applied to the abstract objects that constitute meanings.

What does the abstract objects framework say about the role of concepts in mental processes? Although this framework takes concepts to be abstract objects (which aren't in the mind) rather than mental representations (which are), it still makes reference to an agent's concepts in explaining aspects of her psychology. To see how this can be, let's adopt the assumption that even though a proponent of the abstract objects view doesn't identify concepts with mental representations, she still supposes that mental representations occur in episodes of thinking.[29] Now in the mental representations framework, concepts (understood as mental representations) are directly involved in mental processes such as categorization. When an agent is confronted with a polar bear, and categorizes it as a polar bear, the concept POLAR BEAR (a mental representation) is activated in the process. In the abstract objects framework, the psychological import of a concept is just as real but indirect, deriving from the semantic relation it bears to mental

regarding the abstract objects framework, the mental representations framework, and whether it pays to mix the two, see Margolis and Laurence (2007a).

[29] While we think there are good reasons for proponents of the abstract objects framework to adopt this assumption, here we are just making it for simplicity. In any case, the general point that follows in the text applies whether the assumption is made or not.

states. In this case, a mental representation is also activated, but crucially it must be one that has the concept *polar bear* (an abstract object that is its meaning) as its mode of presentation, as opposed to one with some other mode of presentation. Either way, a concept factors into the categorization process and is a key part of the explanation of why the agent responds as she does. So, even though concepts aren't psychological entities in the abstract objects framework, they can still play a significant role in psychological processes. It is just that this is the indirect role of being the meanings of the mental representations that are directly involved in such processes.

In much the same way, the rationalism-empiricism debate about the origins of concepts can also be reconstructed within the abstract objects framework. In Chapter 2, we characterized the contrast between rationalism and empiricism in terms of their differing views of the acquisition base—the collection of psychological structures whose acquisition is not mediated by more fundamental psychological traits. And we noted that rationalist accounts of the origins of concepts hold that concepts in more than just a few different domains are either innate or acquired via rationalist learning mechanisms, while empiricist views of the origins of concepts hold that domain-general learning mechanisms largely suffice for acquiring all concepts. Now a theorist advocating the abstract objects view of concepts can accept almost all of this as is, but rather than understanding the debate between concept nativism and concept empiricism to be about the psychological origins of *concepts* (which they take to be abstract objects that are the modes of presentation associated with mental representations), they will instead understand it to be about the psychological origins of *the mental representations that have concepts as their meanings*. Likewise, rather than understanding learning mechanisms to be mechanisms that support the acquisition of new concepts, they will understand them to be *mechanisms that support the acquisition of new mental representations that have concepts as their meanings*. And rather than understanding innate concepts as mental representations that are part of the acquisition base, they will understand them to be *the meanings of mental representations that are part of the acquisition base*.

What all this shows is that the rationalism-empiricism debate regarding the origins of concepts doesn't force theorists to adopt a particular view about whether concepts are psychological entities or not. It can be formulated to accommodate either framework—the mental representations framework or the abstract objects framework—and other frameworks as well, as long as suitable adjustments are kept in mind.[30]

[30] We have focused on the mental representations and abstract objects frameworks for illustrative purposes and because they are the two most prominent frameworks for understanding the nature of concepts. Our aim here is not to provide an exhaustive discussion. In any case, much the same points apply to other frameworks. (See Margolis and Laurence 2007a for broader discussion of different frameworks for understanding the nature of concepts.)

Let's turn now to the second area of disagreement about concepts that we want to look at, the issue of how to distinguish *conceptual* states from *nonconceptual* states. The guiding idea behind this second area of disagreement is that mental representations should not be seen as forming a homogeneous category and that conceptual representations (concepts) are a special or distinctive type of mental representation. Different ways of drawing the conceptual/nonconceptual distinction that philosophers and cognitive scientists have suggested highlight different features as marking what is special about conceptual, as opposed to nonconceptual, representations. We will just present some of the main approaches in order to give a sense of what the debate looks like, and because these different ways of drawing the distinction will come up again later in the book.[31] As we will see, the different ways of drawing the conceptual/nonconceptual distinction are not equivalent with one another; a representation that counts as a concept on one account may count as a nonconceptual representation on other accounts, and vice versa.

One family of approaches to the conceptual/nonconceptual distinction is centred around a cluster of properties relating to the *richness* and *fine-grainedness* that are characteristic of some types of representations by contrast with the abstractness or coarse-grainedness of other types of representations. Consider, for example, some of the representations involved in perceptual experience. A characteristic feature of perceptual experiences is the richness of detail that they represent (see, e.g., Dretske 1981; DeBellis 1995; Carruthers 2000; Peacocke 2001). Imagine seeing a vase of roses. This visual experience would incorporate enormous amounts of detail pertaining to the particular shapes of the petals, the relative heights of the stems, the pattern made by the roses' thorns, the shadows that appear below the flowers towards the front, and so on. Now contrast this rich experience with simply having a belief that some roses are in a vase. Such a belief would invariably abstract away from most of the perceived detail. It may capture that there are roses there, but it will gloss over the detailed particularities about how they look at that very moment that are all part of the visual experience of the roses in the vase. In this way, the richness of experience is similar to the idea that is behind the old cliché that a picture is worth a thousand words; experiences, like pictures, pack in far more detail than a verbal description (or a corresponding thought) about the very same situation.

A different, but related, contrast focuses not on richness but on the fine-grained character of some of the representations involved in perceptual experience—for example, the fact that a given perceptual representation represents a highly specific shape or colour (see, e.g., Peacocke 1992; Tye 1995; Heck

[31] To simplify our presentation of these views, we will continue to frame the discussion in terms of the mental representations framework. It should be noted, however, that some participants in the debate don't subscribe to this framework. In addition, this whole debate is sometimes put as a debate about how to distinguish between two types of content (conceptual content vs. nonconceptual content) as opposed to how to distinguish between two types of psychological states (conceptual states vs. nonconceptual states). For present purposes, these differences won't matter. See Laurence and Margolis (2012a) for reasons for framing this issue in terms of states rather than content and for discussion of further ways of distinguishing the conceptual from the nonconceptual.

2000). Take colour. Human vision can distinguish between millions of different shades of colour. By comparison, there are *far* fewer colour concepts that are expressed in natural language words. English, for example, is a language that is relatively rich in colour terms, but it still has only about eleven basic colour terms. And even if we include non-basic colour terms, including ones that rely on compound expressions (e.g., "salmon pink"), the number of colour terms is orders of magnitude lower than the number of discriminable colours. This observation is closely related to facts concerning our ability to identify and recall colours. If shown a specific shade of red (red_{27}), it's easy to recognize that it is red, to say that it is "red", or to identify it as the same general colour as a previously seen sample. But when it comes to the precise shade that is experienced—not just red but the very specific shade of red that it is—the situation changes. We might be able to see that it differs from a nearby shade of red (red_{29}) when they stand side by side, but then not be able to say which of these two was previously seen even for a sample that was presented to us just a few moments ago.

Taken together, the richness of perceptual experience and its fine-grained character motivate a family of views about how to distinguish conceptual representations from nonconceptual ones. Such views say that nonconceptual representations are the ones that are responsible for these aspects of perceptual experience—they are the representations that account for how perceptual experience is able to pack in so much detail at a given time, including extremely fine-grained details. Or alternatively, they are the representations that support rich perceptual experience but which are unable to support the ability to remember or reidentify these fine-grained details. By contrast, concepts would be representations that are less rich, more coarse-grained, or capable of supporting the ability to remember or reidentify the representational content that they express.

A second family of views concerning the conceptual/nonconceptual distinction focuses on whether a representation is constrained by, or factors into, what makes someone a *rational* agent. Different accounts in this family of views adopt different standards on the requirements for being rational. One relatively modest account emphasizes the need to avoid manifest contradictions (Crane 1988). On this type of account, if F is a concept, it would not be possible to judge of the same thing (at the same time, in the same respect) both that it is F and that it isn't F—for example, that it is red and that it isn't red, or that it is a flower and that it isn't a flower. In contrast, if F were nonconceptual, it might be possible to entertain contradictory representations of this sort. Crane points to the motion aftereffect illusion as a possible example where observers simultaneously entertain contradictory representations. This illusion occurs when you stare at a scene that contains motion in one direction (like a waterfall) and then shift your attention to a motionless object. What happens is that the object appears to be moving in the opposite direction of the original motion and, at the same time, appears to remain still. Since this experience is inherently contradictory—the object looks

like it is both moving and not moving—Crane suggests that the representation of the movement is nonconceptual.

Adopting a somewhat different standard of rationality, one could focus on a representation's being able to enter into reasons for actions as a mark of it being conceptual. This idea is connected to the discussion of reasons for actions in Hurley (2003). On Hurley's view, reasons for actions might be thought of as emerging as soon as means-end decision making is in place. That is, if an organism is able to select a course of action based on its representing that action as a way to achieve its goals given how it represents the world to be at that time, then the representations that are involved in this action-guiding process count as reasons for actions. A hallmark of means-end decision making is that it involves a certain degree of cognitive and behavioural flexibility—being able to use different means to achieve the same end and being able to use the same means for different ends. This contrasts with inflexible forms of behaviour, such as a fixed-action pattern response that is simply triggered when the representation for a particular stimulus condition is activated (as in the case mentioned in Chapter 4 in which, in frogs, a visual representation of a small object in motion triggers a rigid type of feeding response). Hurley (2003) stops short of taking any reasons of the type that go with means-end decision making to involve concepts, instead seeing these representations as concept-like, or as coming close to being fully conceptual. But others might want to use the very distinction she makes regarding having and not having reasons as the basis for drawing the conceptual/nonconceptual distinction. In that case, the representations that are capable of entering into means-end reasoning would be considered concepts. By contrast, the representation of a stimulus that simply activates a fixed-action pattern response would not.

A related but more general form of this type of view might take concepts to be distinguished by their ability to enter into general forms of abstract reasoning. For example, a representation might count as conceptual if it isn't confined to operations in a relatively isolated psychological mechanism but is part of a system of representations that is inferentially integrated with higher cognitive processes such as planning and decision making, episodic memory, interpreting discourse, counterfactual reasoning, causal-explanatory reasoning, and problem solving. For instance, on an account like this, if a representation participates in an inferential process in which an agent is working out a causal model for understanding a puzzling feature of the world, it would be considered conceptual. In contrast, representations that couldn't possibly enter into such a process (e.g., because they are locked inside a system that is dedicated to representing the prosodic contours of spoken language) would be considered nonconceptual.

Another view in this family emphasizes a much more stringent connection to rationality considerations by taking conceptual representations (but not nonconceptual representations) to be associated with a highly reflective capacity to justify one's use of a concept and ultimately to justify one's beliefs about the world

(e.g., Brandom 1994; McDowell 1994). On theories of this type, representations that are not integrated with self-reflective and linguistic capacities would generally be deemed nonconceptual.

A third family of views about the conceptual/nonconceptual distinction turns on what is known as the *Generality Constraint* (Evans 1982). Evans describes this constraint in terms of the range of thoughts that someone must be able to entertain on the assumption that they have a given concept-involving thought. "If a subject can be credited with the thought that *a is F*, then he must have the conceptual resources for entertaining the thought that *a is G*, for every property of being *G* of which he has a conception" (p. 104). More intuitively, what this means is that if a thought is truly composed of concepts, then these concepts must be able to freely recombine with other concepts to form other related thoughts—not necessarily thoughts that the subject actually believes, but at least thoughts that she is capable of entertaining. For example, if a mental representation with the content that *Jack is artistic* is to be taken to be made up of concepts, including the concept ARTISTIC, then someone who can entertain this representation and who can also represent Alice must also be capable of entertaining a representation with the content that *Alice is artistic*, as well.

There are different ways of developing a view of concepts that is organized around the Generality Constraint. The main variable that differentiates between such accounts has to do with how extensive the combinability requirement is taken to be (Camp 2004). In the limit, it might be held that for a representation to be conceptual, it must be able to freely recombine with all of the other representations in higher cognition with few constraints. Alternatively, it might be held that there are semantic constraints on the ways that predicative representations recombine with representations of individuals and with each other, to rule out what might be thought to be incoherent thoughts, such as the thought that *the number two is artistic*. In this case, the concept ARTISTIC would be restricted to the sorts of things that have minds and agency. Regardless of which of these or related constraints are employed, the idea behind all accounts based on the Generality Constraint is that representations that fail to meet the specified combinability condition don't count as concepts; they are nonconceptual representations.

The last family of views concerning the conceptual/nonconceptual distinction that we will mention is broadly focused on the distinction between *discursive* and *iconic* representations. Within this family of views, conceptual representations are thought to be discursive and nonconceptual representations to be iconic (Fodor 2007, 2008). How to draw the discursive/iconic distinction is itself very much a matter of debate. Sometimes discursive representations are simply said to have a digital format and iconic representations to have an analogue format. Sometimes discursive representations are said to be distinguished from iconic ones in that iconic representations simultaneously represent multiple properties

of an object, as when a part of a picture of a face—the part corresponding to just the nose—simultaneously represents its colour, shape, and texture. In contrast, a discursive representation that might be used to describe the same object would use distinct words, each representing a particular property (e.g., words like "dark", "smooth", "pointed"). Fodor links the discursive/iconic distinction with several others. Here we will focus on Fodor's primary way of drawing the distinction, which has to do with the way that the parts of a representation relate to the whole. He suggests that discursive representations, unlike iconic representations, have a canonical decomposition. In practice, this means that there is a correct way to subdivide a discursive representation into its representational parts and that not every division would yield parts that are interpretable or that contribute to the semantics of the whole.

Take the sentence "Sue kicked the red ball". Some subdivisions constitute interpretable parts of the sentence ("Sue", "the red ball"), while others don't ("Sue…red", "kicked the"). In contrast, iconic representations are taken to lack canonical decompositions, making any part of an iconic representation just as interpretable as any other part. In explaining in more detail what makes an iconic representation iconic, Fodor contrasts sentences (which are discursive) with photographs (which he takes to be iconic). He points out that photographs and other iconic representations obey what he calls the *Picture Principle*: "If P is a picture of X, the parts of P are pictures of parts of X" (Fodor 2008, p. 173). In other words, any part of an iconic representation represents part of what the whole represents. For example, if you have a picture and cut it into three parts in a random way, the three parts are still perfectly interpretable. They would represent the corresponding parts of scene depicted in the photo. But this isn't possible with a sentence.

The different ways of drawing the conceptual/nonconceptual distinction that we have been outlining sometimes differ over whether they consider a particular type of representation to be conceptual. We will discuss a number of examples where this is the case later in the book. For now we will just mention two examples.

The first example involves representations that are confined to low-level specialized processing systems, for example, the system that analyses the sound structure of a sentence. This is usually understood to involve representations of various distinctive features, such as whether a speech sound is voiced [+/-voiced] (whether any vibration in the vocal cords is involved), or nasal [+/-nasal] (whether air flows through the nose). Distinctive features contribute to the representations of phonemes (sound units in spoken natural languages) like the sounds /b/ and /p/ in the words "bad" and "pad", which differ only in that /b/ is voiced and /p/ is not.

Let's first consider the status of representations of distinctive features and phonemes for a theory like Fodor's in which conceptual representations are

discursive rather than iconic. Since these kinds of representations don't conform to Fodor's picture principle, they count as discursive rather than iconic. This makes them conceptual representations for Fodor. Distinctive features and phonemes are also conceptual representations according to the criterion that identifies nonconceptual representations with fine-grainedness. This is because such representations abstract away from all sorts of fine-grained perceptible differences and are readily reidentifiable. The phoneme /b/, for example, represents a wide range of sounds as all being instances of the very same phoneme despite differing in such things as pitch, timbre, volume, and duration. This is the whole point of the basic sound units of spoken natural languages—being capable of abstracting away from many acoustical details and allowing sounds to be reidentified across speakers and in different contexts. If they didn't have these properties, they wouldn't be much use for linguistic communication, as we wouldn't be able to understand the same person speaking in quiet and loud voices, or different people with different voices, as uttering the same words.

Clearly, though, these kinds of phonological representations are not conceptual representations according to some of the other ways of drawing the conceptual/nonconceptual distinction. For example, since such representations don't freely combine with representations outside of the language processing system, they don't satisfy the Generality Constraint. They also aren't involved in decision making, causal-explanatory reasoning, or other forms of higher-level cognition, and so fail to meet the criteria of being conceptual for a variety of accounts that tie concepts to having reasons or being rational. On these other ways of drawing the conceptual/nonconceptual distinction, these kinds of phonological representations are nonconceptual.

Our second example concerns representations of particular fine-grained colours. Imagine looking at a uniform colour patch of a particular shade of blue in good lighting conditions. The fact that you are having a fine-grained visual experience—not just of blue but of this very specific fine-grained shade of blue (say, blue_{32})—means that the representation involved in the experience counts as nonconceptual on some ways of drawing the conceptual/nonconceptual distinction (see, e.g., Peacocke 1992). However, on some other accounts, your fine-grained colour experience in this type of case turns out to be conceptual. For example, McDowell (1994) argues that such experiences are conceptual on his rationality-based way of drawing the conceptual/nonconceptual distinction. He does acknowledge that this type of example may not at first seem particularly conceptual on his account of concepts, since these experiences may not seem suited to providing reasons that can justify one's beliefs about the world. But he thinks this initial impression is misleading in that fine-grained perceptual experiences can actually be rationally integrated into self-reflective forms of thinking. The core idea of his proposal is that fine-grained experiential contents can be singled out by using complex concepts that incorporate a demonstrative

component. The content of these concepts is basically the same as a natural language expression like "that shade", which is used to pick out a particular fine-grained colour shade. He takes the fact that this content can be expressed in natural language to argue that the content of the experience *can* be integrated with paradigmatically high-level cognition and that the representation that supports the fine-grained character of this colour experience ends up being conceptual after all.

As is clear from this brief overview of different theories of concepts, the conceptual/nonconceptual distinction can be drawn in a variety of different ways. And, while they will overlap in many cases, these different ways of drawing the distinction are not all equivalent with one another. There will be cases where a given representation is conceptual on one way of drawing the conceptual/nonconceptual distinction but nonconceptual on a different way of drawing this distinction. In the following section, we consider the question of what the implications of this might be for the rationalism-empiricism debate regarding the origins of concepts.

6.3 Concept Nativism Is about More Than Just Innate Concepts

One of our goals in this chapter is to make clear why our case for a rationalist account of the origins of concepts is robust in the face of different accounts of what concepts are and different ways of drawing the conceptual/nonconceptual distinction. In the last section, we showed that the case for concept nativism doesn't depend on whether concepts are taken to be mental representations which *have* meanings or abstract objects which *are* meanings. In this section, we will show that our account also doesn't depend on any particular way of drawing the conceptual/nonconceptual distinction. The key to seeing why it doesn't is recognizing that concept nativism is about more than just innate concepts.

First, though, it's worth pausing for a moment to see why someone might think there is a problem here—why someone might think that the whole debate about the status of concept nativism can't even get started without an agreed upon way of drawing the conceptual/nonconceptual distinction. Suppose that concept nativism was just a thesis about whether there are any innate concepts and if so, how many. Suppose it was also agreed that the representations R_1–R_n are innate and that these are the only innate representations. Someone might then argue as follows. A theorist who draws the conceptual/nonconceptual distinction in such a way that many of the representations R_1–R_n turn out to be concepts may well be committed to there being a significant number of innate concepts and thereby endorse concept nativism. But a theorist who draws the conceptual/nonconceptual distinction in a different way, according to which none of R_1–R_n are concepts, would hold that there are no innate concepts and thereby be opposed to concept

nativism—even though the two theorists completely agree about which representations are innate. In that case, the shift from *endorsing concept nativism* to *being opposed to concept nativism* would turn entirely on how the conceptual/nonconceptual distinction is drawn. So the status of concept nativism is hostage to how the conceptual/nonconceptual distinction is drawn and cannot be productively debated without first settling the question of how we should draw the conceptual/nonconceptual distinction.

While this argument may seem quite plausible at first, it is in fact deeply flawed. Seeing where and how it goes wrong is of fundamental importance to understanding the debate over concept nativism. As we will show, not only can this debate proceed before settling the question of how the conceptual/nonconceptual distinction should be drawn, it is actually *better* to conduct the debate in a way that remains neutral among different ways of drawing the conceptual/nonconceptual distinction.

The first point to note is that what makes a view rationalist is ultimately a matter of whether it posits a rationalist (as opposed to an empiricist) acquisition base in its account of the origins of the traits in question, not whether it says that these traits are innate. As we emphasized in Chapter 2, there are many ways to be a rationalist regarding the origins of a given trait. Concept nativism is compatible with a large range of views that take there to be a rationalist account of the origins of concepts. In a local debate concerning the origins of a particular concept or a particular conceptual cluster, a concept nativist is simply a theorist who takes the local acquisition base for this concept or conceptual cluster to be a rationalist one, containing characteristically rationalist psychological structures. In the global rationalism-empiricism debate about the conceptual system taken as a whole, a concept nativist would be someone who adopts a rationalist view of the origins of concepts in more than just a few conceptual domains.[32] Our own version of concept nativism holds that characteristically rationalist psychological structures in the acquisition base are involved in the acquisition of many concepts across many different conceptual domains. Put in different terms, our view is that many concepts across many different conceptual domains are either innate (in the sense of being in the acquisition base themselves) or else they are acquired via rationalist learning mechanisms that trace back to characteristically rationalist psychological structures in the acquisition base.

The key point here is that this is a disjunction—concept nativism turns on the acquisition base *either* containing innate concepts *or* containing characteristically rationalist psychological structures that contribute to rationalist learning mechanisms that underpin concept acquisition. Given this disjunction, it is possible

[32] As we noted in Chapter 2, this way of expressing what concept nativism consists in is a convenient shorthand for a more nuanced view that takes into account potential trade-offs between a variety of independent factors that contribute to making an account rationalist to a given extent. We return to this point below, providing a fuller and more accurate account of what concept nativism consists in there.

that most of the concepts that a particular version of concept nativism provides a rationalist acquisition account of aren't actually innate but instead are acquired via rationalist learning mechanisms. In fact, it is possible to hold a very strong version of concept nativism in which *no concepts at all are innate*. Many concepts across many different conceptual domains might still be acquired via rationalist learning mechanisms, and the acquisition base that underpins the acquisition of these concepts might be exceedingly rich in terms of the characteristically rationalist psychological structures that it contains.

This case dramatically illustrates how concept nativism isn't just about innate concepts and, at the same time, points to a key reason why it's a mistake to suppose that the conceptual/nonconceptual distinction must be settled in advance of debates about concept nativism. What matters to concept nativism isn't how many concepts are innate or even whether any are, but rather, the number and diversity of concepts whose origins are explained in rationalist terms, that is, in terms of their being either innate *or acquired via rationalist learning mechanisms*. Our own view is that, most likely, the concepts for which a rationalist account is appropriate will be a mix—some are likely innate, but many others will be acquired on the basis of rationalist learning mechanisms. That is why we have repeatedly emphasized the fact that rationalism is perfectly compatible with learning. The difference between rationalism and empiricism—and likewise between concept nativism and concept empiricism—isn't that one accepts learning and the other does not. It is about the character of the learning mechanisms involved, with rationalists, but not empiricists, taking rationalist learning mechanisms to play a substantial role in development.

A second problem with the argument that ties concept nativism to the question of how to draw the conceptual/nonconceptual distinction is that, as it turns out, for many of the cases where we will be arguing for a rationalist account of the origins of concepts in a particular domain, the representations that we will be discussing will be ones that come out as conceptual on all or nearly all ways of drawing the conceptual/nonconceptual distinction. For example, we will see that there is a debate about whether infants are capable of representing and reasoning about mental states as opposed to responding to low-level perceptual properties among the stimuli they have been tested on. However, there are specific claims within this debate that are understood by virtually everyone in the debate to be about the use of a concept, such as the concept BELIEF. For instance, virtually everyone agrees that *if* it could be shown that infants' expectations about how an agent will act turn on their distinguishing between cases where the agent has a false as opposed to a true belief and that they do so in physically disparate situations, then they are making use of the concept BELIEF.[33]

[33] Of course, there will always be ways of drawing the distinction on which it won't come out as conceptual (McDowell 1994 may be an example). But we think that theorists who deny this are either mistaken about the consequences of their own views or else adopting a view that is so extreme that it

Much the same will apply for a great many of the cases we will be discussing, which involve highly abstract, coarse-grained, non-iconic representations that are employed in reasoning, decision making, and other high-level cognitive processes, freely combining and recombining with other conceptual representations. To anticipate just a few examples, these include moral concepts like FAIRNESS, logical concepts like OR, emotion concepts like PRIDE, living kind concepts like ANIMAL and PLANT, and social concepts like COMMUNICATION and COALITION. In these, and in a great many other examples, the representations we will be discussing are ones that are conceptual on pretty much every way of drawing the conceptual/nonconceptual distinction (though of course not all theorists would endorse a rationalist theory of their origins). While there will be borderline cases where theorists who draw the conceptual/nonconceptual distinction in different ways disagree, these cases do not mean that the status of concept nativism is hostage to how this distinction is drawn or that one must first settle the question of how this distinction is drawn before concept nativism can be productively debated. Given the widespread agreement regarding the conceptual status of the bulk of the cases that we will be discussing, these consequences would not follow even if there were nothing to say about the cases where there is disagreement.

As it turns out, though, more can be said about these sorts of borderline cases, where a given type of representation turns out to be nonconceptual on some theories but not others. Let's assume a rationalist account of the origins of some representations of this sort and an account of concepts where they comes out nonconceptual. Then a third problem arises for the argument tying concept nativism to how the conceptual/nonconceptual distinction is drawn. The point to notice is that if the rationalist account of the origins of this type of representation is right, then the acquisition base is that much richer in light of this case—either these representations themselves are part of the acquisition base or the rationalist learning mechanism that leads to them being acquired traces back to characteristically rationalist psychological structures that are part of the acquisition base. And this richer acquisition base is likely to support the acquisition of other representations that *are* conceptual.

To make the discussion here concrete, let's return to the example of geometrical representations that we discussed in Chapter 2.[34] Recall that we used Spelke et al. (2010)'s account of the origins of geometrical concepts as an example of a rationalist learning mechanism. On Spelke et al.'s account, children acquire Euclidean geometrical concepts by drawing on two sets of geometrical representations that are at the borderline between the conceptual and the nonconceptual,

is of little interest to theorists in cognitive science. The standard for being conceptual on McDowell's view, for example, may well be so high that it implies that many adult humans don't have any concepts (Kornblith 2002). Moreover, the other two points that we make in this discussion would still apply in such cases anyway.

[34] We will discuss a variety of further cases in connection with discussions of different content domains as these come up in later parts of the book.

coming out as conceptual on some ways of drawing the conceptual/nonconceptual distinction, and nonconceptual on other ways of drawing this distinction. One of these sets of representations is involved in representing the geometrical features of an environment. The other is involved in representing the shapes of small moveable and manipulable objects. On Spelke et al.'s account, both sets of representations are innate (or acquired via an innate rationalist learning mechanism). An argument could be made that the representations in either or both of these sets count as concepts and thus directly support concept nativism. However, given a different way of drawing the conceptual/nonconceptual distinction, it could also be argued in each case that the representations do not count as concepts.

Let's first consider how the representations of the geometry of an area (which are used to reorient to an environment following disorientation) are affected by different ways of drawing the conceptual/nonconceptual distinction. These representations will come out as conceptual on at least some accounts in the family of accounts that draw the conceptual/nonconceptual distinction in terms of fine-grainedness (nonconceptual) vs. coarse-grainedness (conceptual). Since they can be stored in memory and used to reidentify geometrical features in order to reorient oneself in an environment after disorientation, they will count as coarse-grained, and therefore conceptual, on accounts that emphasize these aspects in drawing the conceptual/nonconceptual distinction. On the other hand, on accounts like many of those in the rationality family of accounts or the Generality Constraint family of accounts, these same representations would be considered nonconceptual. This is because these geometrical representations are assumed to only be used for reorientation and so are not subject to the sorts of rational or combinatorial constraints that these type of accounts generally require.

The status of the shape representations that are used in the perceptual categorization of the shapes of moveable and manipulable objects—the circularity of a plate, the right angle on the corner of a box, and so on—is similarly affected by different ways of drawing the conceptual/nonconceptual distinction. Interestingly, accounts that draw the conceptual/nonconceptual distinction in terms of a rationality constraint might well take these representations to be conceptual. Categorization judgements about the shapes of objects using such representations could readily be incorporated into decision making and guide actions in a way that is subject to appropriate rational constraints, for example. On the other hand, as visual representations, these representations don't freely combine with other clearly conceptual representations in the way that the Generality Constraint might require for them to count as conceptual. So again, these would count as nonconceptual on at least some accounts.

Let's assume that for both of these sets of representations, the representations come out as nonconceptual. The key point to notice is that even in this case, if these representations were also taken to contribute to the acquisition of Euclidean

concepts, as on Spelke et al.'s account, then these nonconceptual representations would feed into a rationalist learning mechanism for acquiring other representations which are clearly conceptual. In fact, this is the whole idea behind Spelke et al.'s account, that the representations that derive from her two innate special-purpose mechanisms feed into a process which forms the Euclidean geometrical representations that we use in general thought and reasoning—and thereby explains how we acquire paradigmatic geometrical concepts. What makes Spelke et al.'s account of the origins of Euclidean geometrical concepts a rationalist account isn't that these concepts are taken to be innate but that they are acquired via a rationalist learning mechanism that traces back to characteristically rationalist psychological structures in the acquisition base. This nicely illustrates how a rationalist account of the origins of a set of representations can support concept nativism regardless of whether the representations are taken to be conceptual or nonconceptual.

In short, there are a number of problems with the argument that concept nativism is hostage to how the conceptual/nonconceptual distinction is drawn. We have seen that concept nativism in general, and our own version of concept nativism in particular, is largely independent of any particular theory of concepts and is robust in the face of disagreements about how to draw the conceptual/nonconceptual distinction. Concept nativism is not just about innate concepts. It is fully compatible even with there being no innate concepts at all. The bulk of the cases that argue for concept nativism, and for our version of concept nativism, consist of examples where there is very widespread agreement regarding the fact that the representations involved are conceptual. But, even in the borderline cases which are disputed, taking the representations involved to be nonconceptual doesn't mean that they add nothing to our defence of concept nativism. On the contrary, where there is a strong argument for a rationalist account of the origins of such representations, that shows that we need to accept a richer acquisition base and that these representations will likely be involved in rationalist learning mechanisms that account for the origins of other representations that *are* uncontroversially conceptual. And the fact that these concepts are acquired by rationalist learning mechanisms directly supports the case for concept nativism.

To round out our discussion of what concept nativism does and does not claim, we will briefly return to the question of what makes an account of the origins of concepts more rationalist, building on our earlier discussion in Chapter 2. This will also put us in a position to say a bit more about the type of concept nativism that we will argue for in the rest of the book.

Recall that in Chapter 2 we identified seven dimensions along which an account might be rationalist or empiricist to a greater or lesser degree. Box 7 (repeated here from Chapter 2) provides a reminder of those seven dimensions.

Box 7: Dimensions of Variation for Positions in Rationalism-Empiricism Debates

1. **Quantity**—quantity concerns the number of psychological structures in the acquisition base (particularly the number of characteristically rationalist psychological structures).

2. **Complexity**—complexity concerns the complexity of psychological structures in the acquisition base (particularly of characteristically rationalist psychological structures).

3. **Degree of Articulation**—degree of articulation concerns the extent to which psychological structures are already elaborated into their full mature form in the acquisition base (particularly for characteristically rationalist psychological structures).

4. **Diversity of Content Domains**—diversity of content domains concerns the set of domains targeted by all of the domain-specific psychological structures in the acquisition base taken together. Each domain-specific psychological structure in the acquisition base targets just one or a small number of domains, but collectively these domain-specific structures may target a wider range of domains. The diversity of content domains is the extent to which this full set of targeted domains is diverse.

5. **Degree of Abstractness**—degree of abstractness concerns the semantic distance between the content of a given representation in the acquisition base and that of the mind's lowest-level sensorimotor representations.

6. **Degree of Domain Specificity**—degree of domain specificity concerns the extent to which characteristically rationalist psychological structures in the acquisition base are domain specific.

7. **Degree of Alignment**—degree of alignment concerns the extent to which two domains are aligned with one another, namely, the target domain (the domain that a learning mechanism is directed at) and the resource domain (the domain that the innate resource which the learning mechanism traces back to is directed at). The more closely related the target domain and the resource domain are, the greater the extent of alignment between them.

We can now see that all seven of these dimensions apply just as much to the debate about the origins of concepts as they do to the general rationalism-empiricism debate. Since concept nativism does not turn solely on the question of which concepts are taken to be innate—it turns more generally on whether a rationalist or an empiricist account of the origins of concepts is correct—an account might be more (or less) rationalist or more (or less) empiricist in light of any of these seven factors.

For example, although Spelke et al.'s account of the origins of Euclidean geometrical concepts is clearly a rationalist account, a competing rationalist account might hold that such concepts are acquired via a learning mechanism that is similar to Spelke et al.'s but one which postulates a further innate mechanism in the acquisition base that is specific to the interpretation of scale models, or via a learning mechanism that is very much like Spelke et al.'s but where some of the links between its innate domain-specific components are preformed in the acquisition base. These would clearly be more rationalist than her account—the first regarding the dimension of quantity, the second regarding degree of articulation. A third possibility might involve a rationalist account of the origins of these concepts which incorporates a small number of innate Euclidean geometrical concepts or an innate system which already represents distance, angle, and direction—so that the resource in the acquisition base that the learning traces back to is also directed at the domain of Euclidean geometrical concepts. In that case, the local acquisition base would contain a resource that is more closely aligned with the target domain than the structures in the local acquisition base envisioned by Spelke et al., making it a more rationalist account of the origins of such concepts along the alignment dimension. And, of course, an account like Spelke et al.'s could likewise be made more *empiricist* by manipulating these dimensions in the direction of being more empiricist rather than more rationalist.

As we noted in Chapter 2, our main purpose in highlighting the range of different dimensions along which rationalist and empiricist accounts can vary is to counteract the unfortunate tendency to focus on just one or a small number of these to the exclusion of the others in considering what makes an account rationalist or empiricist to the extent that it is. That same point applies in considering the range of dimensions along which rationalist and empiricist accounts of the origins of concepts can vary. The idea that concept nativism is only about whether any concepts are themselves innate (or how many innate concepts there are) is a clear example of this type of overly narrow perspective. Also as noted in Chapter 2, we will have little interest in trying to locate positions in the space of possible views in terms of just how rationalist or empiricist they are. Occasionally coarse-grained comparisons will be appropriate—as, for example, between our own rationalist view and Fodor's radical concept nativism (something we do in

Part IV). But detailed comparisons of the degree to which different accounts are rationalist or empiricist won't play any role in what follows.

It's also important to bear in mind that the seven dimensions along which rationalist and empiricists views can vary allow for many types of trade-offs in which some of these lean more in a rationalist direction and others lean more in an empiricist direction. These sorts of trade-offs mean that the simple story about what makes a view fall under the overall framework of concept nativism or the overall framework of concept empiricism will require certain qualifications. This is true both for local debates concerning the origins of a particular concept or conceptual cluster and for the global debate concerning the origins of concepts in general.

The simple story about what it is to be a concept nativist or a concept empiricist in a local debate about the origins of a particular concept (or conceptual cluster) is as follows. To be a concept nativist in a local rationalism-empiricism debate of this sort is to adopt a rationalist account of the origins of the concept, that is, to take it to be either innate or acquired via a rationalist learning mechanism. And what it is to be a concept empiricist in such a local rationalism-empiricism debate is to adopt an empiricist account of its origins; that is, to take it to be acquired via an empiricist learning mechanism.[35] Likewise, the simple story about what it is to be a global concept nativist is to adopt an account of the origins of concepts in general that overall counts as rationalist. This typically involves taking there to be concepts in more than just a few conceptual domains for which there is a rationalist account of their origins. By contrast, to be an empiricist (or a concept empiricist) in this global debate is to adopt an account of the origins of concepts in general that overall counts as empiricist. This typically involves taking there to be few if any concepts for which there isn't an empiricist account of their origins.

Unsurprisingly, however, these simple stories involve certain simplifications, which a longer and more nuanced story would need to take account of. For example, for a local rationalism-empiricism debate about the origins of a particular concept, there will be clear cases where there is no doubt that a given account is rationalist or no doubt that it is empiricist. However, there will also be cases that are less clear, where the account at issue involves a learning mechanism with a balance of settings across the dimensions along which rationalist and empiricist views vary that put it close to the border separating rationalist and empiricist views. An example of this kind would be a learning mechanism that traces back to an innate domain-general mechanism together with domain-specific attentional biases and a number of functionally-specific (but not domain-specific)

[35] Alternatively, we could also say that a rationalist account of the origins of concepts in a local debate is one that traces back to a rationalist local acquisition base, while an empiricist account of the origins of concepts in a local debate is one that traces back to an empiricist local acquisition base.

psychological structures. In such a case, the learning mechanism certainly isn't a paradigmatically empiricist learning mechanism, but then it's also not clear that it should be considered a rationalist learning mechanism either, lying as close as it does to the border that divides these approaches.

The longer story for global rationalism-empiricism debates regarding the origins of concepts is that we have to move beyond the quantity and diversity of the content domains of concepts for which there is a rationalist acquisition account. These factors matter, but they aren't the only factors that contribute to making a view more rationalist or empiricist. And different types of trade-offs involving these factors will be consistent with a view being rationalist overall or empiricist overall. For example, a view might count as a form of global concept nativism in light of its positing characteristically rationalist psychological structures pertaining to a very large number of different domains even when these are relatively lacking in complexity, or relatively unarticulated, or not especially closely aligned with their target domains. At the same time, a view might also count as a form of global concept nativism in light of there being characteristically rationalist psychological structures pertaining to fewer domains but where these are relatively complex, richly articulated, or closely aligned with their target domains.

While the complexities associated with these longer stories are important for a complete understanding of these debates, we can largely ignore them going forward because our case for concept nativism isn't affected by them. The reason for this is that the accounts that we will be arguing for when engaging with particular local concept nativism debates will be unquestionably rationalist; they are nowhere near the borderline between rationalism and empiricism. And the arguments we give in support of these accounts will be arguments not just against all empiricist alternatives, but also against views that might be taken to be borderline positions.[36] Similarly, our view in the global rationalism-empiricism debate concerning the origins of concepts in general is most definitely a rationalist one. Our view is that there is an unquestionably rationalist account of the origins of *many concepts* across *many different conceptual domains*. Here too, our case against competing views won't just argue against the empiricist alternatives; it will also argue against any view that might be taken to be near to the borderline. Since these views will invariably be considerably *less rationalist* than the view we argue for, nothing really turns on whether these views are taken to be empiricist or borderline rationalist, as our arguments will apply against them equally either way.

In sum, concept nativism is about more than just innate concepts. It is about whether there is a rationalist account of the origins of concepts, where such an account may be one in which there are innate concepts but crucially may also—or may instead—be one in which there are concepts that are acquired via rationalist

[36] Later, in Chapter 26, we introduce the term *robustly rationalist* for accounts that are unquestionably rationalist and so definitely not borderline rationalist accounts (or even clearly but minimally rationalist accounts).

Box 8

Local Concept Nativism—a local concept nativist view is a rationalist view in a local debate about the origins of a particular concept or conceptual cluster. This involves adopting a rationalist account of the origins of these concepts; that is, taking these concepts either to be innate or to be acquired via rationalist learning mechanisms.

Local Concept Empiricism—a local concept empiricist view is an empiricist view in a local debate about the origins of a particular concept or conceptual cluster. This involves adopting an empiricist account of the origins of these concepts; that is, taking these concepts to be acquired via empiricist learning mechanisms.

Global Concept Nativism—a global concept nativist view is a rationalist view in the global debate concerning the origins of concepts in general, which typically involves taking there to be a rationalist account of the origins of concepts in more than just a few content domains (especially when these are diverse content domains).

Global Concept Empiricism—a global concept empiricist view is an empiricist view in the global debate concerning the origins of concepts in general, which typically involves taking there to be an empiricist account of the origins of all or nearly all concepts.

learning mechanisms. And a variety of factors can contribute to determining how rationalist or empiricist a view is for local or global debates about the origins of concepts, where there can be trade-offs between these factors, just as there are for rationalist and empiricist views more generally. Finally, our own version of concept nativism holds that *many concepts* across *many conceptual domains* are either innate or acquired via rationalist learning.

6.4 Conclusion

In this chapter, we have further developed our understanding of what innateness is, contrasting our view of innateness with some nearby views and responding to the charge that innateness is a confused notion that should be eliminated. We have also discussed what concepts are and considered different ways of drawing the conceptual/nonconceptual distinction. We showed that there are many

competing ideas about how to draw this distinction but that, in the end, the existence of many ways of drawing this distinction (and the existence of correspondingly many accounts of what makes something a concept) doesn't affect the debate over the status of concept nativism. This means that the question of what a concept is does not have to be settled in advance of building a case for concept nativism. In fact, we have also argued that from the point of view of constructing a robust argument for concept nativism, it is actually better to remain neutral about how the conceptual/nonconceptual distinction should be drawn in making such a case. Finally, we have argued that there is no barrier to proceeding in this way since concept nativism should not be understood as just being about innate concepts. Concept nativism is fully compatible even with there being no concepts at all that are innate. What matters for concept nativism is only whether there is a rationalist account of the origins of concepts, and such accounts do not depend on positing innate concepts.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. DOI: 10.1093/9780191925375.003.0006

7
Conclusion to Part I

Since the earliest days of cognitive science, contributions to the rationalism-empiricism debate have inspired some of the most creative and productive research in cognitive science, with far-reaching impact on virtually every facet of the study of the mind. At the same time, however, foundational questions about the debate have haunted it throughout its long history, during which it has been regularly dismissed as vacuous, incoherent, or pointless, and ridiculed as amounting to little more than a collection of obscure metaphors. If these negative reactions were appropriate, our project—a defence of concept nativism—could never get off the ground.

In Part I, we have aimed to undertake a long overdue systematic rethinking of the theoretical foundations of the debate. This has involved two complementary tasks. The first is spelling out precisely what we take the debate to be about and clarifying the key theoretical notions that play a role in it. The second, which is at least as important, is clarifying what the debate is *not* about—that is, sharply distinguishing our way of understanding the debate from a variety of tempting but mistaken alternatives that turn out to be intellectual dead ends. This second task we have set for ourselves is particularly important in light of the fact that these misconceived alternative ways of understanding the debate are so widely held, especially the view that the rationalism-empiricism debate is about nature versus nurture (or the relative contributions of genes versus the environment). Both critics of the debate and its proponents and participants frequently conceptualize it in these problematic ways, and often conflate several incompatible interpretations of the debate without realizing it. As we see it, the widespread scepticism regarding the value and coherence of the debate stems directly from such misunderstandings.

At the heart of our understanding of the rationalism-empiricism debate are the differing views that rationalists and empiricists have about what we have called the *acquisition base*. The acquisition base is the collection of psychological structures that are not themselves the product of learning or any other form of cognitive development that is mediated by psychological processes. It provides the ultimate, unlearned psychological basis for explaining how all learned psychological traits are acquired. Any theorist who takes there to be any psychological traits is thereby committed to there being an acquisition base of one sort or another, since these psychological structures must either themselves be unlearned (and so be part of the acquisition base) or be acquired on the basis of more

fundamental psychological structures which themselves are ultimately acquired on the basis of unlearned psychological structures, on pain of infinite regress. In Chapter 2, we clarified the essence of the rationalism-empiricism debate in terms of rationalists' and empiricists' competing visions of the acquisition base, where empiricists take it to be relatively sparse and rationalists, by contrast, maintain that it is quite rich.

This way of understanding the rationalism-empiricism debate is not new. But it has never been fully articulated. Systematically developing it involves clarifying existing theoretical notions and distinctions (and, in some cases, introducing new ones) to provide a detailed framework for understanding the diverse range of possible rationalist and empiricist theories and how they relate to one another. We do this, particularly in Chapter 2 (for the rationalism-empiricism debate in general) and in Chapter 6 (for the debate as it applies to the origins of concepts). We argue that rationalist and empiricist views can be seen as varying along seven dimensions. These dimensions have to do with the (1) quantity, (2) complexity, (3) degree of articulation, (4) diversity of content domains, (5) degree of abstractness, (6) degree of domain specificity, and (7) degree of alignment of the psychological structures postulated to be in the acquisition base. Trade-offs of different kinds among these different dimensions are possible, and so two views may turn out to be rationalist (or empiricist) to roughly the same extent despite opting for different ways of weighting these dimensions. Overall, however, a view will count as empiricist if the combination of weights set for these different dimensions is sufficiently in the direction of empiricism, and rationalist if the combination is sufficiently in the direction of rationalism. This framework provides a sound theoretical foundation for the rationalism-empiricism debate, and the theoretical terms and distinctions discussed in Part I will be employed throughout the remainder of the book.

This framework also helps address and avoid many of the most persistent misunderstandings of the debate, particularly the idea that the rationalism-empiricism debate should be identified with the nature-nurture debate, with the difference between rationalism and empiricism turning on the relative importance of genes versus the environment. Both rationalists and empiricists have mistakenly seen the debate in these terms (not infrequently conflating this understanding with other ways of understanding the debate, such as our own). However, our account of the rationalism-empiricism debate makes clear that it has nothing to do with the relative contributions of genes versus the environment to the origins of psychological traits—so it is simply a mistake to identify the rationalism-empiricism debate with the nature-nurture debate. It's a good thing that the rationalism-empiricism debate isn't the same as the nature-nurture debate, since the nature-nurture debate is fundamentally confused. As we argued in Chapter 3, it doesn't make sense to say that either genes or the environment is solely, or even primarily, responsible for the development of a given trait; all

development involves a complex interplay of genetic and environmental factors. But this has no bearing on the rationalism-empiricism debate, as it isn't something that rationalists and empiricists actually disagree about. What they do disagree about is the character of the acquisition base. Empiricists, for instance, generally suppose that the cognitive mechanisms that mediate language acquisition are general-purpose mechanisms that do some form of statistical analysis, where these same mechanisms also account for learning in other domains. By contrast, some rationalists suppose that language acquisition is substantially mediated by a language-specific learning mechanism that incorporates the principles of Universal Grammar. The very real differences between these two types of accounts don't just evaporate because of a shared acknowledgement that gene-environment interactions are essential to development. So, while many have argued in one way or another that the rationalism-empiricism debate should be abandoned because the nature-nurture debate is fundamentally misguided, once we see that it is wrong to identify the rationalism-empiricism debate with the nature-nurture debate, these arguments turn out to simply be beside the point.

Chapters 3–5 addressed a broad range of related objections which purport to undermine the coherence or utility of the rationalism-empiricism debate as a whole, or rationalism in particular. To mention just one of these, it is often thought that rationalist views must hold that innate psychological traits are fixed or unchangeable and immune to environmental influence. However, given our understanding of what the rationalism-empiricism debate is about, it becomes clear that such concerns are misplaced. Rationalism doesn't entail that psychological traits are fixed or unchangeable. Rationalism (like empiricism) is committed to some traits being part of the acquisition base. But being in the acquisition base doesn't mean that they are unchangeable and immune to environmental influence. Moreover, since rationalism claims only that these traits are the product of a particular type of acquisition process, rationalism in no way rules out the possibility that they might be subsequently supplemented, modified, or even overridden depending on the social and environmental context, the operation of competing psychological mechanisms, or the impact of further learning.

One of the most important themes throughout Part I has been the central role of learning in rationalist accounts of the origins of psychological traits. This theme played a major role in developing our account of how concept nativism should be understood. Contrary to a commonly held view, concept nativism isn't just about whether there are innate concepts. It is true that rationalist accounts of the origins of concepts may involve claims that the acquisition base contains particular innate concepts. But equally, such accounts may instead involve claims that the acquisition base contains what we have called *characteristically rationalist psychological structures* that contribute to *characteristically rationalist learning mechanisms* which explain the origins of particular concepts. In this way, the difference between concept empiricism and concept nativism—like empiricism and

rationalism more generally—isn't that one accepts learning and the other does not. It's about the character of the learning mechanisms that each envisions, with rationalists taking this learning to often depend on a rich and varied assortment of psychological structures in the acquisition base that go well beyond the highly limited acquisition base envisioned by empiricists. In the end, whether one is a rationalist or an empiricist, there is no way of getting around the fact that most concepts are learned. However, rationalists and empiricists have very different views about how concepts are learned, with rationalists seeing learning as not being limited to empiricist learning mechanisms but crucially as also involving characteristically rationalist learning mechanisms.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0007

PART II

SEVEN ARGUMENTS FOR CONCEPT NATIVISM

8
The Argument from Early Development (1)

Our aim in Part I was to comprehensively rethink the nature and foundations of the rationalism-empiricism debate. The framework for understanding this debate that we developed in detail in Part I provides the backdrop against which we will make our case for a rationalist approach to the origins of concepts—that is, for concept nativism. While many have charged that the rationalism-empiricism debate is riddled with confusion and that rationalist approaches in particular are not theoretically viable and can be discounted without engaging in a detailed evaluation of their merits relative to competing empiricist accounts, Part I shows that these negative appraisals are based on misunderstandings both of rationalism and of what is really at stake in the debate. Here in Part II we will now turn to our case for concept nativism.[1]

As we have noted, concept nativism, like rationalism in general, is not a single view, but rather a broad framework encompassing many possibilities. The view that we will be arguing for is a relatively strong version of concept nativism which holds that there is a rationalist account of the origins of *many concepts across many conceptual domains*. In line with our discussion in Chapters 2 and 6, we will refer to this as *our version of concept nativism*, or *our concept nativism*, when we want to contrast it with other versions of concept nativism. (For ease of exposition, however, we will often not call attention to this contrast, and instead simply refer to our arguments as supporting *concept nativism* or supporting *our case for concept nativism*, dropping the qualifier "our version of".) The argument that we develop in Part II is aimed at supporting this form of concept nativism, but in arguing for our own strong view, we will also thereby be arguing for the weaker claim that *some* version of concept nativism is true. Likewise, since the case that we present is aimed at supporting a strongly rationalist account, parts of our case can be used independently of the full case to support many forms of concept nativism that are not rationalist to the same extent as our account is, and our arguments would support many of these less strongly rationalist accounts even if significant portions of our case were unsuccessful.

[1] For readers not reading the chapters in order, there are a number of technical terms that were introduced and explained earlier that we will continue to rely on in Part II, including "acquisition base", "rationalist learning mechanism", "characteristically rationalist psychological structures", "articulation", and "alignment". Brief summaries of how we are using these and other terms may be found in Boxes 1–7 in Chapter 2 and in Box 8 in Chapter 6.

Overall our argument for concept nativism takes the form of a single overarching inference to the best explanation. But to appreciate the force of this argument, it helps to look at concept nativism from a number of different vantage points that accentuate different aspects of its explanatory power. With this goal in mind, we will distinguish and develop seven relatively distinct lines of argument, each of which makes a strong case for a rationalist account of the origins of a range of concepts or conceptual domains but which work together to form our single overarching argument for concept nativism. We call these *the argument from early development*, *the argument from animals*, *the argument from universality*, *the argument from initial representational access*, *the argument from neural wiring*, *the argument from prepared learning*, and *the argument from cognitive and behavioural quirks*. The bulk of Part II will be devoted to separate presentations of these seven arguments, as each argument raises its own complex and interesting set of issues and draws upon its own body of empirical evidence. But we want to emphasize that this way of dividing things up is largely for expository convenience. In most instances, the concepts or conceptual domains that are covered by any one of these arguments are covered by others too.

The case that we will be making for our version of concept nativism is, by design, neutral about many fine details that would distinguish between different specific accounts. We will make our case by showing how, in many different conceptual domains, the arguments we present support a rationalist account of the origins of concepts in that domain. As we highlighted in Chapter 6, concept nativism is not just about innate concepts. At least as important to rationalist accounts of the origins of concepts is the idea that many concepts are acquired via rationalist learning mechanisms (that is, learning mechanisms that trace back to characteristically rationalist psychological structures in the acquisition base). In arguing for there being a rationalist account of the origins of concepts in a given domain, we are arguing for the disjunctive claim that concepts in this domain are either innate or acquired via a rationalist learning mechanism. The case for a rationalist account of the origins of concepts in a given domain will often be consistent with both of these two possibilities. Likewise, the case we make is neutral about many of the fine details regarding the form that particular learning mechanisms might have. There will be a large variety of accounts that are consistent with our argument for concept nativism that are worthy of further exploration; it's not our aim here to attempt to adjudicate among these.[2]

While encompassing a broad range of related views, our concept nativism stands in contrast with all empiricist accounts (which hold that concept acquisition is driven by domain-general mechanisms) (discussed in Part III) and with

[2] These accounts will vary along the dimensions that we highlighted in Chapter 2 and Chapter 6 in terms of quantity, complexity, articulation, diversity, abstractness, degree of domain specificity, and alignment.

Fodor's radical concept nativism (which holds that virtually all lexical concepts are innate) (discussed in Part IV). But importantly, it also stands in contrast with rationalist theories of the conceptual system that posit a more austere form of rationalism taking a rationalist account of development to apply to only a small number of conceptual domains. For example, the *core knowledge hypothesis* defended by Elizabeth Spelke, Susan Carey, and their colleagues has at times been held to apply to only a very restricted range of concepts:

> we believe that humans are endowed with a small number of separable systems that stand at the foundation of all our beliefs and values. New, flexible skills, concepts, and systems of knowledge build on these core foundations. More specifically, research provides evidence for four core systems (Spelke 2003) and hints of a fifth one. (Kinzler and Spelke 2007, p. 257)

We don't know exactly how many content domains there are where the origins of concepts in that domain is explained via characteristically rationalist learning mechanisms (or where concepts in the domain are themselves innate), but we are confident that it is considerably more than four or five.[3]

One of our major goals in Part II is to spell out the above seven arguments for concept nativism, to distinguish them from one another, and to clarify the logic behind each of them. For these purposes, the examples that we use to illustrate the arguments are incidental. But since we also hold that a rationalist approach is correct for a broad range of conceptual domains, a second major goal of Part II is to show in detail how these arguments support our concept nativism across a broad range of domains. This means that our discussion needs to have a certain amount of breadth—we have to discuss concepts drawn from quite a few conceptual domains. At the same time, we want to show how deep the case for concept nativism goes in that, in most instances, a rationalist account of a given conceptual domain doesn't rest solely on one type of argument or consideration. Rather, multiple lines of argument and a diverse body of evidence often converge in support of the rationalist viewpoint. To satisfy all of these demands, we will examine some of the same conceptual domains in different chapters in Part II, illustrating how different arguments for concept nativism complement one another and lead to the same rationalist conclusion, but we will sometimes have to confine ourselves to examining certain conceptual domains in the context of just one of the seven arguments. Moreover, for each conceptual domain that we discuss, there is

[3] Individuating systems or mechanisms isn't always easy, especially when one considers that different systems may be viewed as subsystems or parts of a larger and more comprehensive cognitive system. Nonetheless, we will present evidence throughout Part II for a considerable amount of differentiation in the acquisition base, including much differentiation that can't reasonably be accommodated taking these different systems to be parts of a handful of more comprehensive systems.

a great deal of potentially relevant evidence to consider, as well as competing interpretations of the data. To keep things manageable, we have necessarily had to be selective, sometimes considering some conflicting findings and opposing arguments, other times just presenting our positive case.[4]

In proceeding in this way, we recognize that our discussion cannot be exhaustive—not for any of the concepts or conceptual domains we touch on, and certainly not for all of them. It is our claim, however, that the balance we strike between breadth and depth goes a long way towards showing that concept nativism is the right account for understanding the origins of many concepts. And as we will see, the different lines of argument that are considered independently in much of Part II work together in a complementary fashion to greatly strengthen the overall case for a rationalist account of conceptual development. When the seven arguments for concept nativism are viewed collectively—taking into account the range of concepts they span, the wealth of evidence they illuminate, and the way that they interact and support one another—the result is a very strong argument for concept nativism.

The topic of this and the next chapter is the first of our seven arguments—the *argument from early development*—which is often taken to be the quintessential argument for concept nativism. The general logic of this argument is that certain concepts and representational abilities appear too early in life to reliably be acquired solely on the basis of empiricist learning mechanisms. Thus, the best account of their origins is a rationalist one, which takes them to either be innate (i.e., part of the acquisition base) or else be acquired by rationalist learning mechanisms. In these first two chapters of Part II, we will fill out this argument by clarifying what is meant by *too early*, by examining some representative examples of early development, and by responding to prominent empiricist concerns about drawing rationalist conclusions from these sorts of examples. This chapter will focus on the argument from early development in cases where there is evidence for representational abilities being present at birth. Chapter 9 will look at how the argument works when there is no evidence that a representational ability is present at birth.

When a representational capacity is present at birth, the capacity can be present in advance of *any* perceptual access to relevant environmental information. Such cases provide a particularly strong version of the argument from early development and can provide highly compelling evidence for a rationalist account. At the same time, however, there are important challenges to arguments that turn on whether a trait is present at (or soon after) birth.

[4] Also, although the core of our positive case for concept nativism is given in Part II, it is worth noting that the discussion of empiricist views in Part III expands on many of the seven arguments in Part II by extending them to new conceptual domains.

The most pressing of these, of course, is the difficulty of establishing that the trait *is* present this early in life. Given that newborns cannot speak, point to things, move around on their own, or display most of the behaviours that are used to establish representational capacities in older children or adults, it can be exceptionally difficult to show that newborns possess a given representational capacity this early.[5] There is also a question about when the inference from *present at birth* to *innate* is warranted. This inference is complicated by the fact that learning can take place in the womb. For example, it has been shown that newborns have a preference for the native language of the linguistic community that they are born into. Newborns of English-speaking parents prefer English, while newborns of Spanish-speaking parents prefer Spanish (Moon et al. 1993). Yet no one thinks that these preferences are themselves innate, just as no one thinks that the specific language that anyone speaks (French, Spanish, English, etc.) is innate. The only real alternative here is that these preferences are somehow learned prior to birth.

How is it possible for researchers to determine that newborn infants *do* have these preferences or that they are acquired prenatally? First, recordings have been made inside the womb that demonstrate that the general prosodic features of language are audible there, and that 30%–40% of phonetic content is even discriminable (Moon et al. 2013). This shows that the information required for learning is at least present in the foetal environment.

Second, researchers have devised a number of ingenious ways to find evidence for newborn's auditory preferences. One method that has proven especially useful gives infants a way to control what they hear by making it contingent on a modest behaviour that they are capable of producing—sucking. Experimenters arrange things so that the duration of an infant's sucking (on a non-nutritive computer-monitored nipple) determines the duration that an auditory stimulus is emitted. In this way, infants' preference for hearing one language rather than another can be measured by comparing the duration of sucking when they hear a familiar language (one they were exposed to in the womb) with the duration of sucking when they hear an unfamiliar language. Moon et al. found that newborns suck for longer periods to hear the familiar language. This shows both that infants are capable of discriminating between the two languages (since they systematically exhibit different behaviours in response to them) and that they prefer their native language (as they are willing to work harder to hear speech in their native

[5] We saw in Chapter 2 that some theorists understand representational capacities to involve nothing more than a disposition to develop a representational competence (i.e., an infant's possessing a capacity to represent X only requires the infant to have the ability to develop the representation of X later in life). That is *not* what we are claiming. Rather, when we say that infants have a representational capacity, we are referring to a *current* representational competence. For example, when we say that newborns have the capacity to represent numerical quantity (see below), we are claiming that infants already possess the competence to represent numerical quantity at birth.

language than to hear speech in an unfamiliar language).[6] Given that this behaviour is apparent so soon after birth and that it is sensitive to information that is available in the womb, it only stands to reason that this behaviour is shaped by prenatal learning. As surprising as it may be, then, representational capacities and specific preferences that aren't innate may be present at birth all the same.

Examples along these lines show that we have to be careful when we turn to evidence that a representational capacity is present at birth. Still, there are many instances in which we can be fairly confident that the relevant environmental information for a representational capacity is not accessible in the womb, thereby ruling out foetal learning. In such cases, if the capacity is present at birth, it must either be innate or else be learned in a way that requires only the barest exposure to the environment—that is, via the kind of learning supported by rationalist learning mechanisms.

We have seen that a certain amount of auditory information about the external environment is demonstrably available in the womb. Notice, however, that comparable *visual information* about the external environment is not available in the womb. So if we are looking for cases where the possibility of foetal learning can reasonably be excluded, representational capacities that are grounded in vision are a natural place to turn, particularly cases where newborns can be shown to categorize specific types of visual stimuli beyond the most elementary sensory properties. We will discuss two examples that illustrate this form of the argument from early development, the representation of faces and number.

Let's start with faces. Newborn infants have been shown to possess a number of psychological abilities related to representing, categorizing, and remembering faces. For example, when given a choice, newborns preferentially look at schematic face-like stimuli over equally complex non-face stimuli (Goren et al. 1975). The average age of the infants in this experiment was just *9 minutes old*—with some as young as *3* minutes old. The stimuli were either a schematic face or a scrambled face (with the eyes, nose, etc., in the wrong places) on a head-shaped cut-out. In both cases, the stimulus slowly moved in such a way that the babies had to turn their eyes and head to follow its motion. The amount of effort they were willing to expend to continue to see the stimulus (by turning in this way) could then be used to measure their visual preference. The crucial finding was that these newborns exerted more effort to follow the schematic face, suggesting that infants have an innate interest in faces or face-like stimuli.

[6] Another method that has been used to show this type of sensitivity in newborn infants works around their response limitations by placing infants on a skateboard that supports the weight of their head and cradles their body in a way that facilitates leg and arm movement. Using this apparatus, researchers compared the responses of French newborns to spoken French and English sentences. The French newborns rotated their head and trunk to orient towards a loudspeaker when it produced French sentences, but not when it produced English sentences. The infants also initiated significantly more crawling-type motions in response to French as opposed to English sentences (Hym et al. 2022).

This is not the only face-related ability that newborn infants possess, however. In spite of their poor vision and their limited attention span,[7] they are also able to recognize and remember individual faces from a variety of different orientations (Turati et al. 2008) and can do so exclusively on the basis of inner facial features, as opposed to features such as hairline contours (Turati et al. 2006; Leo et al. 2018). Moreover, newborns are subject to an important face-specific visual illusion—the *Thatcher illusion*—which strongly suggests that they process faces in the same distinctively holistic manner as adults (Leo and Simion 2009). Each of these findings offers valuable insights into the way that infants represent faces, but just to illustrate how this type of research works, we will focus on the Thatcher illusion.[8]

In this illusion, changes to the orientation of the mouth and eyes look dramatically different depending on whether the face is presented right-side up or upside down. This can be seen by viewing two otherwise identical photographs of a face side by side where the eyes and mouth in one photo are rotated 180° relative to the rest of the photo. When the two photos are upside down, the photos have a broadly similar appearance, even though they look marginally different from one another (see Figure 8.1). But when they are right-side up, the difference between

Figure 8.1 The Thatcher illusion. The two images here look broadly similar when viewed in this upside down orientation, but if you turn this page over so that the images are viewed right-side up, they look dramatically different from one another. (Photo courtesy of Chris Collins of the Margaret Thatcher Foundation, adapted by the authors to illustrate the illusion. https://commons.wikimedia.org/wiki/File:Margaret_Thatcher.png.)

[7] For example, newborns' visual acuity on some measures has been found to range between 20/200 and 20/1200 (Held 1979).

[8] The illusion derives its name from the fact that the researcher who discovered the illusion used images of Margaret Thatcher to illustrate it (as we have here) (Thompson 1980).

them is remarkably vivid—the face with the inverted features looks grotesque (as can be seen by turning the page upside down and viewing Figure 8.1 again). This is because of the holistic character of face perception, which is adapted to the right-side up orientation. When a face is upside down, the normal face-specific processes are disrupted and other, general-purpose visual processes are recruited, ones that aren't sensitive to the nuanced relations among the elements that make up a face.

Now it is easy enough to determine that an adult sees this face illusion. You can just ask. But obviously you can't ask 1-day-old infants what they see. To get around this difficulty, Leo and Simion used what is known as a *habituation procedure*. The basic idea behind this procedure is that infants are presented with a succession of different stimuli which are all of the same type. This continues until their attention to the stimuli diminishes by a predetermined measure. (This drop in attention can be ascertained in a number of ways. With visual stimuli, the crucial measure is often a decrease in the amount of time the infants spend looking at the stimuli.) At this point, an infant is said to have *habituated*. Next comes the test phase, where a different type of stimulus is presented to the infant. If infants' attention increases significantly more to the novel type, this shows that they discriminate between these two types of stimuli—since they had lost interest in the first type of stimulus, the renewed interest shows that they see the new stimulus as being of a different type.[9]

Leo and Simion adapted this procedure by habituating newborns to a photograph of either a regular face or a "thatcherized" version of this same face (with inverted features), and then testing them with both faces presented simultaneously. In one condition, the faces were oriented right-side up, while in another they were oriented upside down. The result was that newborns looked longer at the novel face but only when the faces were right-side up. This indicates that when the faces were upside-down, infants didn't see much of a difference between the original photo and the thatcherized version, but when the faces were right-side up, the difference was evident. Again, this is just like the pattern found with adults, suggesting that newborns process faces in broadly the same distinctive holistic manner as adults.

What does this work on newborn infants' processing of faces show? In the first instance, we take it to show that infants possess various face-related preferences

[9] We saw an example of this method with older infants in the discussion of colour discrimination in Chapter 5. Another classic example of this method in the auditory domain showed that 1-month-old infants discriminate the basic sounds used in their native languages (Eimas et al. 1971). Infants heard speech sounds that varied in their acoustic properties yet continued to instantiate the same phoneme (*ba, ba, ba…*). After reaching habituation, they heard a different phoneme (*pa, pa, pa*), or else further instances of the same old phoneme (*ba, ba, ba*). Recovery of interest was measured in terms of sucking on a non-nutritive nipple, which controlled the presentation of the stimuli (as in Moon et al. 1993). Infants showed a significantly greater recovery of interest for the phoneme change even when the contrasts involved acoustical changes that were no greater than the acoustical changes within a given phonemic category in the habituation phase.

and abilities that are present too early in development to be acquired via domain-general learning. Thus, in addition to whatever domain-general learning mechanisms may be part of the acquisition base, it also includes characteristically rationalist psychological structures pertaining to the domain of faces (and, as we will see in a moment, to some other closely connected domains).

Consider, for example, a type of empiricist account briefly mentioned in Chapter 2. On this type of account, there aren't any innate face-specific psychological structures in the acquisition base. Face-specific processing is taken to be a product of domain-general learning that is supplemented by a modest innate bias to attend to a low-level perceptual property (such as curvilinearity) that loosely correlates with the presence of a face. Empiricist approaches of this sort can't account for the face-specific processing of face stimuli that newborns already exhibit, as seen in the Thatcher illusion. On an empiricist account, face-specific processing only arises after extensive experience with faces—in this case, experience in which disproportionate attention is given to faces owing to associated low-level perceptual properties. But since the data we have reviewed is concerned with *newborns*, this precludes the possibility of this type of face-specific processing being due to domain-general learning based on extensive experience of faces. In this way, establishing distinctive processing in newborns for a given domain makes for a particularly compelling argument against such empiricist domain-general learning accounts.

An important question for future research regarding infants' processing of face stimuli is to determine precisely what characteristically rationalist psychological structures in the acquisition base account for these and related findings. For our purposes in this book, what matters is the overall type of account that is needed—the fact that it is a rationalist account of one kind or another—and not the computational details of a fully worked out theory of face-related abilities. As we noted earlier, in general, our goal throughout Part II isn't to argue about the fine-grained structure of the acquisition base. The space of theoretical possibilities is too large, and the evidence that is currently available is not sufficient to settle questions at this level of detail. Instead, our goal is to engage with the broader dialectic between rationalist and empiricist approaches to conceptual development and to make the case for rationalist approaches by clearly distinguishing our seven arguments for concept nativism, clarifying the logic of each of the arguments, and identifying examples of content domains where they are likely to apply. With this goal in mind, we will often claim that the evidence supports a rationalist treatment of the origins of concepts in a given conceptual domain without taking a stand on the precise form the rationalist theory should take in terms of the specific elements it takes to be in the acquisition base, or without committing ourselves to a particular rationalist theory among the competing rationalist options.

Nonetheless, it is worth noting that in the case at hand, the studies we have reviewed so far speak to some of the controversies that would need to be settled

by a fully worked out rationalist theory. For example, there is a question about whether there are innate face-specific *mechanisms*, innate face-specific *representations*, or both. Notice, though, that the fact that newborn infants can represent novel faces and are subject to the Thatcher illusion is reason to suppose there is an innate special-purpose mechanism capable of productively generating new face representations, one that processes these representations in much the same holistic manner found in adult face perception.[10] And the fact that newborns prefer non-scrambled faces to scrambled faces indicates the existence of an innate face schema, a type of representation. So it looks like the acquisition base contains at least one face-specific mechanism and at least one face-specific representation. This helps to clarify the character of the elements of the acquisition base bearing on the origins of psychological structures in this domain to some extent, though of course it still leaves a great many questions of detail to be addressed.

There are also interesting questions about the content of what infants represent as they respond to faces, for example, the content of what we have called a *preference for faces*. Is this the right way to describe the way that infants represent faces and face-like stimuli? Given that the representations in question are prelinguistic, one might wonder how faithful any description can be that employs familiar linguistic terms and phrases (like the English term "face") for what is going on inside an infants' mind. What's more, there are bound to be a number of kindred rationalist explanations, invoking different candidate rationalist mechanisms and/or innate representations, that explain the evidence nearly as well as one another and that are difficult to tease apart experimentally. In that case, one might want to be cautious about describing infants' abilities in a manner that prejudges the way they represent stimuli pertaining to faces.[11]

We take these questions and concerns about representational content to raise some interesting issues. However, they shouldn't be construed as objections to this sort of infancy research or to rationalist claims based upon it, since one can have good reason to accept a rationalist account for a given domain before all of the details regarding the rationalist account have been settled. Rather, they should be seen as ongoing research questions in a productive research programme. Further work—in some cases *much* further work—will be required to get to the point where we can say which particular rationalist account turns out to be the right one, and which particular set of characteristically rationalist psychological structures it takes to be in the acquisition base.

[10] The representations of particular faces that this innate mechanism produces are good examples of representations that are learned via a rationalist learning mechanism. They are clearly learned in response to environmental stimuli, but the learning is mediated by an innate mechanism that is specialized for acquiring representations in this particular domain.

[11] These sorts of issues about the content that can be ascribed to infants are by no means unique to face perception—or to rationalist approaches. Similar questions come up for every domain that the argument from early development touches on, and for rationalist and empiricist approaches alike.

Also, although we will often set such questions of detail to one side, it is important to see that research that aims to clarify the exact content of what infants represent can be illuminating and is being pursued in accordance with rationalist approaches to all of the domains where the argument from early development has gained traction. For example, in the domain of faces, we have pointed to evidence that there are innate mechanisms that are specific to faces. But this leaves open the question of precisely which domain they are specific to. Are they innately geared towards the representation of human faces, or do they function equally well for faces of other species (at least initially)? This question has been investigated, and the evidence suggests that face perception is not restricted to human faces at birth. It turns out that not only do newborns distinguish one human face from another. They also distinguish one monkey face from another. Moreover, they have a preference for upright monkey faces compared to inverted monkey faces, and they have no preference between monkey and human faces (Di Giorgio et al. 2012). So at a minimum, the innate face-specific mechanism involved here would seem to be broad enough to cover primate faces generally.

The fact that infants start out with this more general yet still domain-specific capacity—one for representing a broader class of faces than just human faces but which is nonetheless restricted to faces—brings us back to a point that we were at pains to emphasize in earlier chapters. This is that concept nativism is perfectly compatible with learning and development taking place.[12] Rationalists are in no way committed to the view that every aspect of an adult competence—for faces or any other domain—is innate. What makes an account of the origins of any given concept a *rationalist* account is the fact that it is grounded in a rationalist acquisition base, which contains characteristically rationalist psychological structures. Such an account is fully compatible with a great deal of conceptual development and conceptual change taking place in the course of development (see Chapter 2).

In discussing what this research shows, we have so far focused on how it bears on questions about the contents of the acquisition base. But we also need to consider the question of which concepts it argues would be best explained in terms of a rationalist account. Before we turn to this question, however, it is worth noting that infants' face preferences and face recognition abilities are not isolated, but rather are part of a larger collection of related abilities and preferences geared towards social interaction and communication. Newborn infants don't just prefer faces to non-faces. They prefer faces that directly gaze at them (Farroni et al. 2002), faces that have previously spoken to them (Coulon et al. 2011), and faces of individuals who have communicated with them in a non-linguistic manner (Cecchini et al. 2011). These preferences all exhibit further representational abilities beyond simply the classification of faces as such or the classification of

[12] In fact, face memory and recognition don't reach adult-level maturity for many years (Weigelt et al. 2014).

specific faces. They suggest that the innate resources in the acquisition base exhibit considerable articulation (in the sense elaborated in Chapter 2) in the form of innate links with innate social and communicative capacities.

As with the studies we began with, further research is needed to determine how best to characterize the interrelated capacities at issue in this research, but the work thus far suggests that, in representing faces, newborns construct intermodal representations of individual people. Findings with slightly older infants supports this broader picture. For example, gaze following in 6-month-old infants is contingent on communicative cues such as prior direct eye gaze or infant-directed speech (Senju and Csibra 2008). Five-month-old infants treat melodies as social cues. They selectively respond to a new individual who reproduces a familiar melody if the melody was previously sung to them by a parent, but not if it was previously heard as simply coming from the environment (i.e., with no social engagement) (Mehr et al. 2016). Seven-month-old infants expect members of a social group to behave similarly, but fail to have this expectation of things that are not social agents (Powell and Spelke 2013). And infants as young as 3 months old discriminate between agents and non-agents, using a variety of different types of cues to discriminate agents (Johnson 2005; Luo and Baillargeon 2005; Luo 2011).

In sum, there is a great deal of evidence that infants at birth possess a schematic face representation, have the ability to productively form and retain face representations for novel individuals, and process faces in the same distinctive (holistic) manner as adults. Moreover, there is also considerable evidence indicating that these abilities are innately linked to a range of social and communicative functions.[13]

Let's turn now to the question of which conceptual domains are likely to involve either innate concepts or concepts acquired via rationalist learning mechanisms in light of these findings, and which concepts within these domains look to be promising candidates for a rationalist account. In the first instance, this work bears on the domains of faces, sociality, and individuals. Some of the candidate concepts that it suggests are either innate or acquired via rationalist learning mechanisms are concepts like FACE, AGENT, INDIVIDUAL HUMAN/PERSON, SOCIAL INTERACTION, EYES, GAZE, ATTENTION, and COMMUNICATION.[14] In some cases, there is reason to suppose that the concept in question is not innate but is acquired via a rationalist learning mechanism instead—such as the concept HUMAN FACE, which, as we noted, is a product of early development. The critical

[13] We will return to infants' social and communicative knowledge and abilities in the next chapter.

[14] In many of the cases we have discussed, it is not clear what precisely the infants are representing—for example, whether they represent particular humans as individual humans/people, as agents, as both individual people and agents (or if the content of their representation(s) differs systematically from these concepts). As we noted earlier, this is to be expected. It is also worth noting that the innate concepts that infants have in a given content domain will likely have substantially less rich conceptual roles than those of adults, and may not be the same concepts as adults have; innate concepts may well undergo substantial conceptual change during development.

point in this and similar cases is that while this development means that the concept is not innate, this development is grounded in a prior capacity for responding to faces that is rooted in domain-specific elements in the acquisition base. This means that a rationalist account of the origins of HUMAN FACE is appropriate even though this concept is learned.

In addition to these general concepts, we also need to consider the origins of representations of particular entities, such as a specific person, a specific face, or a specific voice. The work we have highlighted above suggests that in each case the origins of these kinds of representations are best explained in terms of rationalist learning mechanisms that are able to productively generate representations of numerous particular instances of the general category in question. Some of these representations—notably those of specific people—are clearly conceptual representations in adults. And even if the precise content of these representations in infants differs from those of the concepts that adults possess, it is overwhelmingly likely that the rationalist learning mechanisms that account for infants' acquisition of representations of particular people will also play an important role in the acquisition of concepts of particular people later in life.

The representation of specific human faces and specific human voices raise interesting questions because they are seen as being conceptual on some theories of concepts and nonconceptual on others. Perhaps it shouldn't be all that surprising that some of the representations that can be seen to be present at birth—based on studies with newborn infants—end up being near the conceptual/nonconceptual border. After all, the obvious difficulties of conducting experimental studies that newborn babies could participate in means that such studies are likely to employ relatively simple stimuli, making it more likely that the representations at issue will be relatively simple and concrete, and so, in at least some cases, closer to the conceptual/nonconceptual border. So it's worth pausing for a moment to discuss how the fact that these representations come out as conceptual on some accounts of what concepts are and nonconceptual on others affects the issues regarding concept nativism. We will focus our discussion of this issue on infants' representations of particular faces (Alice's face, Bill's face, and so on) and the fact that such representations might be seen as being conceptual by some theorists and nonconceptual by others.

One type of theorist who might see representations of individual faces as nonconceptual is a theorist who draws the conceptual/nonconceptual distinction in terms of the Generality Constraint. Such a theorist might argue that these representations aren't sufficiently integrated with the rest of the conceptual system to be considered concepts. Another type of theorist who might see representations of individual faces as nonconceptual is one who draws the distinction in terms of whether a representation is iconic or discursive. Such a theorist might argue that face representations are not concepts since they are iconic and more like pictures than sentences.

These grounds for thinking that representations of individual faces are nonconceptual are debatable, however. Regarding iconicity, while the representations of particular faces may not seem especially sentence-like, it's not true that they have no privileged decompositional structure—faces are represented as having components (eyes, nose, mouth) that stand in particular relations to one another. So face representations don't satisfy Fodor's picture principle for iconicity (see Chapter 6). And regarding the Generality Constraint, while it is true that one cannot combine a representation of a particular face with every imaginable conceptual predicate (e.g., the representation ALICE'S FACE IS DIVISIBLE-BY-7 makes no sense), much the same is true of clearly conceptual representations (THAT CHAIR IS DIVISIBLE-BY-7 makes no sense either). Moreover, for predicates which *are* applicable to the category *face*, it seems that representations of particular faces can be freely combined with such representations. Employing a representation of a particular face, we can judge that face to be long and narrow, smiling, attractive, or tired looking. And, of course, such representations figure in inferences too, for example, inferences regarding the identity of a particular person. Finally, it should also be kept in mind that like the general concept of a face, representations that are involved in the recognition of a given individual's face are highly abstract. A given individual's face can be recognized from different angles, different distances, in different lighting conditions, and across a wide variety of dynamic changes—opening the mouth, blinking, smiling, frowning, changes in eye direction, and so on. Taken together, we think that these points make a reasonable case for considering such representations to be concepts, even if they are not prototypical of the sorts of representations that figure in standard general accounts of what concepts are.

However, our concern here is not to settle the question of whether representations of particular faces are or are not concepts. Rather, we will consider the implications for concept nativism both when such representations are taken to be concepts and when they aren't. If individual face representations are concepts, then this means that the set of concepts and conceptual domains that a rationalist account is appropriate for includes not only general concepts like FACE, AGENT, INDIVIDUAL HUMAN/PERSON, SOCIAL INTERACTION, EYES, GAZE, ATTENTION, and COMMUNICATION but also concepts for particular faces. Of course, these particular face concepts wouldn't be innate. But they would still be explained in rationalist terms as being acquired via rationalist learning mechanisms.

What are the implications for concept nativism if individual face representations are nonconceptual? In this case, they wouldn't be on the list of concepts acquired via rationalist learning mechanisms. But importantly, they would almost certainly contribute to other representations (which uncontroversially are concepts) being on this list of concepts acquired via rationalist learning mechanisms. For example, they would almost certainly be involved in the acquisition of the

concept HUMAN FACE. They would also typically be involved in the acquisition of the concepts of particular people. We take this example to illustrate a general point. Even if it is unclear whether a given set of representations whose origins are best explained in rationalist terms is conceptual or nonconceptual, such representations (and the characteristically rationalist psychological structures they tell us are part of the acquisition base) can contribute to the acquisition of other related representations, many of which will be clearly conceptual. So, regardless of whether the representations involved are taken to be conceptual or nonconceptual, they will be contributing to a rationalist account of the origins of concepts.

These considerations also serve to highlight a fact that was emphasized in Chapter 2—that we shouldn't be exclusively focused on which concepts are present at birth. In the end, what matters is how any given concept is acquired, not when it appears. For example, whether or not the general concept INDIVIDUAL HUMAN/PERSON or the general concept AGENT is present at birth, it is clear that adults have both of these concepts. And, if these concepts are learned and not present at birth, then the characteristically rationalist psychological structures in the acquisition base that the research we have been reviewing shows to be associated with them would undoubtedly make important contributions to the acquisition of these concepts. Accordingly, a rationalist account of both the general concept INDIVIDUAL HUMAN/PERSON and the general concept AGENT will be appropriate even if newborn infants lack one or the other or even both of these concepts.

This shows that the argument from early development shouldn't simply be seen as an argument about *concepts appearing too early* to be acquired on the basis of empiricist learning mechanisms. Sometimes the argument plays out this way. But importantly, sometimes it plays out in terms of the critical rationalist psychological structures that support the acquisition of a concept—where it's these psychological structures that appear too early in life to be acquired on the basis of empiricist learning mechanisms. As a result, the best way to see the argument from early development is not simply as an argument about concepts appearing too early in development to reliably be acquired on the basis of empiricist learning mechanisms, but rather as an argument about *concepts or the characteristically rationalist basis for acquiring concepts* appearing too early to reliably be acquired on such a basis.

Let's now briefly consider a second example where the focus on newborns has been fruitful, namely, the representation of numerical quantity. It may seem extraordinary to attribute representations of such an abstract thing as numerical quantity to infants, much less to newborn babies. Before we get to the evidence regarding newborns, it is important to address this issue. Since number is typically correlated with a host of more concrete, non-numerical properties, why think that infants are responding to number rather than these

other properties? For example, three oranges are numerically greater than two, but three oranges will also typically have a greater volume than two oranges, and a greater surface area, greater total perimeter, greater luminance, and so on. Likewise, as we noted in Chapter 2, number can correlate with duration, as when a sequence of ten taps on a keyboard takes longer than a sequence of four. In the literature on numerical representation, properties like these, which typically correlate with numerical quantity, are referred to as *continuous magnitudes*.[15] So the question is, why should we think that infants are responding to numerical quantity as such rather than to some type of continuous magnitude?

In hindsight, it is clear that many of the early and widely cited studies of infants' numerical abilities didn't impose sufficient experimental controls to disentangle numerical quantity from continuous extent. For example, Antell and Keating (1983) claimed to show that infants can discriminate 2 vs. 3 dots. But because the dots were all of the same size, the 3s systematically contained more total surface area, total perimeter, etc., than the 2s. More recently, however, researchers have worked hard to eliminate such confounds, often with great ingenuity.

To see how this is possible, let's start with an important study that helped to establish that 6-month-olds can discriminate larger numerical quantities, that is, numerical quantities greater than 3 (we will come to studies with newborns shortly). Xu and Spelke (2000) tested for whether infants at this age can distinguish between 8 vs. 16 solid black dots, using a habituation paradigm, much like that used by Leo and Simion above. Infants were presented with a succession of different displays which each contained the same number of dots (8 for some infants, 16 for others), with the sizes and positions of the dots varying across displays. Once the infants habituated, they were then tested on both numbers (8 and 16) (see Figure 8.2). As before, if infants look longer at the new stimulus—in this case, the novel number of dots—this indicates they can discriminate it from the habituating stimulus. In fact, this is what Xu and Spelke found. Six-month-olds looked longer at 8 dots if they had habituated to 16, and looked longer at 16 dots if they had habituated to 8.

But again, it is not enough that infants can discriminate between stimuli with different numbers of dots. To show that they are discriminating between them in terms of their *numerical* quantity, we need to be sure that there aren't other features of the stimuli that systematically correlate with their numerical quantity that the infants could be using to discriminate between the stimuli.

Xu and Spelke's procedure dealt with this problem by equating certain continuous magnitudes associated with the 8s and 16s in the habituation trials and

[15] Other terms used for these non-numerical properties include *continuous stimulus properties* and *continuous extent*.

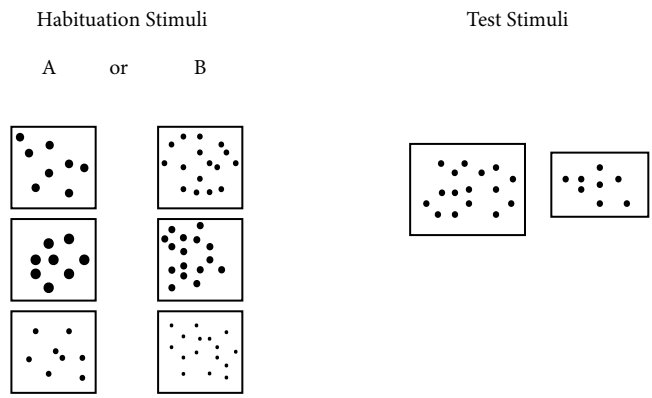

Figure 8.2 Sample stimuli from Xu and Spelke (2000). 6-month-olds were habituated to either 8 or 16 dots, and then shown both 8 and 16 to see if they would dishabituate more to the novel numerical quantity. To ensure that the infants weren't responding to changes in the continuous (non-numerical) features of the stimuli, the stimuli in the habituation trials were equated for area, and the stimuli in the test trials were equated for element size and density. (Figure based on Xu and Spelke 2000, figure 1.)

equating others in the test trials. For example, to make sure that infants couldn't simply respond to changes in the cumulative area of the stimuli—one of the major confounds in earlier work on number—the size of the dots were varied in such a way that the cumulative area of the 8s and 16s were the same across the habituation trials. This precaution led to the individual elements composing the 16s to be, on average, half the size of the individual elements composing the 8s. And because the displays were the same size, it also led to the 16s being more crowded (i.e., having less space between the dots). So these factors were equated in the test stimuli, which used dots that were always the same size and had the same element density regardless of number. Given this arrangement, infants who had habituated to 8 (or 16) dots couldn't dishabituate to the other number by attending to the position of the dots, their cumulative area, differences in brightness (which correlate with area), dot size, or element density.

Despite these meticulous precautions, there is still one non-numerical property that wasn't equated in either the habituation trials or the test trials, namely the total perimeter of all the elements (i.e., the sum of perimeters of all the dots). Notice that you can't equate the total perimeter if you equate the total amount of area covered. (The area of a circle is πr^2; the circumference is $2\pi r$.) But further experiments incorporated another control condition, equating the total perimeter in the habituation stimuli, and infants were found to dishabituate to the same numerical contrasts as before (Xu 2003; Xu et al. 2005). So, surprising as it may be, it looks as though 6-month-olds are able to discriminate 8 from 16 specifically in numerical terms.

An interesting feature of this numerical-representational capacity—one that was mentioned briefly in Chapter 2 and that we will return to in Chapter 10—is that it isn't as precise as the conventional counting list that is learned later in life ("one, two, three, four, etc."). Instead, it is grounded in a distinct system for representing numerical quantity—the *approximate number system*—which is ratio dependent in the sense that what matters to whether it can discriminate between two numerical quantities isn't their absolute numerical difference but the ratio between them.[16] For example, we have seen that 6-month-olds can discriminate between 8 and 16 dots (a 1:2 ratio). They are also capable of discriminating between 6 and 12 and between 12 and 24. However, it turns out that they fail to discriminate 8 from 12. One might have thought that this failure is just because 8 and 12 are too close to one another—a mere difference of 4. But 6-month-olds can discriminate 4 from 8. The key fact here is that 4 vs. 8, 6 vs. 12, and 12 vs. 24 all instantiate the same 1:2 ratio. Six-month-olds can discriminate quantities at this threshold, but not differences that instantiate the more fine-grained ratio of 2:3 (Xu et al. 2005).[17] The same pattern with these ratios has been found as well using auditory stimuli (different numbers of tones) and visual depictions of events (puppet jumps), not just static visual objects (Lipton and Spelke 2003; Wood and Spelke 2005).[18]

What about newborns? Is there any evidence that the approximate number system is present this early in development? We will present two studies that argue that it is. The first of these tested whether newborn infants can detect numerical correspondences between stimuli in different sensory modalities (Izard et al. 2009). This required using an intermodal matching procedure in which infants are initially familiarized with a numerical quantity in one sensory modality (e.g., hearing) and then given numerically matching and non-matching

[16] It turns out that the approximate number system continues to function throughout the human lifespan even though most adults have access to more precise, conventional ways of representing numerical quantity (Halberda and Feigenson 2008). And as we will see in Chapter 10, it can be found in many non-human animals as well.

[17] Other work has documented that the ratio at which infants discriminate between differing numerical quantities becomes more fine-grained in development (see, e.g., Xu and Arriaga 2007). There is also a question about whether the discriminative capacity of the approximate number system is context dependent (Wang et al. 2018).

[18] Interestingly, when the same sorts of experimental procedures are used to determine how well infants discriminate continuous magnitude or continuous extent properties (as opposed to numerical properties), infants have been found to have a more fine-grained ability to discriminate number than continuous extent. For example, using the same sort of habituation paradigm as Xu and Spelke (2000) used for number, Cordes and Brannon (2008) showed that 6-month-olds require a 1:4 ratio to discriminate differences in surface area, but only a 1:2 ratio to discriminate differences in numerical quantity. And in related work using a change detection paradigm, 6-month-olds have been found to require a 1:3 difference to detected changes in perimeter (Starr and Brannon 2015) and even more of a difference to detect changes in area (Libertus et al. 2014), despite the fact that they can detect a 1:2 change in numerical quantity (Libertus and Brannon 2010). Moreover, when three-fold changes in number were pitted against five-fold changes in area, infants preferred to look at changes in number (Libertus et al. 2014). It would appear that infants find numerical quantity to be both easier to process and more interesting than continuous extent.

stimuli in a different sensory modality (e.g., vision). The newborn infants, in this case, were familiarized with a given number of syllables (4 for some infants, 12 for others) and then simultaneously presented with two visual collections—one with the same familiar number of items (e.g., 4 green triangles if they had heard 4 syllables), the other with a novel number (e.g., 12 pink circles if they had heard 4 syllables). The result was that the infants preferred to look at the matching number. Thus, not only do newborn infants represent numerical quantity, but their representations of numerical quantity are highly abstract, applying across different sensory modalities and allowing them to detect cross-modal numerical correspondences.[19]

Notice that, in this study, there is no question of the infants' behaviour being explained in terms of their making use of correspondences involving non-numerical features such as total surface area or surface luminance, since one set of stimuli—the syllables—doesn't have these properties at all. Likewise, the infants couldn't have been responding to a correspondence in, say, the cumulative amount of acoustical energy (or most other auditory properties) since the visual stimuli didn't have any auditory properties at all. Moreover, the stimuli were designed so that the infants could not even use an analogy to connect what they heard with what they saw (e.g., mapping duration to area). This was achieved by varying the individual syllable length so that the total duration of the 4-syllable stimuli was the same as the total duration of the 12-syllable stimuli (i.e., each individual sound in the 4-sound sequences lasted as long as three of the sounds in the 12-sound sequence). In this way, the total amount of sound was equated in the familiarization phase of the experiment and any analogy between amount of sound and amount of surface area wouldn't work.[20]

[19] For related neurological evidence for abstract numerical representations in slightly older infants, see Gennari et al. (2023).

[20] Some have argued that the ratio-dependent representation of numerical quantity is grounded in a generalized magnitude system, one that represents different types of quantities in an undifferentiated way, not just numerical quantity (Walsh 2003; Leibovich et al. 2017). For example, Walsh's ATOM theory ("a theory of magnitude"), which we briefly discussed in Chapter 2, claims that numerical, temporal, and spatial stimuli are processed by the same mechanism. This view is motivated, in part, by people's performance in dual-task experiments that require representing two of these magnitudes simultaneously (e.g., performance in certain temporal tasks is impaired by a concurrent numerical task; see Casini and Macar 1997). What should we make of this? Many theorists associate this sort of proposal with empiricist approaches to explaining the origins of number concepts. However, even if it were granted that the ratio-dependent representation of numerical quantity is grounded in an innate generalized magnitude system, there would still be a question about how empiricist this system is. Notice that if it were confined to the representation of just these three magnitudes—space, time, and number—representing them in an undifferentiated way as spatio-temporal-numerical magnitudes—then as we suggested in Chapter 2, this system would still be an innate domain-specific system. It would just be less well-aligned with the target domain of numerical quantities than a number-specific system (Margolis and Laurence 2023). Accordingly, it would still support a rationalist view of the origins of representations of numerical quantity, albeit a somewhat milder form of rationalist view. But there is also a further question of whether the evidence that has been cited in favour of a generalized magnitude system of this sort really does argue for representations that fail to differentiate between spatial, temporal, and numerical magnitudes. In our view, this evidence at best

The second study with newborns that we will discuss shows not only that newborns represent numerical quantity but also that their representations of numerical quantity have a signature property that has often been taken to be a by-product of cultural conventions (Di Giorgio et al. 2019). This is the fact that numbers (for adults) are represented as if plotted on a number line with smaller numerical quantities on the left and larger numerical quantities on the right. Newborns were first habituated to different arrangements of a given number of squares (4 for some infants, 36 for others) and then simultaneously presented with two collections of 12 squares, one to their left and one to their right. (The number 12 was chosen in light of the fact that newborns can discriminate a 1:3 difference in numerical quantity and hence ought to be able to detect both 4 vs. 12 and 12 vs. 36.) Notice that if newborns associate smaller numbers with their left side and larger number with their right side, then which collection of 12 squares they prefer to look at should depend on whether they represent 12 to be smaller or larger than the numerical quantity they were habituated to. This is exactly what the researchers found. Infants who had been shown instances of 4 in the habituation phase preferred to look at the collection of 12 squares on the right, while infants who had been shown instances of 36 in the habituation phase preferred to look at the collection of 12 squares on the left.

Of course, there is always a question about whether infants are responding to numerical properties of the stimuli as opposed to differences in their continuous extent. To deal with this issue, the researchers equated the total perimeter of the habituation and test stimuli for both groups of infants. For example, the total perimeter of the 36 squares used in the habituation phase was identical to the total perimeter of the 12 squares used in the test conditions. So the infants could not have been ordering the habituation and test stimuli on the basis of total perimeter. What's more, as a consequence of equating the total perimeter in this way, the total surface area and luminance of the 36-square habituation stimuli was *smaller* than the total surface area and luminance of the 12-square test stimuli. Thus if infants were ordering left to right in terms of either of these continuous properties, their preference in the test condition would have been to look at the opposite display than the one they actually preferred to look at.

So, incredible as it may seem, newborn infants are able to represent approximate numerical quantities. Their representations of such quantities are highly

shows that there are strong links between these three types of magnitudes and that the neural systems involved in their representation may overlap. In fact, there is direct evidence that human newborns represent numerical quantity independently of space and time (and vice versa) and hence that the representation of numerical quantity as such is in place well before children could learn how to disentangle numerical quantity from other magnitudes (de Hevia et al. 2014). For this reason, we'd argue that the early links between spatial, temporal, and numerical representation aren't a reflection of a single generalized magnitude system but are instead evidence for an even more rationalist account involving several innate domain-specific systems (for space, for time, and for number) with articulated innate links between them (articulated in the sense of Chapter 2).

abstract (applying across different sensory modalities) and even exhibit a context-sensitive spatial mapping that has been found in studies with adults (which places smaller quantities on the left and larger quantities on the right).

Much more could be said about infants' numerical abilities. As we saw with the representation of faces, the innate abilities we have focused on are part of a cluster of related abilities. For example, there is evidence that newborns can represent small numbers of entities too (in the 1–3 range) (Turati et al. 2013). There is a lively debate about the format of these representations (see, e.g., Simon 1997; Uller et al. 1999; Clearfield and Mix 2001; Hurford 2001; Spelke and Tsivkin 2001; Feigenson et al. 2002; Le Corre and Carey 2007; Barner 2017). While some theorists have supposed that infants distinguish between small numbers of objects without representing their numerical quantity per se, these so-called non-numerical accounts often find themselves crediting infants with the ability to compare collections in terms of which collection has numerically more (or less) by performing a one-to-one mapping between their elements. This of course requires representations that support such mappings and the inferential apparatus for deducing that the one collection has more than, or is numerically different from, the other. Also, a case can be made that infants actually represent small-number numerical quantities as such, and that this further ability traces back to a second domain-specific innate system, one that is more precise than the approximate number system yet is limited to small numerical quantity (Laurence and Margolis 2005, 2007; Margolis and Laurence 2008; Margolis 2020).

There are also interesting representational abilities relating to the input to infants' numerical systems of representation and to the processes that operate on their specifically numerical representations. For example, newborns aren't just able to respond to quantities of individuals. They can also respond to quantities of *groups* of individuals, distinguishing between 2 groups of three and 3 groups of two (Turati et al. 2013; see also Wynn et al. 2002). And older preverbal infants have been shown to recognize the difference between correct and incorrect additions and subtractions for large approximate numerical quantities (McCrink and Wynn 2004). At the very least, then, infants' numerical abilities arguably involve not just representations for approximate numerical quantities but also representations of small precise numerical quantities, groups, numerical sameness/difference, being numerically greater than/less than, addition, and subtraction.

As with the face domain, there are questions about whether some or all of the representations that have been found to be present at birth in the number domain are nonconceptual, rather than conceptual. If representations of the numerical quantities *one*, *two*, and *three* were present at birth—that is, if representations of these small precise quantities and not approximate or non-numerical surrogates for them were present at birth—these representations would uncontroversially count as conceptual, given their continuity with the adult conceptual system for representing natural numbers. In Chapter 10, we will also see that a good case can

be made that the representations that come with the approximate number system are concepts. For now, we will just note that regardless of whether these representations are taken to be concepts, the work we have been discussing provides strong support for a rationalist account of concepts of natural numbers, as well as the various sorts of numerical relations we have mentioned (addition, subtraction, numerical sameness/difference, being numerically greater than/less than), and concepts of groups, objects, events, and sounds. This is because, even if the representations supporting numerical cognition at birth aren't themselves concepts, these representations and the characteristically rationalist psychological structures in the acquisition base from which they derive would undoubtedly play an important role in the acquisition of such concepts.

Suppose, for example, and for the sake of argument, that the approximate number systems' numerical representations aren't concepts. Still, to the extent that the approximate number system and its representations of numerical quantity are a part of the story of how other numerical representations are learned—ones that are uncontroversially considered to be concepts—the overall account supports concept nativism by supporting a rationalist account of the origins of these concepts. And we think it is all but inevitable that the approximate number system plays a significant role in the acquisition and use of the numerical concepts that are associated with the counting terms and with other conventional numerical symbols. This is because traces of the approximate number system's representations can be found in just about any task where adults find themselves using these things (Dehaene 1997).

For example, when adults are given pairs of Arabic numerals, it takes them longer to judge which represented quantity is numerically larger when they represent numerical quantities that are closer to one another. Moreover, this happens even for pairs like 65 vs. 72 and 65 vs. 79, where the comparison can be made without even attending to the second digit in the numeral expressing the larger number—that is, it takes longer to judge that 72 is larger than 65 than it takes to judge that 79 is larger than 65, even though both judgements could be made as soon as one has the information that the number being compared to 65 is a number in the 70s. As Dehaene (1997) explains, the natural explanation of this and related phenomena is that processing these symbols automatically activates the corresponding approximate numerical values and automatically compares them to one another. The reason it takes longer to compare pairs of numbers that have smaller numerical differences is because these differences are less pronounced in the analogue representational format in which these approximate representations are presumed to occur, making the comparison more difficult.[21]

[21] Another indicator of the link between approximate and precise numerical processing stems from the fact that individual variation in formal mathematical abilities—conceptual abilities par excellence—are strongly correlated with corresponding individual variation in the approximate

So there is good reason to suppose that approximate numerical representations are constitutively tied to adult numerical concepts and actively involved in the processes underpinning conscious judgements about relations between numerical quantities expressed in natural language. As a result, even if these approximate representations are nonconceptual, then, given the evidence for a rationalist account of the origins of approximate numerical representations, they will nonetheless be involved in a rationalist account of the origins of numerical representations that everyone agrees are conceptual, namely, concepts of natural numbers.[22]

Our discussion in this chapter has been confined to instances in which there is evidence for a representational ability at birth which couldn't plausibly have been learned prenatally. Our examples to illustrate this particular dimension of the argument from early development looked at the case for characteristically rationalist psychological structures in the acquisition base that factor into rationalist accounts of the origins of concepts pertaining to humans/persons, faces, individual people, specific faces, communication, and a variety of related concepts involved in social interactions, as well as numerical concepts and concepts of groups, events, and various types of numerical relations. Although evidence for a representational ability at birth contributes to a particularly strong form of the argument from early development, it isn't essential. In the next chapter, we show that the argument from early development can provide compelling grounds for a rationalist account of the origins of concepts even when there is no evidence that a representational ability is present at birth.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0008

number system. For example, 14-year-olds who have been found to have greater acuity in their capacity to make approximate numerical discriminations have been shown to have greater achievement in standardized mathematics examinations going all the way back to kindergarten (Halberda et al. 2008). Such differences also appear to explain certain mathematical learning disabilities. Ninth-graders with a history of difficulty with mathematics have been found to have lower acuity in their capacity to make approximate numerical discriminations even when researchers controlled for domain-general abilities that have been thought to be critical to mathematical performance (e.g., working memory) (Mazzocco et al. 2011).

[22] We saw in Chapter 2 that Carey's (2009) rationalist account of the origins of precise numerical concepts doesn't make use of the approximate number system in that she doesn't think the approximate number system is involved in the earliest stages in which children begin to understand the meanings of counting terms. But even Carey sees a role for innate approximate numerical representations factoring into the full conceptual role of our concepts of natural numbers. For Carey, adding these later in development supports numerical comparisons and the ability to do mental arithmetic, and so they play an important role in the acquisition of the full adult competence underpinning numerical concepts even on an account like hers which doesn't take them to play a role in the earliest stages of number concept acquisition.

9
The Argument from Early Development (2)

The argument from early development supports concept nativism through its focus on representational abilities that emerge too early in development to reliably be acquired solely on the basis of empiricist learning mechanisms. What it aims to show is that the timing regarding the development of these abilities offers compelling evidence for some concepts either being innate (that is, part of the acquisition base) or being acquired on the basis of rationalist learning mechanisms. In the previous chapter, we examined one facet of the argument from early development—cases where there is evidence that a representational ability is present at birth. However, we also noted that the argument isn't restricted to these sorts of cases and that a strong version of the argument from early development can be formulated for representational abilities that appear later in development as well. In this chapter, we turn to this second facet of the argument from early development.

To begin, it's important to recognize that concept nativism isn't just—or even primarily—about the minds of newborns. While some innate psychological traits may be present at birth, others may not appear until later with further biological development. Also, as we emphasized earlier, concept nativism isn't simply about innate concepts—concepts acquired via rationalist learning mechanisms are no less important. Finally, even when innate concepts and representational abilities are present at birth, they may not be evident until later due to performance factors preventing the underlying competence from manifesting itself.[1]

But how could evidence that a trait is present later in development—at 6, 12, 18 months, or even later—provide strong grounds for thinking that the trait is possessed *too early* in development to reliably be acquired solely on the basis of empiricist learning mechanisms? The key point is that this might still be too early because for one reason or another the kinds of information a learner with only domain-general learning mechanisms would need to have to reliably acquire the trait isn't available to such learners by this point in development. For example, if a domain-general learning mechanism required extended exposure to information that is only readily conveyed through language, then acquisition of the trait by prelinguistic infants would provide a powerful argument that the trait is present

[1] See Chapter 2 for a discussion of the competence-performance distinction. Newborn infants have little control over their bodies and have a poor attention span and highly limited memory (see, e.g., Oakes and Luck 2014), and so there is the very real possibility of a major gap between their cognitive and conceptual capacities (their competence) and their ability to selectively display these capacities in their overt behaviour (their performance).

too early to reliably be acquired solely via domain-general learning mechanisms. It's for this reason that empiricists themselves expect certain traits to not be present at birth or until much later (for example, not until the child is old enough to have acquired much of her language). These brief remarks should suffice for now to convey the core idea behind how the argument from early development can be extended beyond the case of traits that are manifestly present at birth. We will say more about the form of the argument later in the chapter.

Much of our discussion in this chapter will be focused on a single example: the development of concepts for mental states. The representation of what might be going on in other people's minds is a ubiquitous feature of adult cognition. We do this when we try to figure out someone else's motivation, predict their actions, or make sense of their behaviour. Many mentalistic concepts play a role in reasoning about these matters, including concepts for beliefs, desires, intentions, emotions, moods, preferences, memories, personality traits, and perceptual states (among others).[2] Though there is experimental work bearing on many of these concepts, the most intensively investigated has been FALSE BELIEF.[3]

The development of this ability has been at the very centre of the rationalism-empiricism debate about the origins of concepts for many years. It is perhaps the single most important focal point of this debate and a paradigmatic example of how the argument from early development has been used in cases where a trait isn't present at birth. It is also one of the most intensively studied domains, with an enormous body of experimental results and theoretical work by both rationalists and empiricists. So, while this example is only meant to illustrate the broader form of argument, to do justice to it will require an extended discussion of a range of empirical and theoretical considerations. Towards the end of the chapter, we will briefly consider the wider context of the debate in neighbouring domains concerned with other facets of social cognition, and beyond.

Until relatively recently, the orthodox view was that the ability to represent false beliefs doesn't appear until children are around 4 years old and that the first reliable sign that children have this ability is that they begin to pass a traditional false-belief test, for example, Baron-Cohen et al.'s (1985) Sally-Anne test. In this test, children are asked questions about a depicted sequence of events in which one character (Sally) sees a desirable object (e.g., a marble) in one location, leaves

[2] Beliefs and desires are often taken to be paradigmatic *propositional attitudes* (and BELIEF and DESIRE to be among the most basic propositional attitude concepts). These are mental states that have full propositional content in that it is of a type that might be captured by a sentence as opposed to a word or phrase. The "attitude" part of a propositional attitude refers to the way a thinker is related to this content, for example, whether they take it to be true of the world (as with a belief) or whether it is a way they want the world to be (as with a desire).

[3] Why has so much attention been paid to attributing *false* beliefs as opposed to *true* beliefs? Notice that if someone appeared to predict an agent's behaviour on the basis of the agent's true belief, she might actually be predicting the behaviour without really thinking about what the agent believes at all and instead just be thinking about the constraints on successful action in light of how the world happens to be—a reality-based strategy. On the other hand, a reality-based strategy is blocked if the agent's action is rooted in a mistaken view of the world (Dennett 1978).

the room, and while she is gone, the object is moved to a different location by another character (Anne), who then leaves the scene. When Sally returns, the child is asked where Sally will look for the object. The right answer, of course, is that she will look in the original location, since that is where she would (falsely) believe it to be. However, children younger than 4 years old generally say that she will look in the new location, where it actually is (Baron-Cohen et al. 1985). This finding, which became the cornerstone for much research on mentalizing, led most theorists to conclude that a major conceptual change occurs around the age of 4.[4]

But notice how demanding this false-belief task is. Not only does it require children to keep track of a considerable amount of information (where the marble was, where it is now, whether Sally could have seen the marble in the new location, etc.) and to ignore their own knowledge regarding the object's actual location, but all of this also happens in a social context in which children need to make sense of the question they are being asked and provide an appropriate verbal response (Bloom and German 2000; Baillargeon et al. 2010). What if a simpler test could be devised which doesn't involve all of the complications associated with traditional false-belief tasks but which could nevertheless show that children are representing that an agent has a false belief? Until 2005, no one had figured out a way to do this, but that changed with Kristine Onishi and Renée Baillargeon's ground-breaking work with 15-month-old infants (Onishi and Baillargeon 2005).[5]

Onishi and Baillargeon used what is known as a *violation of expectation* (VOE) methodology. Since this methodology resembles some of the other methods for studying infants that we have already encountered in that the pivotal measure is infants' looking time, it is worth saying a few words about how looking time can figure in different methodologies.[6] The key point is that looking time is a behavioural measure that can have different significance in different contexts, just as other behavioural measures can (e.g., pressing a button). When we say that researchers employed a particular looking-time methodology (e.g., the VOE method as opposed to the habituation method), what we mean by this is that their experiment is designed in such a way that particular patterns of looking times, across the various conditions in the study, will provide reason to interpret

[4] A number of different terms are used in philosophy and cognitive science to refer to the ability to attribute and reason about mental states. We will often use the term "mentalizing", but other commonly used terms include "mindreading", "theory of mind", and "folk psychology".

[5] Following Scott and Baillargeon (2017), we will use the term *traditional false-belief task* to refer to tests (like the Sally-Anne test) in which children must provide an explicit verbal answer when asked about what an agent believes or something that turns on what she believes (e.g., Where will Sally look for her marble?). In contrast, a *non-traditional false-belief task* is one in which the test doesn't require answering such questions. As we will see, the innovation in Onishi and Baillargeon's study was its reliance on infants' spontaneous responses to various situations involving true and false beliefs. For this reason, non-traditional false-belief tasks are often also referred to as a *spontaneous-response tasks*.

[6] For more on the history of the VOE method and further analysis of its advantages and limitations, see Margoni et al. (2022).

infants' looking-time behaviour in line with the interpretation associated with this methodology (while ruling out alternative interpretations). In a habituation study, for example, longer looking in the test trials is meant to track the perception of novelty. Alternative interpretations—for example, that the infants look longer simply because they have a preference for a particular type of stimulus or a contingent preference for stimuli presented on their right side—are systematically evaluated and ruled out through control conditions.[7] The same goes for VOE studies, which interpret infants' longer looking times in terms of the infants' expectations being violated. What determines the interpretation of the infants' looking time in such studies isn't a matter of stipulation by the experimenters, but rather is a question of what interpretation best fits the overall pattern of looking behaviour exhibited by the infants. A well-designed VOE study anticipates alternative explanations of longer looking times and is able to exclude them. If infants were to look longer for reasons having nothing to do with their expectations being violated, the experimental design would allow us to see this.

A typical VOE experiment involves presenting infants with a scenario that would violate adult expectations, and the question is whether it also violates the expectations of infants. If it does, this means that infants have relevantly similar expectations to adults. One feature of such experiments is that they often include pretest trials known as *familiarization trials.* In the previous chapter, we noted that habituation studies use numerous pretest trials (*habituation trials*) to reduce infants' interest in stimuli of a given type. It is important not to conflate these two different types of pretest trials. While VOE studies may appear to have a similar structure in that they often use pretest trials too, their pretest trials have a very different aim. The familiarization trials used in VOE studies function solely to prepare infants to encode and respond to events in the test trials that might be too complicated for them to take in without this extra support. The point of a familiarization trial isn't to make infants bored (indeed, doing so would be counterproductive). It is to introduce them to the overall experimental setup (e.g., to the objects and agents they will encounter) and to facilitate their processing and remembering relevant aspects of the events that follow. Familiarization trials are not a mandatory component of a VOE experiment—some VOE experiments don't use them. But they can help make experiments involving complex scenarios more intelligible for infants.

The typical structure of a VOE study involves both experimental and control conditions. The experimental condition begins with one or more familiarization trials. These are followed by a short series of events that would set up certain

[7] Or they are ruled out using standard experimental precautions, such as presenting the target stimuli on different sides to different groups of infants.

expectations for an adult observer. Finally, infants are presented with a test trial that either fits with these expectations or conflicts with them. If infants have similar expectations to adults, they should look longer at the unexpected event, which is more interesting since it violates their expectations about what should happen. Control conditions follow a similar general structure, while ruling out alternative interpretations.

In Onishi and Baillargeon's study, infants were first familiarized to a setup involving an agent, two boxes, and a small toy. In the familiarization trials, they saw the agent pick up the toy and put it in one of the boxes, and then, in successive familiarization trials, reach into this same box (Box-1). Next, they saw the agent leave the scene, followed by the toy moving to the second box (Box-2) while the agent was absent. Then the agent returned, and infants saw her reach into either Box-1 or Box-2. The key finding was that 15-month-olds looked longer when they saw the agent reaching into Box-2—where the toy really is, not where the agent last saw it. This suggests that they expected her to reach for the toy where she falsely believed it to be and so were surprised when she reached into the box where it actually was instead.

Of course, this isn't the only possible explanation of the infants' looking behaviour given just the result mentioned so far. To establish that infants are truly representing that the agent has a false belief about the toy's location, low-level alternative explanations have to be ruled out. But given the results of the other conditions in their study, such alternatives aren't especially plausible. For example, take the proposal that the infants were merely responding to a preference for novelty. When the agent reaches into Box-2 in the test trial, this isn't just a different box than the one she should falsely believe to contain the toy. It is also a different box than the one she had repeatedly reached into during the familiarization trials. A simple alternative hypothesis might hold that infants prefer to see a novel action (or even just a novel type of motion) and that this leads them to look longer when the agent now reaches into Box-2 for the first time.

We can see that this isn't a viable explanation, however, by attending to one of the other conditions in the study. In this other condition, everything was the same as in the original condition except that the agent never leaves the scene and so she sees the toy move to Box-2. Given just this one small change, the simple novelty hypothesis and the belief attribution hypothesis are easy to tease apart—in fact, they make opposite predictions. The belief attribution hypothesis predicts that infants won't attribute a false belief to the agent in *this* scenario since the agent saw the change of location. Instead, they should attribute to the agent the *true* belief that the toy is in Box-2 and so they should be surprised, and so look longer, if the agent reaches into the *other* box (Box-1). In contrast, the simple novelty hypothesis continues to predict that infants should look longer when the agent reaches into Box-2, for the same reason as in the original condition—reaching into Box-2 is more novel after seeing the agent previously only reach

into Box-1 during the familiarization trials. It turns out that the results clearly favour the belief attribution hypothesis. When infants could see that the agent was in a position to see the object move to Box-2, infants looked longer if she reached in Box-1, suggesting that novelty isn't the driving factor.[8]

We will have more to say about a more sophisticated version of the novelty hypothesis shortly. But for the moment, we just want to remark on how extraordinary Onishi and Baillargeon's study was in that it not only devised a non-verbal method for investigating whether infants and very young children are able to attribute false beliefs to an agent but also came to the conclusion that this ability is present in children as young as 15 months old. This was an astounding finding, pushing back the initial appearance of false-belief representation from kindergarten to infancy.

Subsequent work, using a variety of different research methods, has since corroborated this core finding with infants and toddlers. For example, another type of non-traditional false-belief task asks whether young children will *spontaneously point* in a helpful way when the agent is about to make a mistake because of a false belief (Knudsen and Liszkowski 2012a, 2012b).[9] In one such task, 18- and 24-month-olds saw an agent look for an object among four containers, eventually finding it in the last of the four containers. Following this, in one condition, another individual switched the object's location to one of the other containers while the agent was away and so could not see what was happening (false-belief condition). In another condition, the switch was made before the agent left the room and so the agent saw the switch happen (true-belief condition). During the test phase, the agent re-entered the room and proceeded to a central location in front of the four containers. Then the experimenters recorded whether infants spontaneously pointed to the container that actually held the object before the agent reached for one of the containers. In fact, they did, but only when the agent hadn't been present to see it had been moved—that is, only in the situation in which she was poised to reach for the wrong container because she falsely believed it contained the object she was looking for (Knudsen and Liszkowski 2012a).

In a related form of non-traditional false-belief task, young children are encouraged to assist an agent in a situation where being helpful requires an appreciation that the agent has a false belief (Buttelmann et al. 2009; Buttelmann et al. 2014). For example, in one task, 18-month-olds saw some boxes that contained blocks and saw an agent removing several of the blocks to create a block

[8] Other low-level alternative explanations that were addressed in Onishi and Baillargeon's original study include the possibility that infants prefer looking at a particular box or in a particular direction, or that their attention is drawn to where the agent last reached, where the agent's attention was last, or where the object is actually located.

[9] Other work has shown that infants and toddlers in this age range spontaneously help others through pointing and other means (see, e.g., Warneken and Tomasello 2006).

tower. Next, while the agent was away (false-belief condition) or before the agent left the room (true-belief condition), the experimenter revealed that one of the boxes (which we'll refer to as *the target box*) contained a spoon, not a block, even though it had a picture of a block on the outside just like the other boxes. Then the agent returned with an empty bowl, making it unclear whether she might want a spoon (which goes with a bowl) or another block (to continue building the block tower). In what followed—in the test phase—the child was given access to both a spoon and a block while the agent indicated she needed help as she reached for the target box. Notice that in the false-belief condition, the agent wouldn't realize that the target box contained a spoon, so the agent should be interpreted as wanting a block. In contrast, in the true-belief condition, the agent would realize that the target box contained a spoon, so the agent should be interpreted as wanting a spoon. The question was whether the children would respond differently in the two conditions. The result was that they did. They gave the agent a spoon when the agent held the true belief and a block when the agent held the false belief (Buttelmann et al. 2014). In other words, they recognized that, to be helpful, they shouldn't necessarily give the agent the type of object that was *actually* inside the target box; they should hand over the type object that the agent *believed*—and sometimes *falsely believed*—was in the target box.

We will mention just one other non-traditional false-belief task. This one used a *preferential-looking* paradigm, another type of experimental method, taking advantage of the fact that children, like adults, spontaneously look at pictures that match the content of a story (Scott et al. 2012). In this study, 2.5-year-olds viewed a picture book while they heard a simple false-belief story with a narrative that closely mimicked the classic Sally-Anne task. (Emily placed her apple in one of two containers, but while she was taking a nap, Sarah transferred it to the other container.) After the initial introduction of the two characters, each line in the story was followed by showing the children two pictures at the same time, one that matched the narrative and one that didn't. As expected, the children looked longer at the matching pictures than the non-matching pictures. The crucial question was which picture the children would spontaneously look at given the final line of the story when they were told that Emily looked for her apple. One picture depicted Emily searching in the original location (the false-belief solution), the other depicted her searching in the new location that the apple had been moved to when she was asleep (the reality-based solution). The genius of this experiment is that since the children had looked at the pictures that matched the story up until this point, the one they looked at more at the end of the story tells us where they thought Emily would search for the apple—eliminating the need to explicitly ask them where she would look. It turns out that they looked longer at the picture of Emily searching in the original location (which is where adults would expect her to search, given her false belief that this is where the apple is). In contrast, in a further experimental condition in which Emily was

present when the transfer took place, they looked longer at the picture of Emily searching in the new location (which is where adults would expect her to search, given her true belief that this is where the apple is). Taken together, these results indicate that the children spontaneously anticipated her behaviour would be in accordance with her true or false belief, and hence that they are capable of attributing both true and false beliefs agents.[10]

Dozens of further experiments using these and other methodologies have now shown that children significantly younger than 4 years old can represent false beliefs, in some cases, succeeding in non-traditional false-belief tasks at as young as 6 or 7 months of age (Kovács et al. 2010; Southgate and Vernetti 2014; Hyde et al. 2018).[11] Taken at face value, these findings provide evidence for a rationalist account of false-belief understanding, and they have been widely interpreted as providing such evidence. They strongly suggest that performance limitations are responsible for children's failure to pass traditional false-belief tasks at older ages and that their difficulties with these more demanding tasks do not accurately reflect their conceptual competence.

We now need to consider some criticisms and alternative interpretations of this work. Much of the criticism of this work has centred around two general approaches, which we will refer to as *deflationary accounts* and *dual-system accounts*. Deflationary accounts deny that infants' and toddlers' apparent success with non-traditional false-belief tasks has anything to do with them possessing an understanding of others that involves explaining what someone does in terms of their mental life. On a deflationary account, what might initially look like the representation of false beliefs turns out to be something far more modest. We will discuss two representative examples of this type of approach. On the first, children are merely forming an association between an agent, a particular object, and a location (Perner and Ruffman 2005); on the second, they are merely responding to low-level properties of the stimuli (e.g., looking longer simply because one scene involves an interesting movement or colour that captures their attention) (Heyes 2014). Both sorts of deflationary accounts take infants' apparent success

[10] This study makes use of yet another reason why infants might look longer at a given stimulus—namely, that they prefer to look at pictures that they take to match what is happening in the story that they are listening to. As we note in the text, the researchers were able to confirm that infants' looking is guided by this preference by comparing how much they look at relevant versus irrelevant pictures as the story unfolds, where the question being investigated of whether they understand false belief is not at issue at all.

[11] Such research isn't confined to behavioural measures. For example, Hyde et al. (2018) used a neuroimaging technique (functional near-infrared spectroscopy) that is suitable for studying the brain activity of infants while they are watching videos. Seven-month-olds were shown scenarios in which an agent held a true or false belief about the location of an object. The results showed that 7-month-old infants exhibit the same overall pattern of neural activity as appears in adults engaged in tasks involving the attribution of beliefs—selective activation in the temporal-parietal junction (TPJ), including more activation when the agent in the video held a false belief than when she held a true belief.

in false-belief tasks to have nothing at all to do with the infants trying to account for an agent's behaviour in terms of attributing psychological states to them. Dual-system accounts, in contrast, do grant that the experiments demonstrate the presence of a representational ability that is specifically targeted at explaining behaviour. But this ability is taken to be grounded in its own early developing system for explaining behaviour, which is distinct from a later developing system that subserves the ability to represent false beliefs (and other propositional attitudes) and that comes online around the time that children begin to pass traditional false-belief tasks. Dual-system accounts interpret this early representational ability to consist in the representation of behavioural rules (Perner and Ruffman 2005) or the representation of more basic types of mental states that fall short of being fully belief-like or desire-like (Apperly and Butterfill 2009; Butterfill and Apperly 2013).[12]

A comprehensive reply to these criticisms would require an extensive review of far more experimental work than would be reasonable for us to cover here. Instead, we want to provide enough of a response to indicate why we think that the argument from early development still applies to the case of FALSE BELIEF and also to illustrate some common problems with the sorts of alternatives that have been raised against more rationalist interpretations of infancy work. We will start with the deflationary accounts.

Consider again Onishi and Baillargeon's (2005) study with 15-month-olds in which an agent is seen to place a small toy in one location only to have it moved without her knowledge. Perner and Ruffman claim that the pattern of looking found in this study doesn't require attributing to infants' expectations about where the agent will reach, expectations that are violated in the false-belief condition. Rather, all we need is the supposition that infants represent person-object-location associations (that is, a representation that associatively links a person, an object, and a location). The thought is that forming new associations of this kind takes longer than the alternative in which a previously formed person-object-location association can continue to be used. And according to Perner and Ruffman, in the false-belief condition, infants have to form a new person-object-location association, and this, in turn, causes the longer looking times in this condition. Why do infants need to form a new association in this case? Infants first see the agent place the toy in Box-1, so this leads them to initially form the association <agent, toy, Box-1>. Then, in the false-belief condition, when the location of the toy is switched while the agent is absent and the agent subsequently reaches into Box-2, Perner and Ruffman see this as requiring the infants to form the new association <agent, toy, Box-2>.

[12] Burge (2018) develops a dual-system account that is in many ways similar to Butterfill and Apperly's more widely discussed view, though unlike Butterfill and Apperly, Burge considers the states involved to be non-mental. For critical discussion of Burge's view, see Carruthers (2020) and Jacob (2020).

This account (in terms of new associations) just doesn't work, however. If changes to three-way associations are driving infants' looking time, then infants should look longer if *any* of the three associated elements change. The problem for Perner and Ruffman's account is that an element changes in *both* of the tested outcomes in the false-belief scenario (reaching into Box-1, or reaching into Box-2). Reaching into Box-2 breaks the association in terms of *location* <agent, object, *new location*>, but reaching into the Box-1 equally breaks the association in terms of *object*, since the infant knows that there is no object in this box <agent, *no object*, original location>. So the three-way association account actually makes a different prediction than the one that Perner and Ruffman propose. It predicts that there shouldn't be any difference in looking time regardless of which box the agent reaches into. However, that is not what happens—infants look longer at the test event that is at odds with the agent's false belief about the object's location. So the association-based account cannot even explain the core finding in Onishi and Baillargeon's original spontaneous-response false-belief task.

There are other deflationary accounts to consider, however. Heyes (2014) gives an account that is particularly noteworthy since it provides an extensive and influential critical response to infancy research employing non-traditional false-belief tasks. Heyes presents deflationary alternatives for a broad range of findings that have been taken to support a rationalist account of the origins of the capacity to represent and attribute mental states to others, including false beliefs. The alternatives she proposes are ones which focus on low-level properties of the test stimuli, including seemingly small details regarding how the stimuli move, the timing of these events, and how these factors relate to infants' memory and perceptual preferences.[13]

In evaluating Heyes' account, a good place to start is with what she has to say about the results reported in Onishi and Baillargeon's classic (2005) study. There are four conditions in this study. To appreciate both the structure and prospects of Heyes' account, we will need to look at how it plays out for each of these conditions and how it compares to Onishi and Baillargeon's rationalist account. This means looking carefully at multiple potential explanations of a variety of conditions. So our discussion will necessarily get somewhat complicated. But please bear with us since, as Heyes herself would say, the devil is in the details.

The four conditions in Onishi and Baillargeon's (2005) study varied in two key dimensions. The first has to do with the location of the toy in the test trials—whether it ends up in Box-1 or Box-2. The second has to do with whether the events leading up to the test trials leave the agent with a true or false belief about

[13] As we will see in later chapters, this general strategy is a major theme in empiricist research. Rationalist theories of early conceptual development are regularly criticized on the grounds that they fail to sufficiently take into account the possibility of a low-level perceptual explanation. We will return to this type of criticism in connection with other conceptual domains in Part III (see particularly Chapters 17 and 18).

the toy's location. Following Onishi and Baillargeon, we will use the term *true-belief condition* for the conditions where the agent ends up with a true belief about the toy's location and *false-belief condition* for the conditions where the agent ends up with a false belief about its location. Of course, whether or not infants *represent* the situations to themselves in terms of the agent's beliefs is exactly what is at issue between the competing explanations of the study, not something that can be assumed. Nonetheless, as long as we bear this in mind, it is helpful for understanding the logic of the study and for keeping track of the various conditions to label them in this way.

First, let's look at the two true-belief conditions (the two in which the agent ends up with a true belief about the toy's location), since these are the ones where Heyes' account requires the fewest assumptions. In both true-belief conditions, the agent never leaves the scene and so is perfectly situated to see if the toy's location switches from one box to the other and to form a true belief about its location.

True-belief condition 1. In the first true-belief condition, following the familiarization trials (which are the same in all conditions), the agent puts the toy in Box-1.[14] Next, in clear view of the agent, the toy moves from Box-1 to Box-2. Then finally the scenario ends with the agent reaching into either Box-1 or Box-2. Since the agent is present the whole time and can see the toy move from Box-1 to Box-2, she has the true belief that the toy is in Box-2. In this condition, infants look longer when the agent reaches into Box-1 (where there is no toy).

True-belief condition 2. In the second true-belief condition, following the familiarization trials, the agent remains present the whole time, just as in the previous condition, but this time the toy remains in Box-1—it doesn't move to Box-2. (What happens in this scenario is that the agent puts the toy in Box-1, and while she remains in the room and can see what is happening, the toy just stays in Box-1.) Instead of the toy moving, Box-2 moves towards Box-1 and then back to its original position. In this way, infants still see some movement, but the movement doesn't involve the toy switching location. Following this, the scenario ends with the agent reaching into either Box-1 or Box-2. The agent also ends up with a true belief in this condition, although in this case the true belief is that the toy is in Box-1. In this condition, infants look longer when the agent reaches Box-2 (where there is no toy).

How do Onishi and Baillargeon explain the looking-time patterns in these conditions? On their rationalist account, infants expect the agent to reach into the box where she believes the toy to be. When this expectation is violated, they are surprised. In the first true-belief condition, the agent has the true belief that the toy is in Box-2, so infants are surprised and look longer when she reaches into Box-1. In the second true-belief condition, the agent has the true belief that the

[14] As we noted above, in the familiarization trials, the infants first see the agent pick up the toy and put it in Box-1, and then, in two successive familiarization trials, reach into this same box (Box-1).

toy is in Box-1, so they are surprised and look longer when she reaches into Box-2. In both cases, the explanation is perfectly straightforward. Infants are responding to the scenarios in much the same way as any adult would—by representing and reasoning about the agent's beliefs about the toy's location.

Now let's look at Heyes' (2014) deflationary explanations of these same results. We will start with the second true-belief condition (in which the toy remains in Box-1). Heyes' explanation is essentially the same as the alternative we mentioned when we first introduced Onishi and Baillargeon's study—she suggests that the infants are exhibiting a novelty preference.[15] Given that the agent only interacts with Box-1 in the familiarization trials, her reaching for Box-2 in the test trial is perceptually novel. Heyes' proposal is that this all by itself makes the agent's reaching into Box-2 more interesting than her reaching into Box-1. Infants look longer when the agent reaches into Box-2 not because their expectations about what the agent will do are violated but simply because this event is more novel and they find novel events more interesting to look at. In other contexts, infants sometimes do prefer novel stimuli, so this would seem to be a perfectly reasonable proposal for this one finding.

What about the first true-belief condition? In this condition, where the agent sees the toy move to Box-2, Heyes also takes the novelty preference to be in play. In this case, however, the situation is more complicated. This is because, although the infants are given the same familiarization trials in which the agent repeatedly reaches into Box-1, they *don't* look longer in the test trial in which the agent reaches into Box-2. Instead they look longer when she reaches once more into Box-1. On the face of it, this looks like the very opposite of a novelty preference; infants are looking longer at what would seem to be the more familiar action (once again reaching into Box-1). But Heyes claims that the low-level features of the scene are more subtle than they may seem at first in that the movement of the toy from Box-1 to Box-2 is "visually similar" to the movement of the agent reaching into Box-2. The impact of seeing this visually similar event is to "[reduce] the novelty of…the test event" in which the agent reaches into Box-2. Heyes takes this reduction in novelty to explain why infants now look longer when the agent reaches into Box-1 (Heyes 2014, p. 650).

Heyes does not provide any reason to believe that infants perceive the toy's moving into Box-2 as visually similar to the agent reaching into this same box. However, even if we suppose that it is and that this reduces the novelty of seeing the agent reach into this box, there is still a major problem for Heyes' explanation of the results for the second true-belief condition. This is that *reducing the novelty of reaching into Box-2* isn't the same thing as *making reaching into Box-1 novel*. So even if we suppose that the movement of the toy from Box-1 to Box-2 reduces the

[15] Our earlier suggestion of an explanation in terms of a simple novelty preference was directed to a different experimental condition, but otherwise the explanation is essentially the same.

novelty of the agent reaching into Box-2, this doesn't explain why the infants look longer when the agent reaches into Box-1. After all, *this event* isn't novel either. Recall that the infants, at this point in the experiment, have just seen the agent reach into Box-1 repeatedly in the familiarization trials (which is precisely why Heyes can point to the novelty of reaching for Box-2 in explaining the first true-belief condition). It would seem that, at best, neither reaching into Box-1 nor reaching into Box-2 is especially novel in this circumstance. So, if anything, Heyes' account actually predicts that infants should look equally at the two outcomes. Yet this isn't what happens, as infants look longer when the agent reaches into Box-1.[16]

False-belief condition 1. Let's turn now to the two false-belief conditions in the Onishi and Baillargeon study. The first of these is just the original condition we discussed when we first introduced Onishi and Baillargeon's study. In this condition, the agent puts the toy in Box-1, she leaves the room, and while she is away, the toy moves from Box-1 to Box-2, leaving her with a false belief about the toy's location. When the agent returns, she reaches into either Box-1 or Box-2, as in the other conditions. In this condition, infants look longer when she reaches into Box-2 (where the toy actually is) compared to Box-1 (where she should falsely believe it to be).

False-belief condition 2. The other false-belief condition is slightly more complicated. We just saw that in the first false-belief condition, the toy starts in Box-1 (as in the two true-belief conditions) and the agent ends up with a false belief that the toy remains located in Box-1. The point of the second false-belief condition is to cover the case in which the toy again starts out in Box-1 but the agent is led to have a false belief that the toy is in *Box-2* (rather than a false belief that it is in Box-1). To cover this possibility, the toy moves *twice* in this condition—once when the agent is still present and so can see the change of location, and then a second time when the agent is away and can't see the change. So, in this condition, the agent puts the toy in Box-1, and while she is still able to see what is happening, the toy moves from Box-1 to Box-2. Then she leaves, and while she is away and can no longer see what is happening, the toy moves back to Box-1. Since the agent saw the first move from Box-1 to Box-2 but did not see the second move from Box-2 back to Box-1, she ends up with the false belief that the toy is in Box-2. When the agent returns, she reaches into either Box-1 or Box-2. In this condition,

[16] Heyes (2014) suggests that there may be a further factor in play here (and likewise in relation to the second false-belief condition, discussed below). While the infants have seen the agent repeatedly reach into Box-1 a total of three times but have only seen a single event that is (according to Heyes) visually similar to the agent reaching into Box-2, the latter is more recent. Heyes claims that a recency effect makes reaching into Box-1 more novel as it will have "more than compensated for the higher frequency" of the reaches into Box-1 (Heyes 2014, supporting information, note 1). However, she gives no reason to suppose that the precise values of recency and frequency in play here lead to the desired result—she simply assumes this.

infants look longer when she reaches into Box-1 (where the toy actually is) compared to Box-2 (where she should falsely believe it to be).

What is Onishi and Baillargeon's rationalist explanation of the infants' looking-time patterns in these two false-belief conditions? On their account, these patterns are explained in exactly the same way as the infants' looking-time patterns in true-belief conditions: In both cases, the infants expect the agent to reach into the box where she believes the toy to be. The only difference is that in the false-belief conditions, the infants take the agent to have a false belief, rather than a true belief, about where the toy is. This means that they will expect the agent to reach into the box that the toy isn't in, since that's where the agent thinks the toy is. When this expectation is violated—when the agent reaches into the box that the toy *is* in—they are surprised and look longer. In the first false-belief condition, the agent has the false belief that the toy is in Box-1, so infants are surprised and look longer when she reaches into Box-2 (where the toy actually is). In the second false-belief condition, the agent has the false belief that the toy is in Box-2, so they are surprised and look longer when she reaches into Box-1 (where the toy actually is).

Now let's look at Heyes' deflationary explanations of the results in these conditions. As in the other conditions, the first false-belief condition involves the agent repeatedly reaching into Box-1 in the familiarization trials. Given Heyes' reliance on the assumption that infants prefer perceptually novel stimuli, this repetition might be thought to induce infants to look more at the agent reaching into Box-2. But this condition also involves the toy moving from Box-1 to Box-2, which, we saw earlier, Heyes took to be visually similar to the agent's reaching into Box-2, when she argued that this reduces the novelty of the agent reaching into Box-2. So this suggests that Heyes' account should predict that the infants should look longer when the agent reaches into Box-1.[17] In fact, though, they look longer when the agent reaches into Box-2. To explain this outcome without postulating that the infants attribute to the agent the false belief that the toy is in Box-1, Heyes brings in a new factor. She suggests that the infants become distracted by the agent's reappearance after the toy switches location and that this distraction cancels the effect of the toy moving from Box-1 to Box-2. Heyes' claim is that once the effect of the toy's movement is cancelled, infants should be back to preferring to look at the agent's reaching into Box-2 because it is more novel.

This explanation is consistent with the results in this condition, but what reason is there to believe it is the right explanation or that it is better than Onishi and Baillargeon's simpler rationalist explanation? Heyes doesn't cite any evidence that directly supports her claim that seeing an agent reappear in a scene after a brief absence has this effect. And it is doubtful that such a familiar type of event *would*

[17] Or, as we argued above, that they shouldn't have a preference one way or the other.

have this effect. Infants regularly experience people returning after a brief departure—this is a ubiquitous feature of daily life. It is also worth recalling that in the second true-belief condition there is an unusual event that might be thought to be at least as distracting to infants—when one of the boxes moves towards the other box and then back again. Although we didn't note this earlier, this movement is striking since the box moves on its own accord. If the agent's reappearance is distracting enough to make infants forget the toy's change of location in the first false-belief condition, one might easily suppose that the box's motion might be at least as distracting and make the infants forget some, or all, of the preceding familiarization trials in which the agent reaches into Box-1. But then Heyes' earlier explanation of this true-belief condition wouldn't work. She would no longer be in a position to say that the agent's reaching into Box-2 attracts the infant's attention because of its perceptual novelty. In the end, we see no reason why one of these events, but not the other, should be deemed a "salient distractor". To the extent that Heyes' explanation requires the unsupported stipulation that one of these events is distracting and other not, her overall account would seem to be inconsistent and open to the charge of making ad hoc assumptions.

Finally, let's look at Heyes' explanation of the second, more complicated false-belief condition. As in the other conditions, the second false-belief condition begins with the agent repeatedly reaching into Box-1 in the familiarization trials, so in accordance with Heyes' explanations of earlier conditions, this should set up a novelty preference for the agent reaching into Box-2. But, as in the previous condition, the toy then moves from Box-1 to Box-2, which Heyes takes to be visually similar to the agent's reaching into Box-2, thereby reducing the novelty of the agent's reaching into Box-2. But then, while the agent is away, the toy moves back to Box-1. If we assume, with Heyes, that the toy moving from Box-1 to Box-2 is visually similar to the agent reaching into Box-2, then presumably we should likewise suppose the toy moving from Box-2 to Box-1 is visually similar to the agent reaching into Box-1. So one might have thought that Heyes should maintain that this second move by the toy would reduce the novelty of reaching into Box-1, swinging the pendulum back so that reaching into Box-2 would be perceived to be more novel once again. But we also need to take into account the impact of the agent's reappearance in the final stage of the scenario. As before, Heyes takes this to be highly distracting to the infants. The crux of her explanation of this condition is that the agent's return cancels the effect of the last movement of the toy (i.e., its movement from Box-2 back to Box-1 while the agent is away). With this last assumption in place, Heyes is able to predict that infants should find the agent's reaching into Box-1 more novel and hence that they will look longer at this event, which is precisely what they do.

This explanation is consistent with the results in this condition, but it is again open to the charge that it relies on ad hoc assumptions. For example, we noted

earlier that there is no reason to suppose that infants should be so distracted by seeing an agent reappear after a brief departure that they invariably forget what happened previously. But even if this assumption were granted, why suppose that the agent's reappearance is not only distracting enough to affect the infants' memory of what they have just seen, but *precisely so distracting* as to make them forget the second time the toy moves but not the first—that is, to forget just the toy's last change of location? Only the assumption that it is precisely this distracting will get the required result. What's more, even if it were granted both that the agent's return has the effect of cancelling just the second move of the toy and that the toy's moving from Box-1 to Box-2 reduces the novelty of reaching into Box-2, there remains the problem that reducing the novelty of reaching into Box-2 is not the same thing as making it novel to reach into Box-1. Arguably, the most plausible thing to say is that neither reaching for Box-1 nor reaching for Box-2 is a novel action at this point, and that novelty is simply not a factor one way or the other. In that case, though, Heyes' account would fail to explain the results of this condition as well.

Heyes' deflationary account aims to explain the results of Onishi and Baillargeon (2005) solely in terms of such things as perceptual novelty, distraction, and memory failure. Though her deflationary account provides a reasonable potential explanation of one of the four conditions in this experiment, it fails for the remaining three conditions, either making incorrect predictions about the results in those conditions or relying on what appear to be ad hoc assumptions to achieve consistency with the results. By contrast, Onishi and Baillargeon's rationalist explanation, which takes infants to form expectations about an agent's behaviour based on her beliefs (both her true beliefs and her false beliefs), is simple and direct: In each condition, infants expect the agent to act on her belief about the location of the toy, and they look longer when the agent doesn't reach into the appropriate box (the box that she believes the toy to be in).

It doesn't bode well for Heyes' deflationary account that it isn't even able to explain the results of Onishi and Baillargeon's (2005) study, which initiated the use of non-traditional methods for determining whether infants have the ability to attribute false beliefs to others. Still, despite its problems explaining these core findings, there is a question about whether Heyes' deflationary explanation has any advantages over the rationalist explanation when we pan out and take into account further studies. Heyes' account claims to provide a general deflationary alternative for a broad range of findings that have been inspired by, and claim to support, Onishi and Baillargeon's view that the ability to attribute beliefs is in place early in development. As it turns out, consideration of additional studies only serves to further reduce the appeal of deflationary accounts like Heyes'. We will confine ourselves to a brief discussion of two of the other experiments we discussed earlier, which we take to raise representative difficulties for Heyes' approach to explaining these types of studies quite generally.

First, let's consider the study in which children as young as 18 months old spontaneously help an agent by pointing at the location of a hidden object that has been moved (Knudsen and Liszkowski 2012a). We saw earlier that whether the children pointed to help the agent locate the object as she approached the containers depended on whether the agent had a true or false belief about the location of the object. Anticipating that the agent will look in the wrong place when she has a false belief, the children spontaneously pointed to help the agent before she reached for any of the containers, providing further evidence in favour of the rationalist account that takes young children to represent and attribute true and false beliefs to agents. Heyes doesn't offer a deflationary explanation of this study (or of any of the other studies that make use of other experimental measures apart from looking time, such as spontaneous pointing behaviour).[18] And since the experimental measure in these studies doesn't involve looking times, Heyes' previous explanation (in which infants are supposed to look longer at a visually novel event) isn't applicable.

Nonetheless, these kinds of studies can still shed light on Heyes' deflationary account. In Heyes' deflationary explanation of the Onishi and Baillargeon (2005) study, she took the reappearance of the agent after a brief absence to cause the infants to forget that the object was moved. Since the agent in Knudsen and Liszkowski (2012a) also leaves the scene and then reappears, Heyes would presumably take the agent's reappearance in this study to have a comparable effect. There are two problems with this type of response, however. First, if we assume that the reappearance of the agent causes the children to forget that the object has been moved, this shouldn't hold just for the false-belief condition (in which the switch happens after the agent leaves the room). It should also hold for the

[18] Though Heyes (2014) doesn't offer deflationary accounts of such studies, she does briefly discuss them. She argues that while these sorts of studies (e.g., Buttelmann et al. 2009 and Southgate et al. 2010) provide convergent evidence in favour of the rationalist account, this evidence is balanced out by convergent evidence in favour of her deflationary account, leading to an evidential "tie". In particular, she cites evidence that she takes to support low-level deflationary accounts of studies of infants' understanding of objects (e.g., Diamond 1990) as providing convergent evidence for her deflationary account. We discuss work on infants' understanding of objects in later chapters, including both rationalist and deflationary accounts of this work (see particularly Chapters 10, 15, and 17). The important point for present purposes, however, is that the two types of convergent evidence that Heyes takes to lead to an evidential tie are not in fact comparable. The convergent evidence Heyes cites in favour of the deflationary account comes from a different conceptual domain which she thinks can also be explained in terms of low-level perceptual variables and things like a novelty preference. In contrast, the convergent evidence that she grants is supportive of the rationalist account comes from the very same domain at issue—it just employs different (and complementary) experimental measures to support the claim that infants and young children can attribute false beliefs. What this means is that Heyes' deflationary account simply fails to provide an explanation of the same breadth of studies directly bearing on true- and false-belief attribution as the rationalist account. Surprisingly, Heyes goes on to suggest that the way to break the "tie" that she sees between the deflationary and rationalist theories of the evidence regarding infant belief attribution is to "use new experimental strategies in an attempt to break the tie" (2014, p. 655)—ignoring the fact that this is precisely what is provided by the studies that give convergent evidence for the *rationalist* account using non-looking-time measures.

true-belief condition too (in which it happens before the agent leaves). Either way, the agent's return should cause the children to forget that the object had been moved. But the children treated these cases very differently. They were far more likely to point to the container that actually held the object when the agent wasn't in a position to know that it had been moved. The second problem is that when children point to help the agent (i.e., in the false-belief condition), they point to the container where the object is actually located. Notice, however, that they wouldn't be able to do this if they had forgotten that the object had moved. So, the data, in this case, directly contradicts the assumption that the reappearance of an agent is so distracting as to make children forget that the object has moved. This calls into question Heyes' claims regarding the effects of the reappearance of the agent in Onishi and Baillargeon's study, which her explanations depend on.[19]

Finally, let's consider the study in which 2.5-year-old children hear a false-belief story and spontaneously look at its matching picture (Scott et al. 2012). This study does turn on a measure of children's looking times, so in principle a novelty preference explanation might be applicable. But several considerations argue against this interpretation. Unlike in the Onishi and Baillargeon (2005) study, there were no familiarization trials in which the agent repeatedly interacted with one of the two locations before the test trials. There was also no difference between the true-belief and the false-belief conditions regarding the relative novelty of the agent searching in one or the other of the two locations in this study.[20] So it is difficult to see how an explanation in terms of a novelty preference could even get off the ground. However, there is an even more important reason for rejecting a deflationary explanation of the results of this experiment in terms of low-level perceptual properties. Children's looking times are monitored throughout the experiment, and the measurement of looking times *prior to* the test phase in the experiment provides strong evidence for the interpretation of looking times *during* the test phase. In particular, the children's looking times prior to the test phase supports the interpretation that in this experiment they look longer at the

[19] We should emphasize that our claim here is not that it is impossible to produce a deflationary account that is consistent with this data. Heyes could, for example, try to claim that the infants in Knudsen and Liszkowski (2012a) are not subject to this effect, but those in Onishi and Baillargeon (2005) are, because the infants in Knudsen and Liszkowski's study were on average three months older than those in Onishi and Baillargeon's study. Our claim is only that this study provides some positive reason to question Heyes' assumption about the effect on infant memory when an agent reappears after a brief absence. Heyes simply asserts that the agent's reappearing has this effect based on the general possibility that memories (in non-human animals, human infants, and adults) can be affected by later events. But now that we have seen that this particular type of event doesn't have this detrimental effect in a closely related study, this just underscores the burden on Heyes to provide evidence that the reappearance of agents has this effect on 15-month-olds.

[20] And since the stimuli were static pictures accompanied by a story, the infants do not actually see the agent leave or reappear, so there is no question of an effect of the agent's reappearance on the children's memories of the events.

matching picture—and the matching picture in the test phase indicates that they take the agent's behaviour to be guided by a false belief. This means that any alternative deflationary explanation of these results not only will have to provide an alternative account of the looking times during the test phase of the experiment, but also will need to provide an alternative account of the looking times throughout the rest of the experiment.

To see this, it may help to work through the experimental setup one more time. Recall that the children hear a story in which Emily puts an apple in a box, and Sarah subsequently moves it to another location while Emily is asleep (false-belief condition) or while Emily is still watching (true-belief condition). After the initial introduction of the two characters in the story (Emily and Sarah), each line in the story is followed by the simultaneous presentation of two pictures only one of which matches the story line. For example, when they are told "Look! Emily is putting her apple in a box" early in the story, this is followed by the simultaneous presentation of a picture of Emily playing with some toys and a picture of Emily putting an apple in a box. Like adults, the children look briefly at both pictures but end up looking longer at the picture that matches what they have just heard—in this case, the picture of Emily putting the apple in the box. This pattern continues throughout the course of the story prior to its final line. So, what this tells us is that we can use the picture they look at the most at the end of the story (when they hear "Emily is looking for her apple") to determine where they think Emily will look for the apple. As we noted above, the children take the matching picture to be Emily searching in the apple's original location in the false-belief condition (the condition in which Emily didn't see the apple being moved), but take the matching picture to be Emily searching in the new location in the true-belief condition (the condition in which Emily did see it being moved). This experiment involves essentially the same basic scenario as the one in Onishi and Baillargeon (2005). But the interpretation of the children's looking times in the test phase (the final story line) is supported by the evidence from looking times following all the earlier story lines, ruling out a deflationary account of the experimental results and providing strong convergent evidence in support of the rationalist interpretation of Onishi and Baillargeon (2005).

In sum, consideration of a wider range of non-traditional false-belief task studies shows that Heyes' deflationary account doesn't provide any explanation of a number of studies that use different experimental measures or different methodologies to provide converging evidence for the rationalist account. Of course, Heyes could try to find other deflationary explanations of these studies. And since *something* is always different when comparing a true-belief condition with a false-belief condition, she could try to claim that this difference explains the contrasting results in these two conditions. But, as we saw with Heyes' explanations for the different conditions in the Onishi and Baillargeon (2005) study, merely pointing to a difference between the conditions isn't enough to actually explain

the data. A case has to be made that the difference in question explains the differing results in the true- and false-belief conditions in a way that isn't ad hoc—a point that is further underlined by the fact that studies like Knudsen and Liszkowski (2012a) directly call into question assumptions made by Heyes' deflationary explanation. To pose a challenge to the rationalist interpretation of the data, a deflationary account needs to provide an equally compelling account of an equally broad range of experimental results from non-traditional false-belief task studies and cannot simply turn on ad hoc assumptions. Heyes' (2014) deflationary account—one of the most ambitious and highly regarded deflationary accounts available—fails to do this.

More could be said about deflationary accounts, but we will leave it here so that we can examine dual-system accounts. Earlier we noted that dual-system accounts have the feature that, although they credit young children with representational abilities that aren't confined to low-level perceptual properties, these are still supposed to fall short of representations of belief-like states. One type of dual-system account—Perner and Ruffman's behavioural-rules approach—has been extensively discussed elsewhere (e.g., Song et al. 2008; Carruthers 2013; Scott 2014). The general problem with this approach is that the varied situations involved in non-traditional false-belief tasks show that infants and toddlers are able to respond in comparable ways to false-belief situations that are very dissimilar physically and behaviourally. Rationalist models that credit infants with the ability to represent false beliefs have no difficulty in accounting for this fact, but the behavioural-rules approach ends up having to maintain that infants learn numerous specific behavioural rules that are completely independent of one another. The result is an unrealistic learning model with a complex assortment of behavioural rules that really only hang together because they are chosen in a post hoc manner to mimic the behavioural consequences of understanding that an agent can act on the basis of a false belief. In any case, the behavioural-rules approach is now widely seen as discredited. For this reason, we will focus instead on Apperly and Butterfill's dual-system approach, which is far more sophisticated than the behavioural-rules approach and, in our view, a much more interesting proposal.

Apperly and Butterfill (2009; Butterfill and Apperly 2013) take the results of the non-traditional false-belief studies to show that infants possess an early-developing system for understanding and predicting actions that they call a *minimal theory of mind*. This system operates automatically in a rapid and efficient manner and independently of general cognitive resources, such as working memory. It is also taken to continue to operate through adulthood unaffected by later-developing and more sophisticated mentalizing abilities that are grounded in a second (not minimal) system.

A key difference between this early-developing system and later-developing mentalizing is that the early-developing system doesn't represent beliefs and

desires as such. Instead, it makes use of representational resources that represent different kinds of states—ones that are not part of our common-sense understanding of how minds work but that are often correlated with perceiving, believing, and so on. By representing states that are closely correlated with such more familiar mental states as seeing or believing, these representations can track these states and serve as proxies for representations of them. At the same time, they are meant to be less conceptually sophisticated and less computationally burdensome. One type of representational resource that the minimal theory of mind takes infants to have—*encountering*—approximates perceiving. Apperly and Butterfill characterize it as a relation between an agent, and object, and a location, where (roughly speaking) the object is in the agent's proximity and the situation meets a number of other constraints (there is sufficient lighting, there are no intervening opaque barriers, and so on). Encountering will correlate well with perceiving, since objects that are near agents, in good lighting, and so on will typically be perceived. But at the same time, representing encountering doesn't require young children to attribute to others a mental state that involves representing an object using a mode of presentation. Another type of representational resource that the minimal theory of mind takes infants to have—*registration*—approximates believing, where "one stands in the registering relation to an object and location if one encountered it at that location and if one has not since encountered it somewhere else" (Apperly and Butterfill 2009, p. 962). Registering will correlate well with believing, since if an agent encounters a particular object in a particular location (that is, if the object is near the agent, they are facing it, there is good lighting, etc.), the agent will typically form the belief that the object is in that location. But at the same time, representing that an agent has a registration doesn't require young children to take the agent to be representing the object using a particular mode of presentation. Registrations are further understood to determine where an agent will search for an object, and, like beliefs, can fail to correspond to how the world currently is (they have "correctness conditions"). As a result, an infant who represents others' minds in terms of registrations can predict where an agent will look in a false-belief situation by attributing to the agent an incorrect registration.[21]

[21] How are registrations different from beliefs? One difference is that the "theory" about how registrations work and interact with other states that an agent is taken to have is simpler and less computationally burdensome for infants. A second difference, which we have noted in the text, is that registrations aren't supposed to capture the particular way mental states pick out what they represent (i.e., they don't involve attributing a mode of presentation to the agent taken to have a registration) and aren't supposed to be able to represent quantified propositions. For example, since registration, like encountering, involves a direct relation between an agent and an object, it doesn't take into account the particular way of conceptualizing that object. All the registration encodes is the fact that the agent does in fact stand in the relation to the object, not how the agent thinks about the object. So if someone only has a minimal theory of mind available for attributing mental states to others, she wouldn't be able to differentiate between a person registering the morning star being in a particular location and that person registering the evening star being in that location (since the morning star is

To see how this sort of minimal theory of mind is supposed to work, consider *false-belief condition 1* again, from Onishi and Baillargeon (2005) (discussed above). On Apperly and Butterfill's account, we needn't suppose that 15-month-olds represent the agent's false belief. Rather, infants can take the agent to have formed a registration of the toy as being located in Box-1. And while the infants know this registration is no longer correct, they can nonetheless expect it to guide the agent's behaviour. All of this leads them to expect the agent to search for the toy in Box-1, and so they are surprised if the agent searches for the toy in Box-2. This provides a very nice explanation for this experiment and provides equally strong explanations for all the conditions in Onishi and Baillargeon's (2005) study. Apperly and Butterfill's minimal theory of mind account can explain many other experimental results in non-traditional false-belief tasks in a similar manner, so there is much to be said in favour of their account.

Nonetheless, there are a number of findings that it *cannot* explain and that reveal its limitations. One of these comes from a study of the ability of 18-month-olds to reason about an agent's false belief about the identity of an object (Scott and Baillargeon 2009). The children in this study see an agent interact with two identical looking toy penguins, one of which is divisible into two parts (a top and bottom half) and the other of which isn't. The divisible toy is initially seen separated into its two components. The agent puts a key into its bottom half and assembles the toy so that the two toy penguins look exactly the same. After seeing this sequence repeated in a variety of contexts (and so becoming familiar with the idea that the divisible toy starts out in its divided state and with the fact that the agent likes to store her key inside the divisible penguin), the children see a second agent interact with the toys while the first agent is absent. The second agent assembles the divisible toy, puts it under a transparent cover, and puts the indivisible toy under an opaque cover. (At this point, anyone looking at the scene would see a transparent cover and an opaque cover side by side, with an intact toy penguin visible through the transparent cover.) Finally, what happens next is that the first agent returns with the key and reaches for one of the two covers. Notice that since the toys were always encountered with the divisible one initially in its divided state and since the toy under the transparent cover isn't divided—it looks just like the indivisible toy—it would be natural for the agent to *mistakenly believe* that the divisible toy is under the opaque cover. Eighteen-month-olds seem to interpret the situation in just this way. They look longer when the agent reaches for the transparent container.

However, it is hard to see how Apperly and Butterfill's dual-system account can explain this outcome since it doesn't grant that children who are this young have the ability to represent false beliefs. According to Apperly and Butterfill, all

the evening star). The feature of representing objects in a particular way, or under a particular mode of presentation, is a distinctive feature of beliefs and other propositional attitudes.

18-month-olds know about are outdated registrations. The problem is that, in this case, registrations fail to track beliefs (Scott et al. 2015). The children can clearly see that the divisible toy (in its assembled state) is *encountered* by the agent under the transparent cover. After all, they can easily see that the divisible toy is near the agent, that the agent is facing it, that there is good lighting, and so on, so they should take the agent to have encountered the divisible toy as being there. And having encountered it there, the agent should form a new registration that links the toy to that very location—so the children should take the agent to have formed a registration of the divisible toy being there. But if the children attribute this registration to the agent, then they shouldn't be surprised when the agent grasps the transparent cover instead of the opaque cover. In fact, they should expect the agent to behave in precisely this manner.[22]

Another finding that poses a challenge to Apperly and Butterfill's dual-system account turns on the ability of 17-month-olds to represent a deceptive intention in which a thief sets out to induce a false belief in the victim of their theft (Scott et al. 2015). First, the children see a scene in which one character (the thief) is present and another (the victim) enters holding a toy. Sometimes the toy is one that rattles when shaken and other times it is one that doesn't. The victim shakes the toy, and it either rattles or remains silent. Then a bell rings, and the victim places the toy on a table between her and the thief, announces she will be back soon, and leaves the room. Shortly after, the victim returns and moves the toy. If it previously rattled, she places it in her toy box (showing that she values it); if it previously remained silent, she places it in a rubbish bin (showing that she doesn't value it). The test trials that follow proceed in much the same way, except that right after the victim leaves the scene, if the toy on the table is a rattling toy, the thief steals the toy, concealing it in her pocket and putting a silent toy in its place on the table. Crucially, on some trials the replacement toy looks identical to the stolen toy (making the substitution impossible to detect visually), and on other trials the replacement toy looks noticeably different than the stolen toy (making it likely that the victim would recognize the substitution). Seventeen-month-olds evidentially recognize the significance of this difference. They look

[22] Butterfill and Apperly (2013) respond that the children might have a different expectation. Perhaps they reason that since there is always both a divisible and an indivisible toy present and since the toy under the transparent cover appears to be the indivisible one, the agent assumes that the divisible one is under the opaque container. However, this response faces a number of difficulties. One is that the reasoning here depends on attributing representations of quantified propositions to the agent—the semantics for the indefinite pronouns *a* and *an* is quantificational—despite the fact that the minimal theory says that infants at this stage can't represent quantified propositions. But perhaps the biggest problem is that this explanation simply ignores Apperly and Butterfill's own account of when a given registration should be attributed to an agent. Their account claims that "one stands in the registering relation to an object and location if one encountered it at that location and if one has not since encountered it somewhere else" (p. 962). Given this definition, the children should attribute to the agent the registration *Object O [= the divisible toy] is under transparent cover*, since the agent is manifestly encountering the divisible toy in this location (albeit in its assembled state).

longer when the substituted toy's colour and pattern fail to match that of the stolen toy, indicating that they expect the thief to substitute a matching toy. Further tests show that infants also look longer if the victim throws an identical looking replacement toy in the rubbish bin—indicating that they are surprised that the victim would discard what appears to be the desirable rattling toy they left on the table. Infants also look longer if the victim puts a different looking replacement toy in her toy box—indicating that they are surprised that the victim would treat a noticeably different toy as if it were the desirable rattling toy they left on the table. By contrast, infants aren't surprised if the victim throws a different looking replacement toy in the rubbish bin or puts an identical looking replacement toy in her toy box. All of this suggests that young children expect the victim to be deceived only if the thief substitutes a visually indistinguishable toy for the stolen one, and that they expect the thief to anticipate this and to act accordingly.

Can Apperly and Butterfill's dual-system account explain these results? Like Scott et al., we don't see how. The problem once again is that the children ought to suppose that an agent will update her registration of the location of an object when they see this agent encounter it in a new location. Before leaving the room, the victim last encountered the silent toy in the rubbish bin, but upon her return, she manifestly encounters it on the table. (The children know this is the silent toy because they themselves witnessed the substitution.) Consequently the children should attribute to the victim the new registration relating the silent toy to its new location (the table), and shouldn't be at all surprised if the victim discards it even if its colour and pattern match the rattling toy that had previously been left on the table. But again, they *are* surprised—they look longer when the victim discards this toy. Evidently they understand that the victim mistakenly *believes* it is the rattling toy they had left there and that the thief successfully led the victim to have this false belief—in which case, they must have more than a minimal theory of mind.

In sum, like deflationary accounts, dual-system accounts fail to provide a successful alternative to rationalist mentalizing accounts of young children's success at non-traditional false-belief tasks.[23] The evidence does not point to an

[23] Another argument against the minimal theory of mind is that there is independent evidence that children's difficulty with *traditional* false-belief tasks stems from performance factors, not their conceptual competence. By simplifying the task demands and reducing the burden on various performance factors, researchers have shown that children as young as 2.5 years old can reliably pass a traditional false-belief task (Setoh et al. 2016). In addition, the sorts of performance factors involved are also known to affect *adult* performance on false-belief tasks. For example, if adults are asked to perform a simple memory task at the same time as a non-traditional false-belief task, their performance on the false-belief task severely declines (Schneider et al. 2012). And adults and infants do worse on non-traditional spontaneous false-belief tasks when they know the moved object's actual location (and consequently have to inhibit this knowledge) compared to when they don't (Wang and Leslie 2016).

early-developing system for representing such states as registrations. It points to a system for representing *beliefs*—true and false beliefs—that is already operating at 15 months of age and perhaps much earlier.[24]

Let's now consider how all of this contributes to an argument from early development in this domain. Earlier we noted that the argument from early development doesn't require that the representational capacities at issue be present at birth, and that there are reasons why certain innate representations and representational systems won't be evident until later in development. These include performance factors, biological maturation, and learning (i.e., rationalist learning). In the case of FALSE BELIEF, there can be no question that performance factors are an issue. We have seen that reducing these to a minimum—especially through the shift from traditional false-belief tasks to non-traditional false-belief tasks—has revealed that infants represent far more about other people's minds than researchers previously thought possible.

Of course, there is always the possibility that the infants who pass spontaneous-response false-belief tasks don't really have the ability to represent false beliefs after all and that the data are spurious or misleading. Maybe they won't hold up to replication,[25] or maybe there is a stimulus confound with a lower-level property we haven't addressed (e.g., some not yet imagined alternative in line with, but not the same as, Heyes' suggestion about a preference for novel direction of motion). While these alternative possibilities may not have been categorically ruled out, there is no reason to suppose that they can provide a robust alternative to richer, mentalistic hypotheses; no deflationary account comes close to explaining this data as well as a mentalistic account does. And infants show a consistent pattern of success on false belief tasks involving strikingly different stimulus

[24] We should also note that even if Apperly and Butterfill's dual-system account can be made to work, it doesn't offer much hope for a broadly *empiricist* approach to explaining the origins of FALSE BELIEF. Their early-developing system (with its representation of encounterings and registrations) is a domain-specific system par excellence. It would most likely have to be an *innate* domain-specific system, too, to accommodate the data from the spontaneous-response false-belief studies reviewed above. And, assuming the dual-system account, the second more flexible system would need to be available by the age of 2.5 years old, since as we have seen, children can pass traditional false-belief tasks by this age (Setoh et al. 2016). If it is required to explain results such as those in the toy penguins study (Scott and Baillargeon 2009) and the thief study (Scott et al. 2015) to meet the objections above, it must be in place even earlier—by 17 months at the latest. Arguably, then, the best way to maintain a dual-system account would be to adopt an even more rationalist account and claim that *both* of the postulated systems for representing and reasoning about others' actions are innate.

[25] It should be noted that there have been some failures to replicate the results of some non-traditional false-belief tasks with infants alongside other successful replications; see Paulus and Sabbagh (2018). It is important to take such failures of replication seriously and consider each case in detail. But it is also important to remember that there are many reasons why an attempt at replication can fail to reproduce the original result. Work with infants is especially likely to be sensitive to small variations in methodology—variations that can affect whether an experiment is able to expose infants' true representational abilities. For discussion of some of these issues in relation to the "replication failures" in Paulus and Sabbagh (2018), see Baillargeon et al. (2018). See also Chapter 18 for a related example of a case where the charge of a failed replication attempt clearly turns on the experimenters' not attending to crucial details of the original stimuli.

materials and methods of evaluation—violation of expectation, preferential looking for story-picture matches, spontaneous helping, and so on.

Given that these abilities are not manifestly present at birth, though, how does this work support an argument from early development? Can we still say they appear *too early* to reliably be acquired solely by domain-general learning mechanisms? In fact, we can. This is because, as empiricists themselves see it, the type of information that such a mechanism would require is typically thought to be linguistically mediated—in which case, children would have to have a well-developed linguistic competence before they would be in a position to develop an understanding of false beliefs. For example, one standard type of proposal has been that the representation of false beliefs builds on prior linguistic abilities, such as the mastery of grammatical constructions that allow for the representation of multiple perspectives (e.g., de Villiers and de Villiers 2000, 2009), which is thought to come online around the time that children reliably pass traditional false-belief tasks—considerably later than they show understanding of false belief in spontaneous response tasks. As things stand, then, the weight of evidence strongly suggests that infants have an understanding of false beliefs at an age that is too early for a domain-general learning mechanism to reliably acquire it, since such a mechanism wouldn't have access to the information it requires. In contrast, rationalist accounts fit will with the early development of mentalizing and false-belief understanding and explain this in terms of innate psychological structures that are specific to this domain, all without having to claim that these abilities should be evident at birth.

Our examination of this case study has required an extended discussion. But as we noted earlier, the origins of the concept FALSE BELIEF has been at the very heart of the rationalism-empiricism debate regarding the origins of concepts and is one of the most thoroughly investigated examples of an early-developing concept, and so is a particularly important example to consider. FALSE BELIEF has also been widely thought to be especially challenging for rationalists given how sophisticated an ability it is to represent not just mental states but ones that are at odds with the way the child herself takes the world to be.

The data that have informed debates about the origins of FALSE BELIEF also point to a rationalist account of the representation of perceptions, goals, preferences, and intentions (among other mental states). If this isn't obvious, consider the reasoning that must be attributed to children to explain the results in the study in which infants saw the thief set out to steal the rattling toy from his victim. As Scott et al. (2015) note, infants have to appreciate not just that the agent acts on a false belief. They must understand that:

> (a) T [the thief] had a *preference* for the rattling toys; (b) when O [the victim] introduced the rattling test toy, which was visually identical to a previously discarded silent toy, T formed the *goal* of secretly stealing the rattling test toy;

> (c) substituting the matching silent toy was consistent with T's deceptive goal, because O would hold a *false belief about the identity* of the substitute object; and (d) substituting the non-matching silent toy was inconsistent with T's *deceptive goal*, because O would *know* which toy it was as soon as she saw it. (p. 41; italics highlighting mental state attributions in original)

All of this makes FALSE BELIEF a particularly good case to consider in relation to the argument from early development for traits that are not present at birth. But, it is important to note that this type of argument from early development is widely applicable to many other domains. We will close out our discussion of the argument from early development by briefly highlighting some of the findings supporting an argument from early development in another domain broadly connected to social cognition, namely communication. We have already mentioned some of the evidence that infants have early representational abilities for understanding certain behaviour as involving communication. In Chapter 8, we noted that 6-month-old infants treat eye gaze as communicative, selectively following eye gaze when it is preceded by direct eye contact or infant-directed speech (Senju and Csibra 2008). This suggests that well before infants have uttered a single word—indeed, just as they are beginning to babble—they already have some understanding of communication and take eye gaze and infant-directed speech to be communicative.

The view that infants understand the communicative role of language is supported by numerous other findings as well. Consider infants' preference for speech. We noted earlier that newborns prefer to listen to the language of their community compared to a foreign language, a preference that is established by prenatal learning. They also have a generalized preference for language over many other types of non-linguistic sounds. This preference is initially broad enough to include non-human primate vocalizations (rhesus macaques)—which infants have *not* heard before—and is refined in the first three months of life to a general preference for human language (Vouloumanos et al. 2010). By the time infants are 3 months old, they prefer to hear words in a human language—any human language—compared to a variety of other sounds, including non-human primate vocalizations, natural environmental sounds (running water, bells, wind), and various familiar non-linguistic human sounds (e.g., laughter and vocalizations associated with agreement, inquiry, and surprise) (Shultz and Vouloumanos 2010).

Infants don't just have a preference for speech. Speech also guides a great deal of infant learning in a domain-specific manner. For example, 4-month-old infants orient to visual objects in the direction of eye gaze more quickly when the eye gaze is preceded by speech rather than silence or non-speech sounds (backward speech) (Marno et al. 2015). Other work argues for an important role for speech in facilitating categorization in infants as young as 3 months of age (Ferry et al. 2010).

Notice that this effect of speech on categorization at this early age is quite surprising on the empiricist assumption that children initially take speech sounds to be no different than any other auditory stimuli. But at as young as 3 months old, infants are already responding to *visual* stimuli differently according to whether these are accompanied by speech. In this work, infants are shown a number of exemplars of a category (e.g., visual images of different dinosaurs) while either hearing speech or comparable non-linguistic sounds. The infants are then shown two objects, one from the same category (another dinosaur) and one from a novel category (e.g., a fish). Infants who hear non-linguistic sounds (tones) don't discriminate between the two (showing no evidence of having formed a category corresponding to the objects that accompanied the tones), while infants who hear speech prefer to look at the novel instance of the previously seen category (suggesting that the linguistic sounds accompanying the visual stimuli caused the infants to treat these objects as forming a category).[26] The key point is that the category learning here seems to be domain specific in the way that it is facilitated by a particular type of stimulus condition. Such category learning isn't confined to representations in a single domain (e.g., animals), but it is initiated by a specific type of cue—sounds that are represented as *speech* or *language*—suggesting that infants are already employing abstract representations such as SPEECH or LANGUAGE at 3 months of age and that the connection between the systems involved in categorization and the attribution of speech have a fair degree of innate articulation (in the sense explained in Chapter 2).

Categorization prompted by speech would naturally be advantageous when it comes to word learning. And, surprising as it may be, it turns out that by the time infants are 6 months old—long before they have uttered a single word—they already associate a substantial number of common nouns with their correct

[26] At this age, the learning is facilitated both by human speech and by non-human primate (in this case, lemur) vocalizations (Ferry et al. 2013). But by the time infants are 6 months old, non-human primate vocalizations no longer facilitate categorization unless infants have been exposed to these vocalizations in the past (even very brief exposure will do) (Perszyk and Waxman 2016). In contrast, exposure to non-linguistic sounds, such as backward speech or birdsong, fails to facilitate categorization (Woodruff Carr et al. 2021). This pattern of results underscores the fact that the effect is not driven by familiarity, that infants' categorization is prompted in a domain-specific manner by what is taken to be linguistic communication, and that the innate facilitating category undergoes further specification in early development. This conclusion is further supported by work showing that categorization is also facilitated by non-vocal language in the form of signing in American Sign Language and that this facilitation occurs even in infants who have not otherwise been exposed to sign language (Novack et al. 2021). Interestingly, there is one way in which tones can come to facilitate categorization. This is if they are presented in a communicative context that suggests that the tones are part of a type of communication system. Ferguson and Waxman (2016) presented 6-month-old infants with a scene in which two women were engaged in a cooperative activity, supported by joint attention, in which they took turns in a "conversation" where one spoke in English and the other spoke in "tones" (the tones were in sync with the woman's mouth movements). A different group of 6-month-olds saw much the same scene but the women were silent and the same "conversation" could be heard as background noise. In the first condition, the infants were able to subsequently use tones to facilitate categorization, but they weren't in the second condition.

referents. Given the choice between pairs of photos or videos, they selectively look more at the photo or video that matches a heard word (e.g., they selectively look at the corresponding referents for words such as "mouth", "hand", "apple", and "spoon") (Bergelson and Swingley 2012, 2015). Of course, associating a word's sound with a category is one thing, and knowing its meaning is another. But if this learning is guided by domain-specific processes (as seems likely given the evidence we have been reviewing), then even if the learning takes the form of an association, it is not a purely domain-general association. Rather, it is one that requires a specific type of input (much as we saw in the case of animal food aversions in Chapter 4).

Moreover, there is evidence that infants as young as 6 months old even grasp the *communicative function* of speech (Vouloumanos et al. 2014). To show this, researchers familiarized 6-month-olds with two agents, the Communicator and the Recipient, each interacting with a pair of objects. The infants saw the Communicator repeatedly grasp one of the two objects when presented with both, and they separately saw the Recipient interacting with both of the objects. Then, in the test phase, the Communicator (who was no longer in a position to reach for the objects) turned to the Recipient (who had access to the objects) and either spoke a novel word ("koba") or made a coughing sound. For adults, the Communicator's repeated attention to one of the two objects (given a choice of both) signals a preference for that object, and adults naturally interpret the speech, but not the cough sound, as a communicative vocalization. So adults expect the Recipient to hand the preferred object to the Communicator in the speech condition but don't have this expectation in the non-speech condition. Apparently 6-month-olds see the situation in the same way. They looked equally whether the Recipient handed over the preferred object or the non-preferred object in the cough condition, but looked longer when the non-preferred object was offered in the speech condition. Notice that the difference between the two test events isn't one that infants can appreciate by just attending to what they themselves know about the Communicator—in both conditions they presumably attribute to the Communicator a preference for one of the two objects (another example of early mental state attribution). Infants must further understand that the speech sounds in this situation are communicative whereas the cough sounds are not. This is essential for them to form the expectation that the Communicator's vocalizations would influence the Recipient's understanding of the situation and that the Communicator's (non-speech) vocalizations would not. Thus 6-month-old infants seem to appreciate that speech can be used to convey information between people even though they don't understand the speech (the novel word "koba") and even though they have no experience of conveying information through speech themselves.

Further evidence suggests that the language processing areas in the brain are wired at birth to respond specifically to communicative uses of human language

(Forgács et al. 2022). In this work, newborn infants heard voices uttering various pseudo-words. In the communicative condition, two voices took turns speaking different words each of which had some internal syllabic repetition (an indication that the utterance exhibits a structure that may follow a rule and hence might be language). For example, one said *mu-fe-fe* followed by the other responding with *pe-na-na*. The researchers found that in this condition there was increased activity in known language processing areas (left fronto-temporal areas) relative to two control conditions. One control condition was a *non-social* control condition in which the same type of sequence was heard as in the communicative condition but from only one speaker. The second control condition was a social but *non-communicative* condition in which there were two voices as in the communicative condition, but the second voice merely echoed the word produced by the first, reducing the likelihood that information was being transmitted. Only the communicative condition showed elevated activity levels in known language processing areas, suggesting that even newborns are selectively sensitive to indicators of linguistic communication involving turn taking and variability regarding the spoken words and are already disposed to activate known language processing areas in response to these conditions.[27]

In sum, there is a great deal of evidence for the early emergence of concepts and related representational capacities pertaining to language, communication, and a wide range of mental states—evidence that is best explained by an approach that postulates a combination of innate concepts and rationalist learning mechanisms in accordance with concept nativism's vision of the acquisition base. We'll see later that similar evidence exists in other domains broadly connected with social cognition concerning the representation of such things as norms and social groups (see especially Chapters 18 and 21). More generally though, the argument from early development, as extended to cover representational abilities that aren't present at birth but only appear later in development, provides a strong argument for concept nativism across a wide range of further domains (many of which we touch on in later chapters), including ones pertaining to such things as physical objects, groups or collections, events, animals, plants, death, elementary logical concepts, function/purpose, and possibility and necessity.

What counts as *too early* in relation to the argument from early development will naturally vary depending on the particular type of representational ability; it needn't turn specifically on having already come to possess a developed linguistic competence. Highly abstract representational abilities—ones that are far removed from the low-level sensorimotor representations that are available as part of an empiricist acquisition base—will often require learners to first come to possess a range of other cognitive abilities and more abstract concepts and representations

[27] See also Cho et al. (2021) for evidence that 6-month-olds recognize that a person intends to communicate something about an object when she points towards the object while uttering a novel word, but not when making a novel emotional vocalization.

in order to be acquired via a domain-general learning mechanism. Accordingly, possessing the representational ability in advance of any of these traits can form the basis for a forceful argument from early development. We will also see that the argument from early development often forms part of a larger, more powerful argument to the best explanation for a rationalist account by working in conjunction with some of the other arguments for concept nativism we highlight. In later chapters, we will illustrate this using the concepts from a number of domains, including concepts for mental states.

We conclude that the argument from early development offers compelling support for concept nativism. While the argument may be particularly powerful when there is good evidence for a representational capacity being present at birth (Chapter 8), it can also be highly effective in arguing for concept nativism in cases where the representational capacities only appear later in development. But the case for concept nativism doesn't stand or fall with the argument from early development. This is only the first of our seven mutually supporting arguments for concept nativism. So let's turn to our second argument.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0009

10
The Argument from Animals

Our second argument for concept nativism is a development of an argument that came up earlier—the *argument from animals*. In Chapter 4, we presented the argument from animals as an argument for rationalism with respect to cognitive development in general. The form the argument took there was to highlight the fact that human beings are animals. Since rationalist learning systems are widespread in the animal kingdom, we concluded that it is very likely that humans also possess rationalist learning systems—some that are shared with other animals and some that are distinctive of our species. Although we didn't draw out this point earlier, notice that this argument is as much an argument for *concept nativism* as it is for rationalism in general. This is because many of the rationalist systems it covers are likely to play an important role in concept acquisition (systems involved in navigation, foraging, predator avoidance, etc.), making the learning mechanisms for these concepts characteristically rationalist. In this chapter, we will focus on two additional arguments that are rooted in the fact that animals offer distinctive advantages in arguing for the existence of rationalist learning systems. We will argue that establishing the existence of rationalist learning systems in these ways *in animals* strengthens the case for the existence of comparable rationalist learning systems in *humans*.[1]

The first of these advantages focuses on the *input* available to animal learners—in particular, on the fact that it is possible to study the acquisition of representational abilities in animals under highly constrained input conditions. With animals it is at least sometimes possible to arrange their environments in such a way that experimenters have virtually total control over the experiences animals have prior to testing—in the limit, ensuring they have *no* relevant experience in advance of an experiment. Obviously if animals can still be shown to have a given representational ability under *these* conditions, it can't derive from general-purpose learning. By contrast, for both practical and ethical reasons, such control

[1] From now on, we will use the expression *the argument from animals* to cover the whole family of considerations that locate evidence for concept nativism in facts about non-human animals, including variations on the argument from Chapter 4, the two arguments from this chapter, and other related arguments. This inclusive terminology is meant to emphasize that these considerations aren't mutually exclusive and certainly aren't exhaustive and that the study of non-human animals offers a rich and varied source of evidence for concept nativism. In addition, following common usage, we will also often use the term *animal* to mean non-human animals. The expression *non-human animals* is somewhat cumbersome, and context should make it clear enough when a contrast with human beings is intended.

over environmental input is not possible with humans. As a result, it is often more difficult to investigate how a given representational ability is acquired by humans than by animals.

The second advantage that animals offer for establishing the existence of rationalist learning systems focuses on the *cognitive mechanisms* that animals use to acquire a given representational ability, as opposed to the input to such mechanisms. In some cases, rationalists and empiricists agree that a given type of representational ability cannot be acquired by simple general-purpose learning. In such cases, empiricists propose to explain acquisition of the traits in humans through learning that is mediated by linguistic abilities or powerful general-purpose learning systems. But while many animals have communication systems, no such systems have been shown to possess anything like the expressive power of human natural languages, and many types of animals are widely seen as lacking the sorts of powerful non-linguistic general-purpose learning systems that empiricists posit in the human case. So, if it can be shown that such animals can nonetheless acquire the representational abilities in question, this strongly suggests that the acquisition of these abilities does not require such powerful general-purpose learning systems and that the learning (for these animals) is owing to a rationalist learning mechanism.

Each of these two further types of the argument from animals—the input-based argument and the mechanisms-based argument—provides a type of argument for the existence of rationalist learning mechanisms that is more difficult to make in the case of humans than in the case of animals. Where there is a question about whether or not a given representational ability is innate or acquired via rationalist learning mechanisms in human beings, but where it can be shown that the same (or a highly similar) ability is innate or acquired via rationalist learning mechanisms in some animals, then this increases the likelihood that this representational ability is innate or acquired via rationalist learning mechanisms in humans too. At the very least, we would have a *proof of possibility* that evolution is capable of producing innate representations or special-purpose learning mechanisms of this type. Moreover, by taking into account the pattern of presence or absence of the innate representational ability among different species—how widespread the innate representational ability is among different species, which species possess the representational ability, and how closely related to homo sapiens these species are—we would also be in a position to draw reasonable, though of course defeasible, inferences regarding the human acquisition base.[2] For example, if a representational system were shown to be innate across a diverse range of genealogically

[2] Our focus in this chapter is on the human acquisition base. But the rationalism-empiricism debate regarding the origins of concepts and mental representations more generally also arises for other species, and the evidence that is highlighted by these two further types of the argument from animals can clearly also be used to draw inferences about the acquisition bases of other species.

related primate species, including those most closely related to humans, this would certainly lend support to the proposal that it is innate in humans too.

Let's take a closer look at these two further types of the argument from animals. We will start by looking at the version based on input considerations.

Though it is not always possible to control the environmental input that animals receive, in at least some cases it is possible to put in place remarkably stringent restrictions on such input, and, in the limit, to control all of the relevant experiences that an animal might have before being tested for a given representational capacity. In our view, the paradigm for this type of research is a recent body of work with newborn domestic chickens. Baby chickens can be hatched and raised in total darkness so that their *very first visual experiences* are those involved in the experiment. But, in addition, chicks are invaluable research subjects because they are *precocial*: unlike human babies, they are up and about from birth and can express their interests and preferences not only by what they look at but far more dramatically through what they approach and avoid when presented with different options. (By comparison, human infants typically can't even lift their own heads until they are 3 months old, are unable to engage in unassisted goal-directed reaching until their fourth month, and aren't walking on their own until their first birthday.)

One example of this type of research asks whether chicks have an innate special-purpose system for representing biological motion. In the human case, it is known that biological motion can be detected on the basis of fairly minimal cues. You don't have to see someone's full body as they walk past you. Biological motion is evident in point-light animations, in which a small number of lights are strategically placed at major joints on a moving person or animal and reproduced in a video in which only these points are visible (Johansson 1973).[3]

To determine whether newly hatched chicks have the innate ability to detect biological motion, researchers have studied chicks that were hatched in complete darkness and kept in darkness right until they were placed in the middle of a testing apparatus that consisted of a short walkway with two video displays on opposite ends (Vallortigara et al. 2005). The crux of the experiment was to determine whether the chicks would express a preference between two depicted animations—the very first things they saw in their lives. One was a point-light animation of a walking adult hen, the other was either the same points of light moving around the vertical axis (like a rigid statue of a hen being spun round and round) or the same points moving randomly.[4] The chicks turned out to have a

[3] It is impressive how much adults can discern from a point-light animation—not only the type and direction of motion but also things like the sex of the performer (when the videos are of walking people) and the weight of a raised object (when the motion is to lift something) (Blake and Shiffrar 2007).

[4] Open access videos of these stimuli can be viewed on the publisher's website: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030208

clear preference for the point-light animation of the walking hen over these other two displays. They spent the majority of their time in close proximity to the animation that depicted walking. However, when given the choice between a point-light animation of a walking hen and a walking cat (a potential predator), they showed no preference. So while chicks prefer biological motion over other types of motion, the preference isn't grounded in specific features of chicken-like movement; the underlying system of representation is a specialized system but it isn't initially tuned to conspecifics. Further work is helping to clarify the parameters of the system. One interesting feature of the system is that it exhibits a gravity bias, that is, it functions when the movement patterns are consistent with gravity but not when they aren't (Vallortigara and Regolin 2006). When shown a point-light animation of a walking hen changing direction, newborn chicks orient themselves so that they face in the same direction as the "animal" they see moving, and when the hen they see appears to change direction, they change direction to follow the hen. However, when presented with the same stimulus but upside down (and so inconsistent with gravity), the chicks do not face in the direction of the hen in this way.

Subsequent studies have yielded comparable results in newborn human infants using exactly the same stimuli as in these chick studies (Simion et al. 2008; Bardi et al. 2011).[5] This is a prime example of what we mean when we say that the different arguments for concept nativism interact and should be viewed collectively. The infants in these related studies are newborns, tested in the first few days of life (the youngest a mere 10 hours old and the oldest only 5 days old). It is doubtful that infants in this situation have enough time and the right visual experiences for a domain-general empiricist learning mechanism to develop a predilection for biological motion. (And it is especially unlikely that infants in these experiments have had much experience with walking chickens!) But if one were to suppose that it is at least *possible* that newborns do have sufficient experience with biological and non-biological motion prior to testing, this concern is completely eliminated with the chicks. There is simply no question that chicks can, and do, develop this same preference in the absence of such experience. Consequently, the work with chicks shows that an innate special-purpose mechanism for representing biological motion is part of the acquisition base for at least one other species. And given that the representational capacity it supports is so similar to the corresponding representational capacity in human newborns, it strengthens the case that a mechanism for representing biological motion is part of the human acquisition base too.

[5] Of course, the testing procedure had to be adapted to accommodate newborn infants' lack of mobility. In this case, the methods used included a visual habituation procedure and a preferential looking procedure.

What are the implications of this work for concept nativism? Different theorists are likely to take different perspectives on whether representations of biological motion that this work points to are conceptual or nonconceptual. As with our earlier discussion of faces, we will argue that regardless of whether such representations are conceptual or nonconceptual, the case for rationalism with respect to representations of biological motion contributes to the case for concept nativism.

If BIOLOGICAL MOTION is conceptual and is part of the acquisition base (or acquired via rationalist learning mechanisms), then clearly this research feeds into the overall case for concept nativism in that it would directly expand the set of concepts and conceptual domains where a rationalist account is appropriate. And on some accounts of the conceptual/nonconceptual distinction, the representation of biological motion does seem to come out on the conceptual side of the divide. For example, if what makes a representation nonconceptual is being highly fine-grained, like a representation of a maximally specific shade of red, then the representation of biological motion doesn't seem nonconceptual; the representation of biological motion doesn't correspond to a pattern that is comparably fine-grained. It doesn't even correspond to a dynamic pattern that is tied to the details of motion in a given species with its idiosyncratic body plan and gait. Rather, it is an abstract representation that applies equally across visual patterns as different as the walking patterns of chickens, cats, and human beings. This means that the representation here is a coarse-grained one, otherwise it would not be capable of capturing what is common to biological motion in the face of the massive variability regarding how animals move (i.e., how they walk, run, jump, swim, and so on).

On the other hand, on some accounts of the conceptual/nonconceptual distinction, the representation of biological motion seems to come out on the nonconceptual side of the divide. For example, it's doubtful that the representation of biological motion can freely combine with other representations in all sorts of higher cognitive processes. On an account of the conceptual/nonconceptual distinction that requires concepts to have this level of flexibility in thought, biological motion would be deemed to be nonconceptual. What then? What are the implications for the debate about concept nativism in that case? The answer is similar to what we saw earlier with the representation of individual faces in Chapter 8. Concept nativism would still be supported—but not because the account would imply the existence of an innate concept of biological motion. Instead, concept nativism would be supported because the account argues for the presence of other types of characteristically rationalist psychological structures pertaining to biological motion in the acquisition base. These innate resources pertaining to biological motion—in the form of nonconceptual representations and/or rationalist learning mechanisms or resources that such learning mechanisms trace back to—while not themselves innate concepts, would undoubtedly

play a major role in the acquisition of representations that are uncontroversially conceptual on any account of the conceptual/nonconceptual divide. In this case, such rationalist psychological structures would almost certainly be involved in the acquisition of concepts pertaining to the conceptual domains of agents, biological locomotion, and animals.[6]

Another area where research with chicks may help to tell us something about the human acquisition base is the representation of ordinary physical objects. Adult humans have numerous expectations regarding the behaviour of ordinary physical objects. For example, we expect that ordinary physical objects will maintain their physical integrity and not spontaneously split apart or allow other physical objects to pass through them, that inanimate physical objects will not move unless contacted, and that a moving inanimate object will follow a specific, predictable trajectory at a given speed. We also have expectations about what sorts of objects are physically possible and impossible (i.e., impossible in the way that a staircase in an Escher etching may appear to be simultaneously ascending and descending). A key feature of object representation for adults is that we recognize that objects continue to exist when we can't perceive them, a phenomenon known as *object permanence*. When your mobile phone falls behind the bed, you don't suppose that it no longer exists; you realize that it just isn't visible from its present location.

The enormously influential developmental psychologist Jean Piaget maintained that children go through numerous stages before they have anything that approximates an adult-like representation of objects, that they fail to represent unperceived objects as having objective and determinate locations in the early stages, and that they don't reach the stage of object permanence until about 2 years of age (Piaget 1952, 1954). Research with human infants in the past thirty or so years, however, has demonstrated that many of these expectations are present much earlier than had previously been thought (Baillargeon et al. 2011). Work with newborn chicks complements this research with human infants, showing that chicks share many of these expectations prior to having *any* visual experience at all.

Consider *amodal completion*. This happens when an object is partly occluded yet represented as a connected whole. Amodal completion is such a basic feature of the ordinary perception of objects that we normally don't even notice it. When a pencil rolls behind a glass of milk so that all that can be seen of the pencil is the tip poking out from one side and the eraser from the other, people still conceptualize it as a connected, unified object, and fully expect to see the intact pencil when the glass is removed from the table.

[6] For further arguments and evidence for a rationalist account of concepts in these domains, see Chapters 11, 13, 14, 19, and 21.

In a pioneering demonstration of amodal completion in infants, Kellman and Spelke (1983) showed that physical objects are represented in this way by children as young as 4 months old. Infants were habituated to a centre-occluded rod moving back and forth behind a rectangular block (i.e., the block concealed just a portion of the centre of the rod, with the rod's ends remaining visible).[7] Then the infants saw two test conditions without the block, one with the whole rod intact and fully exposed and one with two shorter rods (consistent with the two rod ends they had previously seen poking out from behind the block). The result was that they looked longer at the two shorter rods, indicating that they were perceiving that situation to be new and different from what they had seen before and hence that they had perceived the earlier event as the motion of a fully connected rod.

If the width of the block occluding the rod is reduced in a way that makes it easier to visually connect the exposed edges that are poking out, the same preference shows up even in 2-month-olds (Johnson and Aslin 1995). But what is happening before that, in the preceding two months? Some empiricists take the very fact we can ask this question to show that the infant studies don't provide good reason to suppose that there are innate specialized mechanisms for representing or learning about objects. For all we know, infants might be learning that centre-occlusion has no effect on an object's integrity, using an empiricist (general-purpose) learning mechanism. However, work with chicks makes this option less plausible than it might otherwise seem. This is because newly hatched chicks represent the hidden parts of occluded objects (e.g., the partly occluded rod) just like Kellman and Spelke's 4-month-olds.

This has been shown using essentially the same stimuli that were used with the infants (Lea et al. 1996). In this case, chicks that were raised in complete darkness were transferred to a room where they viewed one of three stimuli—a complete rod, two smaller rod segments, or a centre-occluded rod. The chicks were given a few hours in this room so that they would imprint on the stimulus they saw and develop a preference to be near it.[8] Then they were transferred to an experimental chamber much like the setup for the biological motion studies, with two displays on opposite ends of the chamber. On one display was an image of the complete rod, on the other an image of the two shorter rod segments. Not surprisingly, the chicks that imprinted on the complete rod spent more time near the image of the complete rod, and the chicks that had imprinted on the rod segments spent more time near the image of the rod segments. The interesting question is how the third group behaved—the chicks that imprinted on the centre-occluded rod. The

[7] We will follow standard psychological usage in which an *occluder* is any object or surface that blocks at least part of the view of another object, and a *centre-occluded* object is one in which the middle of a viewed object is occluded.

[8] Imprinting occurs when chicks rapidly form a social attachment to the first moving objects they encounter.

result in this case was that they preferred to spend time with the complete rod. This suggests that although they had never seen anything but a centre-occluded object prior to their time in the experimental chamber, they interpreted what they originally saw—the imprinting stimulus—as a fully connected rod, one whose centre just happened to be blocked from view. Evidently chicks don't need prior visual experience of objects moving in and out of view from behind occluders to develop the expectations involved in amodal completion. These expectations are grounded in an innate mechanism for representing physical objects.

What can this work tell us about human infants' representation of physical objects? It suggests that the early development of amodal completion isn't due to a domain-general learning process but derives, instead, from an innate mechanism for object representation that is as much a part of the human acquisition base as it is a part of the chicken acquisition base. How, then, can the delay of the appearance of amodal completion until the second month in childhood be explained? There are three possibilities: (1) The aspects of physical object representation responsible for amodal completion are innate and present at birth yet not manifested due to performance factors. (2) These aspects of physical object representation are innate yet not present at birth, emerging later in development as a result of non-psychological processes of biological maturation. Or (3) they aren't innate but are acquired at least in part by a characteristically rationalist learning mechanism that requires certain input to produce this outcome.

As it turns out, subsequent work with human newborns, inspired by this research with chicks, has helped to clarify which of these is most likely (Valenza and Bulf 2011). In this work, the researchers went to great lengths to reduce the task demands on the infants using a variant on the Kellman and Spelke experiment with the moving centre-occluded rod. One difference was that the width of the occluder was greatly reduced, making it easier for infants to scan between the two visible edges of the centre-occluded rod to see that they move together. And since it is known that newborn infants have difficulty detecting motion, the motion cues in the stimuli were made stronger by using stroboscopic motion (seeing the stimuli as if under a strobe light). Under these conditions, newborn infants showed clear evidence of amodal completion—the two-month lag disappeared. This argues that the aspects of physical object representation that are involved in amodal completion are innate in humans and present at birth, just as with chicks, but not always manifested due to performance factors.

Amodal completion is just one small part of way that physical objects are represented by adult humans. We will discuss other important aspects of physical object representation in Chapters 15 and 17. But for now, it is worth noting that newborn chicks have a variety of other important expectations about objects that are identical, or closely related, to human expectations. Among other things, newly hatched chicks can discriminate between possible and impossible objects (Regolin et al. 2011). They can extract three-dimensional structure from

two-dimensional shadows (Mascalzoni et al. 2009). They are capable of forming perspective-invariant representations of physical objects based on only a small sample of possible visual perspectives on the very first object that they see (Wood 2013).[9] They are also able to determine which of two occluders an object must be hiding behind taking into account their height, width, and slant (Chiandetti and Vallortigara 2011). This last result is particularly noteworthy because newborn chicks are passing tests comparable to ones that Piaget thought children fail until late in the development of object permanence.

We have been looking at examples of the argument from animals involving input considerations where experimenters impose broad restrictions on the perceptual experience of the animals tested in the experiments—the animals are prohibited from having *any* visual experience of the outside world prior to testing. But this isn't the only way to run this form of the argument from animals. Another is to allow the animals to have a considerable amount of perceptual experience but just not access to the environmental information that empiricist models require. For example, empiricist models of the acquisition of the ability to represent faces suppose that face representations are acquired when general-purpose pattern-detection mechanisms come to single out faces after prolonged exposure to recurrent patterns that appear in *environments that contain visible faces*. As a result, one way to challenge these empiricist models (and to argue for competing rationalist models) is to show that face perception emerges in animals even when they are *not* given the opportunity to see faces but aren't otherwise deprived of visual experience prior to testing.[10]

A nice illustration of this strategy can be found in work that focused on Japanese monkeys (Sugita 2008). The monkeys were raised from birth in a visually rich and socially stimulating environment yet one in which they couldn't see any faces. (To manage the deprivation, caretakers wore headgear that completely

[9] Further work by Wood and colleagues suggests that this ability is contingent upon specific features of the (extremely limited) input. While a perspective-invariant representation of an object can be formed from seeing a single (virtual) object rotate just 60°, such invariant representations fail to form if the object rotates too quickly or the rotation is not smooth or continuous (Wood and Wood 2016; Wood 2016). It is unclear whether this means that there is an innate mechanism for forming invariant representations that breaks down in these conditions, or whether there is an innate, presumably domain-specific, learning mechanism that rapidly learns to form invariant representations based on highly limited evidence.

[10] An even stronger challenge to empiricist models along the same general lines would be to show that a representational ability emerges not only when animals aren't given the opportunity to perceive good examples of the phenomenon in question but are also given considerable evidence that supports a contrary way of representing the world. We aren't aware of any studies that do this for the representation of faces, but one study that takes this form reports what happens when newborn chicks are reared in a virtual world which provides them with thousands of examples in which an object moves behind one screen only to reappear from behind another without having traversed the space in-between them (Wood et al. 2024). Despite this overwhelming evidence that objects "teleport" from one location to another when they are obstructed from view, the chicks behave no differently than when reared in a more natural world (one that exhibits object permanence and no violations), behaving in both cases as if they expect that objects to move continuously even when they can't be seen.

concealed their faces.) Different groups of monkeys were raised in this environment for six, twelve, or twenty-four months. When they were tested (using photographs of monkey and human faces seen for the first time), the monkeys exhibited a spontaneous preference for faces over other equally complex visual stimuli, and were immediately able to recognize individual faces.[11] The monkeys were then placed in a new environment for one month in which they were exposed either to human faces but not monkey faces, or to monkey faces but not human faces. This was all it took for the monkeys to develop a long lasting preference for the type of face that they were exposed to and to become better able to process this type of face. What's more, this preference persisted for at least a year (when they were tested again) even though they were exposed to both human and monkey faces in the intervening period. Notice that the monkey's preferences when first tested were clearly not driven by repeated exposure to faces—they didn't see any faces prior to the initial test (which occurred after the first six, twelve, or twenty-four months depending on the group). And their preferences were quickly set to a specific face type, monkey or human, in a way that was not reset by extensive exposure to other types of faces later on. These results provide a strong argument for an innate face-specific learning mechanism, though one that is not exclusively geared to conspecific faces.[12]

Experiments in this vein—ones that impose significant restrictions on animals' early perceptual experiences—are often called *deprivation experiments*. As we noted in Chapter 1, deprivation experiments aren't without their critics. Some go so far as to claim that they rest on faulty logic. For example, Griffiths and Machery (2008) note that "there is no such thing as raising an animal without an environment, only raising it without access to some specific aspect of the environment" (p. 404). The objection, in other words, is that when the results of a deprivation experiment are used to support a rationalist conclusion, this invariably overlooks aspects of the environment that are still present and that might influence an animal's psychological development in ways that are hard to discern and even harder to anticipate. Consider again Gottlieb's work with mallard ducks (Gottlieb 1997), which was discussed in Chapter 3. It was once thought that deprivation experiments had established that mallard ducks innately imprint on suitable adult vocalizations. Gottlieb, on the other hand, is often credited with demonstrating that this form of behaviour isn't innate by showing that researchers had failed to

[11] Interestingly, monkeys have also been shown to be susceptible to the Thatcher illusion. In this case, the research was with rhesus monkeys (Adachi et al. 2009).

[12] Critics of this work might try to argue that the monkeys, who were taken away from their mothers an hour after birth, may have seen their mother's face in their first hour of life, or that they may have garnered information about faces from touching their own faces. While these are legitimate points to raise, face recognition would have to be grounded in a remarkably robust learning mechanism in order for monkeys to acquire these face recognition abilities on the basis of seeing a single face up to two years earlier or on the basis of touching a single face (their own) over this period.

notice the causal relevance of another aspect of the environment of a mallard foetus—the sounds that the foetus itself produces.

But it is important not to make too much of these sorts of cases. As we explained in Chapter 3, rationalists are not claiming that experience, understood in the broadest possible sense, is irrelevant to development. Again, rationalists and empiricists agree that the development of all traits depends on gene-environment interactions. So what critics who oppose the use of deprivation experiments have to establish is that these experiments fail to restrict *relevant* environmental input, that is, environmental input of the type that would allow for a domain-general empiricist learning mechanism to form the trait in question. To put the point another way, the essence of a good deprivation experiment—the reason why deprivation experiments are an important tool for rationalist research—isn't that they cordon off all environmental influences. Clearly that isn't possible. Rather, it is that the animals are prevented from having the types of perceptual experiences that competing empiricist models claim are integral to the development of specific representational abilities. We think there should be little question that the studies we have been discussing (with chicks and monkeys) do just that.[13]

Turning to the ramifications of these deprivation experiments and the question of precisely what they show to be innate, there is considerable latitude for interpretation. As with the argument from early development, there are also related issues about the content and interrelations among the representations that research on animals has been uncovering. In Chapter 8, we suggested that human infants discriminate faces in ways that are connected to social and communicative capacities and that most likely reflect an interest in individuals. Similar remarks apply to monkeys' representation of faces. Given how important discriminating particular individuals is to parent-offspring identification, territoriality, mating, and hunting, among other things, it should hardly be surprising if the representation of individual agents were to be found in a wide variety of species.[14] In highly social species, one might also expect animals to have

[13] There is also related research with animals that doesn't impose rearing conditions that deprive animals of the sorts of experiences that empiricists claim are needed to acquire a given concept or given conceptual ability. Instead, it simply capitalizes on cases where we can be fairly confident that the experience in question hasn't occurred—as we saw in Chapter 4 with the poverty of the stimulus argument. For example, Chen et al. (2006) examined capuchin monkeys' response to losses and gains when trading tokens for food (in effect, using tokens as money). Although these monkeys had no prior experience with money and trade, they were found to have some of the same decision-making biases as humans. In particular, they exhibited *loss aversion*, in which a loss is deemed more serious than an equivalent gain. As Chen et al. point out, given the evolutionary proximity between capuchins and humans and given this common psychological trait, it is reasonable to conclude that loss aversion is an innate feature of the human mind. If this is right, then this would support a rationalist account of the concepts LOSS and GAIN. See Santos and Rosati (2015) for an overview of related work.

[14] In fact, it has—in, for example, non-human primates (Cheney and Seyfarth 1999), other mammals (Adachi et al. 2006), birds (Kondo et al. 2012), and even in fish (Johnsson 1997), frogs (Chuang et al. 2017), crustaceans (Karavanich and Atema 1998), and insects (Sheehan and Tibbetts 2011).

innate dispositions to categorize conspecifics according to the social relations that are relevant to their form of social organization (Cheney and Seyfarth 2008; Seyfarth and Cheney 2015). Baboons, for example, are known to simultaneously represent group members in terms of kinship and social dominance hierarchies (Bergman et al. 2003). This work is all of great interest for understanding the minds of these nonhuman animals. Here, however, we are concerned with how it can shed light on the human acquisition base. Following the logic of the argument from animals, we take it to provide prima facie grounds to suppose that analogous kinds of capacities are likewise rooted in characteristically rationalist psychological structures in the human acquisition base.

More could be said about work that limits the sorts of experiences animals have prior to testing in these ways. But let's move on to the other form of the argument from animals mentioned previously, the one based on the character of the acquisition mechanisms involved, in which certain animals are shown to possess interesting representational abilities despite lacking the powerful general-purpose learning mechanisms that empiricist models use to explain their acquisition for humans.

Animals are prodigious learners. Even exceedingly primitive animals are capable of associative learning. The nematode worm *C. elegans*, with a mere 302 neurons, can be conditioned to respond to stimuli involving taste, smell, temperature, and oxygen level (Ardiel and Rankin 2010). Honeybees, with roughly a million neurons, are even capable of associative learning that goes beyond straightforward links between perceptual stimuli or between a response and a perceptual stimulus. For instance, they can learn that while two stimulus types are each positive (S1, S2), their combination (S1+S2) is negative (Giurfa 2003).

Now many theorists have supposed that animal learning is largely confined to associative learning. But the examples of animal learning we already touched on in Chapter 4 suggest that this is implausible and that rationalist learning mechanisms are needed too. Consider dead reckoning again, which we noted is used by many types of animals to find their way back to a point of origin after meandering through their environment. It is easy to see how a technologically advanced language-using species might discover and refine this navigational technique and pass it on to novice learners—as human sailors used to do before the widespread use of computer-based navigation and GPS. But how could an *ant* learn that it

Earlier we mentioned that the experiments with newborn chicks take advantage of the fact that chicks imprint on objects seen early in life and that this involves the formation of a social attachment. It is well known that the limbic system (especially the septal nuclei) is involved in social decision making in many animals. Thus another interesting and useful line of investigation that the argument from animals may take is to examine the brains of newborn chicks whose first seen object elicits imprinting. Early studies of this kind have found that the septal nuclei are selectively active when a newborn visually naive chick sees another chick. In other words, even for newborn chicks that haven't seen any objects at all, brain structures for social decision making respond when it sees another chicken (Mayer et al. 2017).

can find its way home by summing distance and direction information in just the right way? How would it even get the idea that direction and distance can be combined to plot a direct and efficient path? The only plausible account for creatures like this is that dead reckoning is grounded in an innate special-purpose mechanism that is part of their species' acquisition base. And once again, if it is true that ants innately possess a learning mechanism of this type, this demonstrates that such a system can be innate and provides reason to suppose that if humans possess a similar system, it may well be innate in humans too.[15]

When we turn to the origins of concepts, this form of the argument from animals is especially strong in cases where animals that lack powerful domain-general learning systems are nonetheless found to represent abstract content—content that is far removed from sensory stimulus conditions. In these cases, there plainly isn't an empiricist learning mechanism to explain how these animals generate such representations. The only serious candidates are rationalist accounts that attribute to them innate abstract representations or rationalist learning mechanisms that are specifically geared towards representing the world in these abstract terms.

Consider, for example, the representation of *same* and *different*. What makes two things the same or different varies considerably from case to case. It may be their colour, their odour, their orientation, their texture, their shape, and so on. Sameness regarding one of these dimensions (e.g., colour) isn't anything like sameness regarding the others (e.g., odour), and hence a truly general representation of sameness is an abstract representation par excellence. For this reason, it is perhaps surprising to learn that animals that lack powerful domain-general learning systems (e.g., one based on a communication system with the expressive power of human natural language) can represent *same* and *different*. Yet some can.

One elegant study has shown that honeybees have this representational ability (Giurfa et al. 2001). In this work, bees were trained in a Y-maze (a simple maze shaped like the letter 'Y'). They first encountered a sample, A, shortly after entering the stem or bottom portion of the Y. Then, after moving further along the straight initial portion of the Y-maze, they came to a chamber in which they had to decide which of two paths to follow (the two arms of the end of the Y-maze), one marked with stimulus A, the other with stimulus B. The bees were rewarded with access to a sucrose solution for taking the path with the matching stimulus, A, and soon learned to choose this path. Following this training, they were subsequently tested on novel stimuli without rewards. For example, if trained to choose between two colours (blue vs. yellow), they were tested on two black and white

[15] Similar considerations apply to the system of geometrical reorientation discussed in Chapter 1, where a comparable system has been found in bees (Lee and Vallortigara 2015). In this case, the previous form of the argument from animals is in effect too, since this system has also been found in animals (newborn chickens) that have been prohibited from having any visual experiences related to reorientation prior to testing (Chiandetti et al. 2015).

gratings (oriented horizontally vs. oriented vertically). What's more, they were also tested on stimuli from a new sensory modality (this time trained on odours and then tested on colours). In each case, the bees were able to generalize sameness to the novel stimuli, whether they were in the same or a different modality. Evidently, what the bees learned was to choose based on sameness in general, whether it involved colour, pattern, or odour, demonstrating that they possess an abstract general representation of sameness. (To test bees for *different*, much the same test was used but researchers instead rewarded choosing the non-matching stimulus during the training phase.)[16]

How is it that the bees were able to make the right choice when confronting test stimuli in a new modality? One could try to claim that the bees somehow learned the representations SAME and DIFFERENT in the course of the experiment, but this is unlikely. In addition to the fact that bees have limited cognitive resources to draw upon, the training procedure itself used a narrowly defined set of matching stimuli—stimuli from just one modality (e.g., vision) that instantiated just one dimension (e.g., colour) and just one contrast within that dimension (e.g., blue vs. yellow). Why—and how—would a general-purpose learning mechanism generalize from this meagre sample to a highly abstract general representation of sameness?

We suggest that the most plausible account here holds that bees have innate representations for *same* and *different*.[17] In this case, the bees *do* learn something in the experiment. They learn that rewards are associated with the stimulus that is the same as the sample encountered near the entrance of the maze (or the one that is different, as the case may be). But this learning depends on the prior

[16] Work with animals has also looked at whether *same* and *different* are available to animals that have been raised without access to relevant perceptual stimuli. One study used this approach with ducklings (Martinho and Kacelnik 2016). Newborn ducklings initially saw a pair of moving objects that were either the same or different in shape or colour. (As with the studies with newborn chicks, this involved just enough exposure so that the ducklings would imprint on the stimulus and develop a preference for it over other stimuli.) Subsequently, the ducklings were shown novel pairs of objects and were found to selectively follow pairs that instantiated the abstract relation—*same* or *different*—that was instantiated by the pair they had imprinted on.

[17] In principle, much the same work could be done here without an explicit representation of sameness (or difference) per se. For example, rather than representing two stimuli as being the same, the bees might instead represent the fact that there exists a property that the two stimuli are both instances of. Notice, however, that while this type of alternative might strictly speaking avoid attribution of the representation SAMENESS to bees, it would nonetheless involve attribution of highly abstract innate representations—for quantification ("there exists"), for representing properties in general, for representing being an instance of a property, and so on—which are part of a small circle of interdefinable representations that includes SAMENESS. The alternative that doesn't involve postulating an explicit representation of sameness effectively exploits the fact that two items *being the same* (in some specific respect) involves *there being a property which the two items are both instances of.* While such an alternative might technically avoid attributing the representation SAMENESS to bees, it is no less strongly rationalist than an account that does attribute SAMENESS to bees.

possession of the representations SAME and DIFFERENT themselves.[18] In any event, the present point is that we can use examples like this to address questions about the human acquisition base. The overall strategy is to establish that animals without the powerful domain-general learning systems that humans have can nonetheless possess abstract representational abilities—abilities that empiricists typically take to be the product of such powerful systems in humans. This argues that the representational content is innate or acquired via rationalist learning mechanisms in these animals. And to the extent that humans possess the same (or highly similar) representational abilities, it gives us reason to suppose that such content is also innate or acquired via rationalist learning mechanisms in the human case, too.

Another domain where this form of the argument from animals adds to the case for concept nativism is the representation of numerical quantity. In Chapter 8 we saw that newborn human infants can represent numerical quantity. But a fuller picture regarding the origins of numerical representation has been emerging in recent years as researchers have taken seriously the possibly that numerical representation is an evolutionary ancient representational capacity. Much of this research has focused on what we have been calling the approximate number system. We saw earlier that this system is functional at birth in human infants and that it continues to operate past infancy and throughout the human lifespan.

Interestingly, comparative studies have shown strong parallels between the operation of this system in humans and animals. For example, one study examined the representation of numerical quantity in the context of a numerical ordering task given to both rhesus monkeys and university students (Cantlon and Brannon 2006). The task used sets of particular numbers of clipart pictures that varied in size, shape, etc., controlling for continuous extent. With monkeys, the task involved first training them to order pairs of numerical stimuli in the range of 1–9 in ascending numerical order. They were then tested on novel stimuli for familiar and novel numbers, including 10, 15, 20, and 30. With human participants, training wasn't necessary, of course, since they could be given verbal instructions on the task. To discourage them from using precise integer representations rather than their approximate number system, they were instructed not to count and to respond as quickly as possible. The results were

[18] See Cope et al. (2018) for a computational model of how sameness and difference could be implemented in the brain of a bee—a model that takes into account some of the known properties of how their brains are organized. This model works by building in structures that differentially respond to matched or non-matched stimuli of a given type (e.g., yellow-yellow vs. yellow-blue) and structures that are responsive to the presence or absence of this within-stimulus-type index of sameness across different stimulus types, allowing the system to represent a generalized form of sameness/difference. Although Cope et al. describe their model as showing how sameness and difference could be learned, we see it instead as showing how bees might learn to determine whether matching (or non-matching) in general (that is, across stimulus types) is being rewarded in the experiment and choose the appropriate action in light of this—or, in other words, how they might learn when their behaviour should be guided by the perception of sameness or difference.

remarkably similar for the two groups—monkeys and humans responded at much the same speed and exhibited a virtually identical pattern of errors in accordance with the approximate number system's ratio dependence.

Work with other species has established that the representation of approximate numerical quantity is widespread in the animal kingdom. It can be found in primates (Brannon and Terrace 1998), other mammals (Suzuki and Kobayashi 2000), and birds (Rugani et al. 2010). But for present purposes, we want to emphasize that it can also be found in animals that are more phylogenetically remote from humans—including fish (Agrillo et al. 2010) and even insects (Gross et al. 2009)—since such animals are even less likely to have acquired these representations using the sorts of powerful domain-general learning mechanisms that humans have.

We will mention just a small sample of the work with fish to convey how researchers have come to hold that animals as lowly as the mosquitofish have representations of numerical quantity. In one study (Agrillo et al. 2010), female mosquitofish were individually placed in a transparent enclosure that had doorways (transparent flaps) leading to the outer tank where there was a group of female mosquitofish they were motivated to join. The doorways were labelled with stimulus patterns composed of abstract geometrical figures that changed from trial to trial but in a way that presented the fish with the same numerical choice (e.g., a doorway labelled 4 versus a doorway labelled 8). By arranging things so that only one of the labelled doorways in each trial actually allowed the fish to pass through to the outer tank (say, the doorway that was labelled 4), the researchers were able to reward the fish for choosing the door labelled with a target number. After the fish learned to respond to the right number (4 or 8), they were tested using new labels—instances of 4 and 8 that they had never seen before. Despite the novelty of the stimuli, the fish managed to continue to respond to the numerical quantity that gave them access to the outer tank.[19] In further experiments, the mosquitofish discriminated between other numerical quantities that stand in this same 1:2 ratio (5 vs. 10) and a ratio of 2:3 (8 vs.12), and also showed that they are able to discriminate large numbers (15 vs. 30, 100 vs. 200).

In considering how the work on numerical representation in animals bears on the human case, it is important to keep in mind that the evidence isn't just that mosquitofish demonstrate an ability to represent numerical quantity and to use it in making decisions (although this is an impressive fact all on its own). It is that mosquitofish represent numerical quantity in the same ratio-dependent way as

[19] The stimuli were chosen to ensure that the fish couldn't succeed by representing non-numerical properties. The position, shapes, and sizes of the geometrical figures were varied, and controls were in place to exclude responses based on the cumulative surface area of the geometrical figures, their density, the space covered by the arrays, and differences in luminance. One non-numerical property wasn't specifically controlled for—the cumulative amount of perimeter in each array—since previous work suggested that it isn't a salient property for mosquitofish in similar test conditions (Agrillo et al. 2009).

humans.[20] Yet mosquitofish also clearly lack the sorts of powerful general-purpose learning mechanisms that empiricists see in humans and that they rely on in order to explain the origins of numerical representation in the human case. For example, a common empiricist strategy is to maintain that children can only represent non-numerical features of stimuli that are confounded with numerical features until cultural practices mediated by language help them to see that numerical quantity can be distinguished from continuous quantity (Leibovich et al. 2017). But of course mosquitofish don't have access to these same cultural practices and can't learn about numerical quantity from learning numerical words. The most plausible account, rather, is that the approximate number system is an evolutionarily ancient system of representation that is part of the mosquitofish acquisition base (and part of the acquisition base for many species of animals). This, in turn, complements and strengthens the case based on the argument from early development that the reason the approximate system can be seen to be operating in early childhood isn't because of fast general-purpose learning but because it is innate for humans too.

In Chapter 8, we noted that there is little dispute that interpreting and working with conventional numerical symbols is a conceptual activity, as when people solve complex arithmetic problems or even just report which of two numerals expresses the larger numerical quantity. Given the pervasive role that the approximate number system plays when people do such things and given, as we have seen, that the approximate number system is most likely a part of the human acquisition base, there is good reason to suppose that the *number* domain should be included in the case for concept nativism regardless of whether representations in the approximate number system are conceptual or not. Even if these approximate numerical representations are nonconceptual, they will support concept nativism regarding concepts such as those of natural numbers. At the same time, though, one might wonder whether these approximate representations themselves are concepts as well. If they are concepts, then in addition to contributing to a rationalist account of the origins of concepts of natural numbers, they would also be concepts in their own right for which a rationalist account is appropriate, providing an even stronger form of concept nativism in the number domain. And since this further question regarding the conceptual or nonconceptual status of approximate numerical representations is of independent interest in any case, we will briefly consider it here.

One type of theorist who might take approximate numerical representations to be nonconceptual is a theorist who draws the conceptual/nonconceptual

[20] Work with newborn chicks illustrates further striking commonalities regarding the way humans and animals represent numerical quantity. We saw in Chapter 8 that newborn humans map numerical quantity to space, with smaller numerical quantities on the left and larger numerical quantities on the right. It turns out that 3-day-old chicks do this too (Rugani et al. 2015).

distinction by appealing to the Generality Constraint. For example, Beck (2012) has argued that an account of the conceptual/nonconceptual distinction that is based on the Generality Constraint has the implication that approximate numerical representations aren't conceptual. Recall that the Generality Constraint concerns whether a representation can be combined with other representations to form complex representations with novel complex contents. As it is often understood, it says that a necessary condition on a representation's being conceptual is that the representation must be able to freely recombine with other concepts to form other related thoughts. For example, on the assumption that someone can think SAM IS TALL, then BILL is a concept for this person only if she can also entertain the thought BILL IS TALL too. Beck argues that this constraint on novel combinations isn't met for approximate numerical representations as can be seen in animals' representations of numerical quantity. A pigeon might be able to represent 40 pecks to be less than 50 pecks, and 38 to be less than 47, but not be able to represent 38 to be less than 40. (This is because the approximate number system in pigeons can only discriminate values that exceed a 9:10 ratio.) According to Beck, this means that the approximate representations that pigeons use when representing the numerical quantity of a sequence of pecks can't be concepts.

Much could be said regarding this argument, but we will focus on just two points. First, while it is clear that the pigeons cannot reliably distinguish 38 from 40 pecks, this fact alone is not enough to show that pigeons can't form the thought that represents 38 as being less than 40 (unless we assume some form of verificationism, in which having a thought requires being able to determine it is true).[21] It may well be that while pigeons can't reliably distinguish 38 from 40, they can nonetheless think that 38 is less than 40, and that the very same representations of 38 and 40 that they use when thinking 38 PECKS ARE LESS THAN 47 and 40 PECKS ARE LESS THAN 50 can be used to think 38 PECKS ARE LESS THAN 40. Of course, being able to think that 38 is less than 40 wouldn't be a particularly useful thought for pigeons if they aren't able to reliably distinguish 38 from 40. For example, it wouldn't enable them to reliably succeed on a task that required them to discriminate between these two values. But it may well be that all this means is that pigeons would have little reason to entertain this thought—not that they were incapable of doing so—and that we lack a readily accessible behavioural criterion for whether they are capable of forming this thought. Their inability to demonstrate to us that they can entertain a behaviourally useless thought says nothing about whether they are capable of entertaining such thoughts in principle.

[21] For ease of exposition, we will continue to use natural number expressions to indicate the relevant approximate numerical representations, rather than more cumbersome expressions such as "represents the numerical quantity of being approximately 38 as being less than the numerical quantity of being approximately 40". Of course, even these more cumbersome expressions provide only a rough and inexact gloss on the representational contents involved.

One way to see the problem here is by noting that the same basic situation can arise for thoughts composed entirely of representations that are uncontroversially conceptual. For example, consider the following thoughts, which should be understood as employing concepts for natural numbers, not approximate numbers, along with the concept ABOUT. It seems clear that as humans we can think *being about 178 seconds long is definitely longer than being about eight seconds long*, and we can think *being about 177 seconds long is definitely longer than being about nine seconds long*. But does the fact that we can't reliably discriminate between being about 177 seconds long and being about 178 seconds long mean that we can't entertain the thought *being about 177 seconds long is definitely less than being about 178 seconds long*? It seems that we *can* readily entertain this thought—you just did!—even if we may not know exactly which intervals the concepts BEING ABOUT 177 SECONDS LONG and BEING ABOUT 178 SECONDS LONG apply to or even whether the thought *being about 177 seconds long is definitely less than being about 178 seconds long* is true.

The second problem for Beck's analysis that we will note is that it is not at all clear the degree to which we should say that concepts must be able to "freely" combine with one another. Even if the representation for 38 that is used when thinking 38 PECKS ARE LESS THAN 47 can't be combined with the representation for 40 that is used when thinking 40 PECKS ARE LESS THAN 50, each of these can be combined with *many* other representations to express a wide variety of number-involving contents. The use of approximate numerical representations in pigeons (and other animals) is not restricted to representing numerical quantities associated with their own actions (e.g., pecking or bar pressing and the like). Animals can discriminate between numerical quantities of many types of things, using the same numerical representations for many different purposes.

We'd suggest that a better way of approaching the question of whether approximate numerical representations are conceptual in animals (and by extension, in humans) is to ask whether such representations figure in a variety of higher cognitive processes such as categorization, planning, and decision making. As it happens, there is much evidence from work with animals to suggest that the approximate number system's representations are indeed conceptual given this approach. This may be surprising to some because it's often thought that animals have little need to represent numerical quantity and that they only do so in very limited contexts or when the reward structure for an experiment forces this upon them after a very long training period. But in fact, animals spontaneously attend to numerical quantity and use numerical information to accomplish a large variety of tasks. These include using numerical information to make decisions about whether to attack an opponent in a territorial dispute depending on the number of allies the opponent has (e.g., Wilson et al. 2001; Bonanni et al. 2011), to form a preference for one individual over another based on which of the two made a larger number of scent marks in a particular context (e.g., Ferkin et al. 2005), to

determine which of several groups is safer to join based on the numerical sizes of the groups (e.g., Agrillo et al. 2007), to keep track of a landmark based on its numerical position in a sequence (e.g., Suzuki and Kobayashi 2000; Dacke and Srinivasan 2008), to detect opportunities for brood parasitism or to thwart it based on the expected number of eggs in a nest (Lyon 2003; White et al. 2009), to represent the rate of return for selecting one food source over another (e.g., Lima 1984; Bar-Shai et al. 2011), among other things.

More speculatively, there are many benefits to tracking numerical quantity that would create selective pressure on animals to use numerical quantity in other ways. To name just a few, an animal might want to track the number of predators on a particular encounter to know whether unseen predators might still be around. Likewise, it might be beneficial to a predator tracking a group of prey to know how many in a group it had been chasing had escaped in order to know how many might now be hiding. It might also be beneficial to track the number of sexual or status rivals in an area, irrespective of imminent battles or mating opportunities, to prepare for future possibilities. And when foraging, it might be beneficial to represent number to track the numerical details of particular types of landmarks or goal objects, for example, for pollinators to track numbers of petals on flowers to help discriminate higher and lower value nectar sources. This list is not intended to be exhaustive, but only to convey some of the breadth of different types of situations where thinking about numerical quantities may be employed and have value. Many of these situations involve planning, decision making, and the making of tactical assessments of various kinds, involving unquestionably abstract representations that figure in processes which in humans would be considered examples of higher cognition. All of this suggests that the numerical representations of the approximate number system may best be thought of as concepts in many animals.

In the end, however, the argument from animals isn't about the status of rationalism regarding animal minds. It's about how research on animal minds argues for rationalist accounts of the origins of concepts in the human case.[22] In light of the considerations we have been reviewing about how the approximate number system's representations may function as concepts for animals, it is overwhelmingly likely that such representations are likewise involved in a wide variety of higher cognitive processes such as categorization, planning, and decision making in humans. And even if there were doubts about whether such considerations were sufficient to establish that these representations were conceptual in some animal species, there is no reason for such doubts to carry over to the human case. Accordingly, these considerations provide strong grounds for supposing that approximate numerical representations are conceptual in humans,

[22] Of course, this isn't to say that the evidence we have presented does not also support rationalist conclusions about the origins of concepts in many animals as well.

and that the origins of such representations in humans should be explained in rationalist terms. In that case, the approximate number system's bearing on concept nativism isn't just that it contributes to a rationalist account of the origins or concepts of natural numbers, but even more directly contributes to concept nativism in the number domain in light of approximate numerical representations being concepts whose origins should be explained in rationalist terms.

More could said about the examples we have mentioned in this chapter to illustrate the different versions of the argument from animals, and we could cite many other representational abilities that connect with one or another form of the argument from animals. But as with the argument from early development, our goal isn't a comprehensive account of all instances of the argument but to offer enough evidence to see how it contributes to the overall case for concept nativism and to our claim regarding concept nativism's breadth and depth.

It is also worth mentioning that regardless of how many examples we were to cite, the argument from animals, in all its forms, is based on research that is still very much in its early days and that its potential has barely begun to be tapped. This is because the study of animal cognition has a history dominated by a reluctance to credit animals with any rich representational abilities. Until relatively recently, researchers weren't even looking for sorts of abilities that bear on the status of concept nativism.[23] It is thus a relatively new and extremely welcome development that anyone would even investigate whether fish represent numerical quantity, whether baby chickens represent the continued existence of an occluded object, and so on. And of course, if you aren't asking questions like these, then you are bound to miss out on important elements in the acquisition base for other animals, and, by extension, you will have a limited understanding of the human acquisition base.

Even when it comes to animals that are far closer to humans (in evolutionary terms and in their general intelligence), the trend is that we are only just beginning to uncover representational abilities that are likely to be grounded in rationalist learning mechanisms that also bear on the structure of the human acquisition base. For example, the conventional wisdom regarding the attribution of false beliefs has long been that only humans can represent false beliefs (and even then, not reliably until around 4 years of age; see the previous chapter). In line with this supposition, chimpanzees had been found to fail what were thought to be critical false belief tasks (see, e.g., Povinelli and Eddy 1996; Tomasello and Call 1997; Call and Tomasello 2008). However, as experimental techniques have become more sensitive to the details of chimpanzees' and other great apes' social

[23] The legacy of behaviourism in psychology lived on considerably longer in comparative psychology than it did in cognitive psychology, and as a result, the growth in the volume of research on aspects of animal cognition beyond those studied by behaviourists lagged behind that seen in many other areas of psychology by as much as twenty years.

lives and have reduced the obstacles stemming from performance factors, researchers have uncovered compelling evidence that chimpanzees, bonobos, and orangutans can all represent false beliefs (Krupenye et al. 2016; Buttelmann et al. 2017; Krupenye et al. 2017; Kano et al. 2019).

Needless to say, humans differ from apes in many ways. Our claim is not that all interesting human cognitive capacities are identical to the capacities found in other primates, much less to ones in birds or fish or insects. Nor is it that human cognitive capacities that are similar to capacities in other species must function in an identical manner or that they are invariably subserved by the very same underlying machinery. Rather, our claim is that findings of shared and overlapping capacities can nonetheless play a key role in a version of the argument from animals and thereby provide strong support for rationalist accounts of concepts. In particular, such findings can provide strong arguments to the best explanation for there being rationalist psychological structures which, in humans, play a significant role in the acquisition of a range of concepts. While other examples could be given to illustrate this line of reasoning, the examples that we have discussed (bearing on the acquisition of concepts pertaining to faces, agents, animals, sameness, difference, and number) illustrate how the argument from animals, in its different forms, can contribute to the overall case for concept nativism.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. DOI: 10.1093/9780191925375.003.0010

11
The Argument from Universality

Our third argument for concept nativism points to a type of evidence that has historically been at the centre of the rationalism-empiricism debate but that has often been misunderstood. Put in the simplest possible terms, the *argument from universality* maintains that cognitive and behavioural universals can count as evidence for the existence of innate concepts and concepts that are learned via rationalist learning mechanisms. We want to emphasize, however, that this initial formulation of the argument—or any comparably succinct formulation—can't do justice to the complex relation between concept nativism and universality. Unpacking the argument from universality, and seeing how cross-cultural data can favour concept nativism, will take some work. In the end, what we hope to make clear is that this argument doesn't turn on a simple rule that links universality with a rationalist acquisition base. Rather, it asks us to attend to the way that similarities across cultures are best explained in particular cases and holds that sometimes the best explanation is a rationalist explanation.

Many critics of rationalism have failed to appreciate the complexity surrounding the argument from universality. In fact, empiricists going back as far as Locke have mistakenly rejected the argument on the grounds that it adopts a naive outlook that is too quick to postulate universals and that turns on the flawed assumption that universals entail innate psychological structures. The idea is that postulating universals too readily overlooks the enormous variability that is found across cultures. And even if a concept or conceptual ability were found to be universal, it needn't be innate—it could instead reflect a readily discovered solution to a widespread problem, a successful cultural product that has spread across the globe, or even straightforward observations regarding features of the world that hold in any habitable environment. For these reasons, many empiricists have claimed that the argument from universality is so flawed that considerations having to do with universals aren't possibly going to help build a case for concept nativism.

As tempting as these criticisms may seem, this empiricist outlook is mistaken in that it rests on an oversimplified picture of how the argument from universals is meant to work and how rationalism is informed by cross-cultural data. There are a number of ways in which rationalism is perfectly compatible with traits that are not evidenced universally across cultures (but instead show a more nuanced form of universality). And while it is possible to explain universality in non-rationalist terms, this doesn't mean that such explanations are the best

explanation for any given instance of universality, or that they should automatically be preferred over a rationalist explanation. Understanding when and how such evidence best supports either empiricism or rationalism can be difficult, but it is a difficulty that cannot be avoided if we want to come to a proper evaluation of the evidence.

A first step to seeing the potential of the argument from universality as an argument for rationalism is to recognize a number of factors that complicate the relation between concept nativism and universality. We will begin this chapter with a brief discussion of five of these, and then will turn to a few examples where recent research suggests that the argument from universality makes a significant contribution to the case for concept nativism. The five complications that we will begin our discussion with concern (1) the existence of conditional and disjunctive universals, (2) the relevance of the competence-performance distinction, (3) the possibility that innate psychological traits can be modified or overridden, (4) the fact that concept nativism includes a commitment to learning, and (5) the impact of growing up in a degraded environment.

1. Disjunctive and conditional universals. Critics of the argument from universals often suppose that cognitive and behavioural universals must be *simple universals*, positing traits that are universally present across all cultures. But this overlooks more subtle types of universals that are just as relevant to the evaluation of concept nativism, namely, disjunctive universals and conditional universals.

Disjunctive universals are perhaps most familiar from rationalist theories of language acquisition (Laurence and Margolis 2001). One widely discussed possibility in linguistics, which we have already encountered, is that Universal Grammar allows for a small number of different options (or *parameters*) regarding certain key features of languages, with different natural languages making different choices among these options. For example, the head of a phrase is the word that determines the core of its meaning and its grammatical properties (e.g., the verb "ate" is the head of the phrase *ate an apple*). The head directionality parameter in linguistic theory is a proposed disjunctive universal in which there is an option between a language's heads appearing at the beginning of a phrase or at the end. English is a head first language (you say *Sally eats an apple*, not *Sally an apple eats*), whereas other languages (e.g., Japanese and Basque) are head final. A similar situation might hold for concepts. Depending on its initial input in development, the conceptual system could realize one or another of a small number of alternatives for a given domain. Some have speculated that this is true for aspects of moral cognition, for example, with implications for moral concepts and intuitive moral and legal thinking (Mikhail 2011).

Conditional universals (or context-sensitive universals) are another important departure from simple universals. In this case, what is universal is a set of conditional or context-sensitive cognitive settings, where a given setting is manifested

only in particular circumstances. Since the conditional responses to different types of environments may be innately specified or the product of innate domain-specific mechanisms, conditional or context-sensitive universals are perfectly compatible with rationalism.[1]

The possibility of disjunctive or conditional universals illustrates one way in which certain types of cross-cultural variation are compatible with the existence of innate universals, and therefore compatible with concept nativism.

2. The competence-performance distinction. A second factor complicating the relation between rationalist accounts of conceptual development and universality is the competence/performance distinction. A conceptual universal may not be apparent due to what we call *competence masking* factors. Often enough, when the competence-performance distinction is invoked, what is intended is to emphasize the way that a psychological competence doesn't directly leave traces on behaviour because it is always limited by such general resources as memory and attention. (Some of the work on false-belief representation mentioned in Chapter 9 turned on this type of consideration.) But, in addition, we need to take into account other psychological processes that are engaged in a competition to influence a person's subsequent thinking or behaviour. In these cases, such psychological processes *mask* the competence, which would otherwise push things in the direction of a different outcome.

Consider the suggestion that there may be a contextually sensitive cognitive universal that holds that harming other people is morally wrong. One reason that empiricists have been sceptical about such a universal is that people do harm each other all the time, and that there is a great deal of cross-cultural variability regarding torture, fighting, and other harm-inducing practices. For example, Prinz writes: "Is there a universal prohibition of harm? The evidence is depressingly weak. Torture, war, spousal abuse, corporal punishment, belligerent games, painful initiations, and fighting are all extremely widespread. Tolerated harm is as common as its prohibition" (2007, p. 373). But in evaluating such potential

[1] It may seem that disjunctive and conditional universals threaten to trivialize the question of whether there are universals. Couldn't any purported counterexample to the claim that a psychological trait is universal simply be described as a more complex conditional or disjunctive universal? If so, then any amount of variation could be said to be rooted in a "psychological universal". It is true that in principle it could; however, it is crucial to remember that the argument from universality comes down to an argument to the best explanation (more on this below). This means that the question isn't whether it is possible to find a disjunctive or conditional universal that is consistent with known patterns of variability across cultures, but whether any given disjunctive or conditional (or, for that matter, simple) universal provides the best overall account of these patterns. This is a highly non-trivial burden to meet. It is also a burden that isn't unique to proponents of the argument from universals. Its critics face a comparable burden. It is not enough for them to show that an account that does not posit universals is consistent with an instance of cross-cultural variability. They likewise need to show that such an account, and the psychological structures that it claims are responsible for the variability, provides the best overall account of the data.

counterexamples to a universal harm norm, it shouldn't be assumed that harm-inducing behaviours are a direct reflection of the agent's harm-related behavioural norms. On the contrary, there is every reason to believe that self-interest, anger, fear, and a multitude of other psychological variables can interfere with the compliance to a norm that one accepts. So failing to conform to a putative norm does not in and of itself show that the norm isn't endorsed.[2]

The significance of this for the argument from universals is that a cognitive universal that is embedded in an underlying competence—including a rationalist one such as an innate domain-specific competence—needn't show up as a simple behavioural universal. Performance factors and competence masking are other ways in which surface-level variation is compatible with the existence of an innate universal (in this case, a universal competence).

3. Innate psychological traits that are modified or overridden in development. A third factor complicating the relation between rationalist accounts of conceptual development and universality is that innate traits might be either modified or overridden in development. In these cases, a universal that is present at an early stage in development may be difficult to identify because it is no longer present at later stages. Concept nativism holds that there is a rationalist account of the origins of concepts in more than just a few content domains (and our own account holds that this is true of *many concepts* across *many conceptual domains*). This means that these concepts are either innate or else acquired by rationalist learning mechanisms that trace back to characteristically rationalist structures in the acquisition base. But, as explained in Chapter 2, this does *not* require that any (much less all) of the psychological structures that make up the acquisition base are permanent or unchangeable. In fact, we know from work with animals that innate psychological traits can be overridden. For example, an innate prey preference in cuttlefish for shrimp over crab can be reversed if they are exposed to crabs but not shrimps immediately upon hatching (Darmaillacq et al. 2006). Even bees can override innate dispositions (e.g., attractions to certain pheromones) through learning (Roussel et al. 2012).

There are a number of ways in which the possibility of innate traits being either modified or overridden might bear on concept nativism in humans. Explicit learning might override innate preferences, dispositions, or other traits in humans in much the same way that it does with other animals. We have already seen examples of innate traits undergoing change and development above—for example, a preference for faces that is initially indeterminate transforms into a preference that is specific to human faces. A related form of representational

[2] There are other difficulties involved in Prinz's list as well. For example, harm norms are plausibly conditional in various ways, and some conceptualizations of corporal punishment or belligerent games may not classify these as harms at all (see Chapter 18 for further discussion).

narrowing is known to take place in language acquisition.[3] Young infants are sensitive to many basic sound contrasts employed in the world's languages which aren't used in the native language of their community and which adults in their community have difficulty discriminating (Werker and Tees 1984; Werker and Hensch 2015). Something comparable may happen at the conceptual level in some conceptual domains. For example, there is evidence of representational narrowing in the way that language encodes spatial representations. English and Korean represent spatial relations in ways that cross-classify one another. While English distinguishes support (ON) from containment (IN), Korean singles out tight fit versus loose fit (e.g., classifying a ring worn on the finger and a cassette tape in its case as the same type of spatial arrangement, namely, tight fit). Yet an investigation of 5-month-olds from monolingual English-speaking families has found that these prelinguistic children readily distinguished tight-fit events from loose-fit events even when (as we'd say in English) both are instances of one object's being in another or one object's being on another (Hespos and Spelke 2004). A plausible model of what happens in conceptual development in such a case is that language learning selects among the concepts available to prelinguistic infants (Hespos and Spelke 2004). In other words, both IN/ON and LOOSE FIT/TIGHT FIT are available to both English-speaking and Korean-speaking children, but English-speaking children eventually settle on IN/ON in conformity with their language community, while Korean-speaking children settle on LOOSE FIT/TIGHT FIT in conformity with their language community. Language learning, it would seem, reduces and hones a conceptual sensitivity that is present prelinguistically.

This is not the only way in which innate concepts might be modified or overridden, though. Another possibility is that concepts or conceptual domains might have default settings that are adjusted in light of cultural input. Infants might come equipped with a conceptual or cognitive starter kit for a given domain, which is later amended or even replaced in the course of development.

Consider the following speculative possibility regarding the intuitive conceptualization of a fair distribution of resources and rewards. There is evidence that infants as young as 4 months old have intuitions about what counts as a fair distribution (Buyukozer Dawkins et al. 2019; see Chapter 18 for further discussion). However, there is also evidence for diversity across cultures in how adults respond in situations that evoke one's sense of fairness (e.g., differences concerning offers and acceptance of offers in the ultimatum game, in which one person proposes to split a pot of money a certain way and another must accept the proposal or else neither receives any of the money) (Henrich et al. 2005). One possibility here is

[3] By *representational narrowing* we mean a process in which a more general representational ability is honed in such a way that earlier representations or representational distinctions are lost or become less accessible in development.

that there are innate default settings regarding what seems intuitively fair, that these are manifest early in life, but that these settings are modified or overridden in later development as children acquire differing local norms pertaining to the allocation of resources and rewards. While there is some cross-cultural evidence bearing on whether this speculative proposal is correct, it is not sufficient to resolve the issue yet.[4] For present purposes, however, our aim is simply to illustrate that there are a variety of different ways in which a psychological structure in the acquisition base could in principle be modified or overridden. Were this to happen, in any of these (or other) ways, it could result in variation across cultures that conceals universality at an earlier developmental stage, providing another way in which variation is compatible with the existence of universality and concept nativism.

4. Rationalist learning mechanisms. A fourth complication when considering how universality bears on the status of concept nativism is that concept nativism isn't opposed to learning. On the contrary, as we were at pains to emphasize in previous chapters, concept nativism claims that learning is of central importance to conceptual development. This matters because learning often allows for different outcomes depending on the particularities of the learner's environment. This holds whether the learning mechanism is the sort of domain-general mechanism favoured by empiricists or whether it is a rationalist learning mechanism.

[4] Some support for this speculative proposal comes from the fact younger children across cultures show greater commonalities in fairness norms and sharing practices than do adults and older children, who have more divergent fairness norms and sharing practices across cultures. For example, in one study, children from six diverse cultures were presented with choices between two different types of distributions of treats between themselves and others (House et al. 2013). One choice involved deciding between a distribution in which both children would each receive a treat, and a distribution in which the child making the choice would receive two treats and the other child would receive none. Across cultures, 3- to 5-year-olds were more likely to choose the equal distribution than 6- to 8-year-olds were, even though this meant that they would have fewer treats for themselves. And while children across cultures showed a similar pattern of being increasingly less willing to choose the equal distribution rather than the one that resulted in their having a greater number of treats, substantial cross-cultural variation emerged in the choice patterns among older children (9- to 14-year-olds), with children in each culture moving towards the societal norms in their communities. In another study, children from seven diverse cultures were also asked about uneven distributions between themselves and another child—in this case, not to decide between different distributions but to accept or reject each in a series of distributions (Blake et al. 2015). Across the seven cultures, children displayed an aversion to disadvantageous inequalities (ones that favoured the other child) between 4 and 15 years of age. However, only children from three societies (USA, Canada, and Uganda) showed an aversion to advantageous inequalities (ones that favoured themselves), and this was at later ages than the aversion to disadvantageous inequality. In both studies, there was greater commonality in the behaviour of younger children in different cultures than older children and adults. However, the interpretations of these results is complicated by the fact that unlike in the Buyukozer Dawkins et al. study with infants that was mentioned in the text, participants in these two studies were making choices between distributions where they themselves were recipients of rewards, and not just third-party observers. Under these circumstances, where children's choices directly affect the rewards that they themselves will receive, their choices may well not simply be a reflection of the norms that they endorse regarding which distributions are fair ones, as self-interest may overpower an accepted moral principle (for experimental evidence that precisely this happens in children in the age ranges of these studies, see, e.g., Smith, Blake, and Harris 2013).

Suppose, for example, that children possess an innate special-purpose mechanism for learning new animal concepts (a proposal we will return to later). Being a learning mechanism, it would respond to the experiences with animals that a learner actually encounters (as well as further sources of information about animals), so that learners with different experiences should end up with different concepts. For example, learners exposed to kangaroos but not penguins should acquire kangaroo concepts, but not penguin concepts, and vice versa. And yet these learners would still share a common psychological mechanism underpinning this diversity. What is universal in this case isn't a particular innate concept or a narrow set of options (as with linguistic parameters) but an open-ended capacity for acquiring concepts of a particular type—animal concepts. Variability regarding the concepts learned may well conceal a shared, universal rationalist learning mechanism.

5. Cognitive development in degraded environments. Finally, a fifth complication in considering how universality relates to concept nativism has to do with the presence or absence of input that is necessary for a particular rationalist learning mechanism to operate properly. If the input that is required for a rationalist learning mechanism to deliver its normal output is absent in a given environment, then this could result in certain concepts failing to be acquired in this environment—despite the universal presence of the rationalist learning mechanism.

It has been well documented that children who aren't exposed to a natural language within a critical period in development have considerable difficulties with features of language that are second nature to the rest of us (Jackendoff 1994). But this fact hardly undermines the claim that there is an innate language-specific acquisition mechanism. These are just unfortunate cases in which the language-acquisition mechanism lacks some of the essential information that it needs to do its job. Something similar could happen with a rationalist learning mechanism for acquiring concepts in a given domain. For example, Atran and Medin (2008) have argued that some of the peculiarities in how contemporary city dwellers in large-scale societies conceptualize and reason about biological kinds trace back to the fact that they have relatively little engagement with the natural world and with other sources of information about plants and animals. They may still have an innate domain-specific mechanism for the conceptualization of biological kinds, but according to Atran and Medin, this mechanism's normal output is only fully realized in contexts where children grow up with a higher level of exposure to animals and plants than is typical in modern cities.

These considerations show that we need to be careful when sorting through the evidence pertaining to the argument from universality, as the relationship between concept nativism and universality is a complex one. The psychological structures in a rationalist acquisition base might not be reflected in readily

discernible universals because they give rise to disjunctive or conditional universals, because of competence masking, because they are associated with universals that are overridden in development or are subject to representational narrowing, because they involve innate domain-specific learning mechanisms that are responsive to environmental variation, or because they lack essential input. As a result, evidence of apparent cross-cultural variation may nonetheless still be best explained in rationalist terms.

The empiricist perspective that we mentioned at the start of this chapter holds that while universality is *necessary* for rationalist accounts of conceptual development to be true (and is undermined by any evidence of variation), universality is not *sufficient* for rationalist accounts to be true (since non-rationalist explanations for universality are possible). The considerations raised so far highlight the weakness of the first of these empiricist claims—the necessity claim. But the sufficiency claim also requires clarification. The rationalist view isn't that universals automatically establish the existence of corresponding innate psychological traits. Rather, rationalists and empiricists both need to ask for any given pattern of universality or variation what the best explanation is for that pattern. The rationalist view is that sometimes the best explanation is a rationalist one—an explanation that postulates innate concepts or rationalist learning mechanisms. Once the five types of complicating factors above are taken into account, we'd suggest that the evidence regarding universals supports rationalist accounts of conceptual development in a wide range of conceptual domains.

It isn't easy to show this, however—not because the argument from universality involves any kind of confusion, but because of a paucity of data available to bring to bear on specific detailed comparisons of people living in disparate societies. This poses a challenge for both claims about universality and claims about variability. For while variation is easier to establish than universality, it is often unclear what drives the variation without a broader perspective that takes into account a range of cultures. And the fact is that most of what is known about the mind is based on data from a shockingly small sample of human societies.[5] The typical participant in a psychology experiment lives in a society that is, in the terms of Henrich et al. (2010), a WEIRD society (Western, Educated, Industrialized, Rich, and Democratic). The overwhelming majority of psychological experiments have never been run in non-WEIRD societies, and the few exceptions have only looked at a handful of non-WEIRD societies. Obviously

[5] As Henrich et al. (2010) note, a 2008 review found 95% of the participants in experiments that were reported in six leading psychological journals from 2003 to 2007 were from the US, Canada, Australia, New Zealand, or Europe (Arnett 2008), and it is very likely that the majority of the remainder were from urban populations in other parts of the world. One of the concerns that Henrich et al. raise about this focus isn't just that this selection effect may distort the field's understanding of whether any given psychological trait is universal. It is that the experimental participants are overwhelming drawn from large-scale Western societies and that individuals in these societies may be outliers among the full human population.

claims that a psychological trait is a human universal are best established on the basis of a significantly broader range of data—ideally quantitative data that directly compares participants from a broad and diverse sample of societies. Unfortunately, the current evidence is highly fragmentary, it is short on quantitative assessments, and it involves different experimental procedures in different locations, making it difficult to perform direct comparisons.

Does this mean that the argument from universality is dead in the water? Not at all. Evidence from a relatively small number of different societies can still provide strong grounds for inferring universality when it is reasonable to extrapolate from the societies studied. Here it is particularly helpful if evidence can be drawn from traditional or small-scale societies. (By *traditional* or *small-scale societies*, we mean communities with low population densities, paradigmatically involving at most a few thousand people, who make their living largely in traditional ways—hunting and gathering, horticulture, and herding—and where many of the norms, practices, and traditions in the societies have not been substantially altered by contact with Western and other large-scale industrial societies; see Diamond 2013 for more on such societies.) It is also useful to focus on small-scale societies whose physical or social environment differs from that of WEIRD societies in ways that are relevant to how the conceptual capacity in question might be acquired, for example, ones where there are significant differences regarding how much caregivers talk to their children about the domain or how salient the category is for life in the community. In looking for evidence of universality, it also makes sense to focus on examples of domains where some preliminary case can be made either for universality or for a rationalist treatment of the representational capacities associated with the domain. This means taking hints from prior ethnographic studies (e.g., Brown 1991), cross-linguistic studies (e.g., Wierzbicka 1996), and from the candidates for innate features of the conceptual system favoured by other arguments for concept nativism.

We have already touched on one example that meets some of these criteria—cross-cultural studies of geometrical representation. In Chapter 1, we mentioned a body of work in which researchers tested Mundurucú adults and children regarding their intuitions about matters connected to Euclidean geometry (Dehaene et al. 2006; Izard et al. 2011). The Mundurucú are a particularly interesting population to examine in evaluating the universality of geometrical representations not only because theirs is a small-scale society. They also have no formal education in geometry and little or no experience with rulers, compasses, or maps. Thus, they provide a strong test for universality in this domain.

Recall that this research used a number of different methods and tested for geometrical knowledge of various kinds. Mundurucú adults and children were asked to estimate solutions to geometrical problems (e.g., given two angles in a triangle, indicate the size of the third angle), were asked to detect the odd one out in depictions of a geometrical property (e.g., one open figure among five other

closed figures), were asked to locate an object with a map that depicted only geometrical information about the environment (no landmarks), and were even asked direct questions about geometrical possibilities (e.g., whether more than one straight line can be drawn through two points).

On many of these tasks, the Mundurucú performed indistinguishably from Western participants.[6] As we noted in Chapter 1, this strongly suggests not only that certain geometrical concepts are likely to be universal but that a rationalist account of the origins of geometrical concepts is correct.[7] The argument here is essentially that the Mundurucú provide a particularly good case of the type of society where one would expect geometrical concepts not to be present if such concepts weren't in fact universal. And at the same time, the sorts of possible empiricist explanations for universality mentioned at the start of this chapter aren't plausible in this case. For example, learners with only domain-general learning mechanisms and low-level perceptual representations to work with could not simply gain this knowledge by noticing obvious features of the world around them. Some of the geometrical knowledge that learners exhibit in this research is concerned with entities and situations that are not straightforwardly perceptually available at all—such as the properties of idealized straight lines which go on forever or the properties of such lines in relation to different types of hypothetical worlds—one that is an endless surface versus one that is spherical. Similarly, it is very unlikely that the Mundurucú are getting this geometrical knowledge from another culture. They do not have access to any of the games, tools, or books that are commonly used to teach geometry in WEIRD societies, for example. Of course, none of this is to say that an empiricist explanation of these universals is absolutely ruled out. However, as we have noted, our aim is not to show that an empiricist account is demonstrably impossible; our argument takes the form of an inference to the best explanation. And we think that it is clear that the best overall explanation of this work is one that takes the geometrical concepts at issue to either be innate or acquired via rationalist learning mechanisms of one sort or another.

[6] This doesn't mean that there wasn't some variability too. As also noted in Chapter 1, there were some tasks where Western adults performed significantly better than Mundurucú adults, who performed comparably to Mundurucú and Western children. These differences most likely reflect the influence of formal education on the development of geometrical knowledge.

[7] We think that the geometrical representations employed by participants in these studies are uncontroversially conceptual. These representations may not have precisely the same content as geometrical concepts employed in axiomatized Euclidean geometry as taught in schools, but the thorough embedding of these representations in clearly conceptual cognitive processes leaves little doubt that they are concepts. For example, the study in which Mundurucú adults and children were asked about various geometrical possibilities required participants to interpret a verbally presented question, to entertain whether the stated situation is possible or impossible, and to reason about situations that fall well outside the scope of perceptual experience (e.g., whether a line passing through a given point can have a parallel line that also passes through the same point).

Let's consider another example, the representation of false belief. Recall that the dominant view in developmental psychology has been that children aren't able to represent false beliefs until they are 4 or 5 years old. This is the age at which they start to pass traditional false-belief tasks, verbal tasks in which they are explicitly asked about the behaviour of an agent who acquires a false belief regarding the location of an object. We saw an important challenge to this work in the context of the argument from early development. This came from researchers who, in an effort to eliminate the high performance demands in traditional false-belief tasks, turned instead to methods that allow children to display their understanding of false belief in their spontaneous responses to false-belief scenarios—a change that greatly reduced the age at which children are seen to pass false-belief tasks. But when we say *children* here, we mean children from Western urban populations, the demographic that dominates research in developmental psychology.

Fortunately, more recent research has taken spontaneous-response tasks into the field, asking whether children in very different parts of the world represent false beliefs at younger ages too (Barrett et al. 2013a). These studies involved children from diverse small-scale and traditional societies (a Yasawan community from a small village in Fiji, a Shuar/Colon community in rural Amazonia, and a Salar community in rural north-west China). Three different spontaneous-response tasks were used, each with a different methodology. Because the full impact of this work comes from the convergence of the results using these differing methods, we should take a moment to look at each of these tasks.

The first involved a verbal *preferential-looking task*—the picture-book task that we mentioned earlier in our discussion of the argument from early development. In this task, children hear a story about a character in a false-belief scenario and follow along while watching pairs of matching and non-matching pictures in a picture book. Recall that children look more at pictures that match a story, and that this presents researchers with clear evidence about how the children are interpreting what they hear. The crucial feature of this experimental procedure, once again, is that the story ends with a statement which leaves open which location she looks in, simply reporting that, after returning to the scene, the character looked for her object. Would the children look more at the picture of the character searching for the object in its original location (the false-belief solution) or the picture of the character searching in the location it had been moved to while she was away (the reality-based solution)? Children in all three of these communities looked longer at the original-location picture, indicating that they attributed a false belief to the character in the story.

The second task was a verbal *anticipatory-looking* task. In this case, there were two experimenters, E1 and E2. E1 introduced the child to a game with stickers. Having shown the child that one container was empty and the second contained a pair of scissors, E1 removed the scissors and cut out a sticker from a sheet of

stickers and gave it to the child to place on a piece of paper. As E1 was about to cut out a second sticker, E2 entered the room and mentioned that someone needed E1. So E1 replaced the scissors in the container where she had originally found them and left the room. While E1 was gone, E2 joined the child and started exploring what was in the two containers on the table, showing the child that one was empty and one contained scissors. E2 then announced that she needed some scissors for a project she was working on, and placed them in her pocket. Next, E2 looked away from the child, assumed a thoughtful pose and said (as if wondering to herself out loud), *When E1 comes back, she is going to need her scissors again…where will she think they are?* Once again, children in all three communities spontaneously indicated their understanding that the agent would suffer from a false belief, this time by looking longer in anticipation at the container where E1 had left the scissors.

The third task was a non-verbal *violation of expectation* task. In this case, the child watched a series of events that began with an experimenter (E1) sitting at a table with three brightly coloured small containers. E1 picked up the one closest to her (E1's object) and shook it, making a rattling sound. One of the remaining objects looked exactly the same as E1's object (identical object) and one looked different (different object). E1 next shook each of these other objects. The identical object didn't make any noise when shaken, but the different object rattled when shook, just as E1's object had. All of this made clear to the child that the unobvious property of E1's object—its rattling noise—was shared with the different object, but not the identical object, contrary to normal expectations. Next, E2 arrived and sat across from the child and facing the identical and different objects. E1 shook her object, demonstrating its rattling noise to E2, and asked E2 "Can you do it?". Finally, E2 grasped either the identical object or the different object and paused. Of course, anyone reading this would understand that E2 should falsely believe that the identical object produces the rattle and thus would find it unexpected if E2 grasped the different object. Evidently the children in these communities agree. They looked longer when E2 grasped the different object.[8]

We have noted that the children in all three of these societies passed these false-belief tasks well before the age at which Western children pass traditional false-belief tasks. Children's performance across these three societies was very similar. In addition, the very same experiments were conducted in the United States, allowing for a direct comparison between children from a diverse group of small-scale societies and WEIRD children (Scott et al. 2010; He et al. 2012; Scott et al. 2012).[9] The result was much the same pattern of responses across all *four*

[8] The data for this last task come from just the Salar and the Shuar/Colon communities, as the data from the Fiji community wasn't useable. See Barrett et al. (2013b) for details.

[9] Given that fewer research participants were available in the small-scale societies, a slightly wider range of ages had to be tested than in the United States. Still, the mean ages in the study were all well below the age at which children succeed on traditional false-belief tasks in the West, and the youngest

groups. But does this show that FALSE BELIEF is a universal concept, better yet an innate universal concept? This cross-cultural work certainly doesn't provide definitive proof that it is, but the evidence is very compelling.

Two points stand out. The first is simply the scope of the social and environmental differences across the four populations. The children in these three small-scale or traditional societies were growing up under significantly different conditions from one another and from the children in the United States. The second is that there are also features about life in these small-scale societies that greatly reduce children's access to the sorts of experiences that drive learning about the existence of false beliefs on non-rationalist accounts (Barrett et al. 2013a). In all three of these societies, parents are far less apt than WEIRD parents to engage in parent-child conversations that are explicitly intended to instruct their children and guide their children's development. People in all three sites also hold an outlook that grants children little personal agency, so children are rarely asked about their thoughts, feelings, and preferences. And children in all three sites typically interact with a smaller and less diverse group of social partners than is often the case in WEIRD societies. Because of these differences, these children enjoy less joint attention with parents, fewer prompts to reflect on their own mental states, and less exposure to competing perspectives. In addition, the Yasawan participants come from a part of the world (in Melanesia and the South Pacific) where there is the widespread belief that it is very difficult or impossible to know what is going on in someone else's mind and consequently that one should refrain from attributing thoughts to others unless these thoughts have been explicitly verbalized.[10] A similar outlook is present among the Shuar, who frown upon speculating about things one does not have direct knowledge of, including other people's intentions—not only is this discouraged, but it is sometimes considered a form of lying. This would of course severely limit the quantity of mental state talk in general in the community that children would be exposed to.

Thus there are major obstacles for learners in these small-scale societies under the assumption that general-purpose learning would have to rely on mental state discourse and related cultural practices to obtain the very idea that there are such things as false beliefs. Given the difference in both the quantity and quality of this linguistic input and related culturally mediated support for thinking about other minds, empiricist models ought to predict a great deal of variability regarding *when* children are able to pass these sorts of false-belief tasks under these different living conditions. But what we have seen is that there is no evidence of such

children tested were as young or younger than the youngest among the US children. Analyses comparing the different populations also found no effect for age in these experiments.

[10] Anthropologists who have documented and commented on this outlook refer to it as *the doctrine of the opacity of other minds* (Robbins and Rumsey 2008).

variability across this range of societies—children perform similarly in all of the tested societies—suggesting that the ability to represent false belief is itself innate or is acquired via a rationalist learning mechanism.[11]

The key to the argument from universality is being able to extrapolate from data concerning people's conceptual abilities in a limited number of different communities. This means looking for cases where children develop more or less the same concepts even when they have strikingly different opportunities to learn about the domains in which these concepts figure. In the examples we have looked at so far, the focus has been on cases where social facts about small-scale societies entail that children will have less access to relevant information than WEIRD children. But sometimes things go the other way—WEIRD children are the ones with less access.

Consider the conceptualization of death. In urban Western communities, children are typically insulated from experiences with death. They see few human corpses, few dead animals, and their experience with death in connection with food and cooking is highly sanitized—to the point where there is no clear link for many young children between meat and the animals it comes from. In contrast, in traditional small-scale societies, children often witness animals dying and being dismembered, gutted, or skinned, and may be allowed or encouraged to "play" with small animals in a way that often leads to the animals' death.

The conceptualization of death is also an especially interesting case study because, despite these cultural differences, beliefs about some form of afterlife are commonplace. So whether children are kept from having experiences related to death or not, they often hear about dead ancestors continuing to exert an influence on current affairs, dying relatives passing on to live in paradise, and so on. This way of talking about death is clearly at odds with the thinking that death is final; hence it should interfere with a biological understanding of death as the total cessation of biological processes.[12]

[11] Some cross-cultural studies, employing traditional false-belief tasks, report greater variation cross-culturally (compare Callaghan et al. (2005) and Mayer and Träuble (2013). However, this isn't particularly surprising given the heavy performance demands associated with traditional false-belief tasks (see Chapter 9). In fact, an important recent study shows that much of the individual variance in traditional false-belief tasks may be explained by one key feature—the fact that such tasks require participants to overcome the "curse of knowledge". The curse of knowledge is a cognitive bias in which people find it difficult to put aside what they know about the world when considering what someone else may know. Variations on traditional false-belief tasks that do not require overcoming the curse of knowledge are as readily passed by 3-year-olds as older children (Ghrear et al. 2021), suggesting that much of the variance on traditional false-belief tasks, both individual and cross-cultural, may be explained by variance in the ability of children to overcome the curse of knowledge. By using non-traditional false-belief tasks, Barrett et al. (2013a) were able to limit the impact of the curse of knowledge and other kinds of competence masking factors.

[12] Developmental psychologists who have studied Western children have thought that death is a particularly hard concept for children to grasp because it is also easily confounded with sleep and requires disentangling the categorical distinction between animate/inanimate and alive/dead. When an animal dies, it loses its animacy, but ordinary inanimate objects (e.g., keys, cups) aren't dead (Carey 1985).

Notice, though, that an evolutionary perspective suggests that there ought to have been selection pressure in ancestral environments for children to rapidly develop an understanding that death terminates an animal's agency. Consider the simple question of what to do if one encounters an inert wild animal, such as a wolf or a coyote. If it is dead, approaching it is an option. But if it is still alive, then a high degree of caution is in order. It may be dangerous or even deadly to approach. There is also a major asymmetry in the cost associated with errors in these circumstances. To mistakenly suppose that a dead animal is alive is a small mistake compared to assuming that a live animal is dead. This line of reasoning led Barrett and Behne (2005) to set out to determine whether children are endowed with an innate special-purpose system that responds to cues of death and that uses these to categorize an entity according to whether it is a living agent or whether it is dead. This categorization would act like a cognitive switch. It would have the default setting that favours the living/animate setting (to err on the side of safety), but once the switch is thrown and an item comes to be categorized as dead, all the former inferences associated with agency would be blocked (e.g., the item would no longer be taken to have the power of engaging in goal-directed behaviour).

To test this hypothesis, these researchers examined 3- to 5-year-old children's inferences about agency-related properties in scenarios involving an animal that falls asleep or dies.[13] The children were asked: *Can it be afraid? Could it move if you touched it? If you walked by, could it know you were there?* Because the motivation behind this study was the hypothesis that the living/dead discrimination system is a human universal, the populations chosen for comparison were ones that have markedly different exposure to animals and death and whose children face stark differences regarding the danger that animals pose in day-to-day life. One was an urban Western population (children from Berlin), the other was a hunter-horticulturalist population living in the forest (Shuar children from the Amazonian region of Ecuador). The result was a remarkable correspondence despite these differences—a developmental trajectory in which, by 4 years of age, children in both populations appreciate that death but not sleep blocks normal agency inferences.[14] This finding is hard to explain on the assumption that this distinction is owing to children's experiences of death or what they hear about death. Why would it develop in the same way, and at the same time in childhood, across these populations? On the other hand, it makes perfect sense if the acquisition base includes a living/dead discrimination system that links representations

[13] It is worth noting that, while this work is with children, the logic of the argument that we are giving is not that of an argument from early development.

[14] For further supporting cross-cultural work on the conceptualization of death, see Astuti and Harris (2008) and Barrett et al. (2021).

of life and death with inferences suitable to animate and non-animate biological entities.[15]

Let's look at one more example of the argument from universality. Evidence from both anthropological surveys (Brown 1991) and cross-linguistic analyses (Wierzbicka 1996, 2015) suggests that basic logical concepts—such as OR, ALL, and NOT—are human universals. However, languages differ in important ways in how they express these concepts. Consequently there are interesting cross-cultural differences in the linguistic and social environments that children experience when learning the meaning of logical terms. We will focus here on disjunction (the concept OR) and the interpretive problem for children learning the language of disjunction and how it is related to other linguistic constructions involving logical concepts (e.g., negation and quantification). Our argument is based on research that has led linguists to conclude that disjunction takes a particular universal form in natural language. But to anticipate the claim we wish to make, it isn't about what may (or may not) be universal among adult speakers. It is about what is universal for children. Our claim is that children universally interpret disjunction in natural language in just one way despite considerable evidence from adult speech for contrary interpretations.[16]

A little background on the form that disjunction takes in language is necessary to understand how the argument from universality works in this case. There are two possibilities to consider. Disjunction could be interpreted inclusively or exclusively:

Inclusive disjunction (inclusive-or)
A or B is true when A is true but B isn't, when B is true but A isn't, and when both A and B are true.

Exclusive disjunction (exclusive-or)
A or B is true when A is true and B isn't, and when B is true and A isn't, but not when both A and B are true.

Interestingly, although there are good reasons for supposing that disjunction takes the inclusive form in natural language, this isn't always obvious from

[15] What about younger children? Although the 3-year-olds in this study weren't reliable in distinguishing the impact of sleep from death, we suspect that the test wasn't sensitive enough to get at the youngest age at which the living/dead discrimination system first appears. A more sensitive test would require providing children with the most reliable cues of death that would have been available in ancestral environments, for instance, seeing an animal being torn apart or seeing the whitened face of a human corpse. For understandable reasons, stimuli of this kind are off limits and we have to be content with studies employing less vivid cues.

[16] Our discussion of this example follows the analyses in Crain and Khlentzos (2010) and Crain (2012), who use a version of the argument from universality to argue for a position they refer to as *logical nativism*. This is the view that "humans have an innate logical faculty that structures thought and assists in the acquisition of language" (Crain and Khlentzos 2010, p. 31).

patterns in adult speech. In fact, there are some systematic patterns in adult speech that lead to the appearance that disjunction should be interpreted exclusively (we will mention a few of these in a moment.) This situation constitutes a major challenge for young language learners on the assumption that language learning is restricted to domain-general learning. The evidence from adult speech should cause children to opt for the exclusive interpretation. On the other hand, if children universally reject this interpretation despite the evidence in its favour, this suggests that they approach language learning with an innate disposition or bias to interpret natural language disjunction inclusively. And if this is how language acquisition proceeds, it would suggest that the logical concepts that stand behind this universal feature of language learning are innate concepts and, in particular, that inclusive disjunction is part of the acquisition base.

Now consider the way English-speaking adults use the word "or". In English, it is natural to interpret the sentence *Alice ate eggs or cereal* as implying that Alice ate one but not the other—suggesting that "or" isn't understood as inclusive disjunction. However, the reason that simple sentences of this sort are treated in this way is that it is pragmatically odd to say *Alice ate eggs or cereal* when she ate both. If Alice ate both eggs and cereal, then a speaker who said she ate one *or* the other wouldn't be complying with the pragmatic principle that requires speakers to be cooperative and hence as informative as possible; a cooperative speaker who knew that Sue ate both should say that Sue ate the eggs *and* the cereal, even if it is technically true (on the inclusive reading of "or") that Sue ate the eggs or the cereal. However, children below the age of 6 or 7 aren't in a position to appreciate that it is misleading to talk in this way, because children this young have a poor grasp of pragmatic implicature. Accordingly, we should expect them to treat the adult usage as directly reflecting the meaning of disjunction, and interpret "or" exclusively. Yet children learning English do *not* do this. They happily accept a description like *Alice ate eggs or cereal* when Alice ate both. That is, they adopt the inclusive interpretation of "or" even when adults, to all appearances, do not. In this way, English-speaking children seem to be acquiring inclusive-or *in spite of* their adult models rather than because of them.

English is of course just one language. To establish that the inclusive interpretation of disjunction is universal, we have to look at other languages, especially languages that are typologically remote from English and from one another. Chinese and Japanese are useful languages to investigate with these criteria in mind. And they are especially pertinent for studying how children interpret disjunction in language. This is because the fact that disjunction is inclusive-or in Chinese ("huozhe") and Japanese ("ka") is even less transparent than in English, making them especially difficult test cases for the universality hypothesis.

To see this, it is useful to consider another feature of inclusive disjunction, namely, the fact that a simple negated disjunction entails a conjunction of negations. From *Alice didn't eat eggs or cereal*, it follows that *Alice didn't eat eggs* and

that *Alice didn't eat cereal*. Adult English speakers readily accept these entailments. However, adult Chinese and Japanese speakers do not appear to accept the equivalent entailments in Chinese and Japanese (even though speakers of these languages are in fact employing inclusive-or in these cases), making it look even more as though such speakers do not interpret disjunction inclusively.[17] This means that young Chinese and Japanese children are regularly exposed to adult speech that appears to treat disjunction as exclusive-or. And so it might be expected that the children follow suit and—at least initially—interpret disjunction in Chinese and Japanese as exclusive-or. However, they don't opt for the exclusive-or interpretation. Instead, they interpret disjunction just as English-speaking children do. For example, given a situation in which Alice has eaten eggs but not cereal, 5-year-old Japanese speakers reject the description *Alice didn't eat eggs or cereal* (just like English speakers)—even though Chinese- and Japanese-speaking adults are happy to accept (the Chinese or Japanese equivalent of) this description. Here, too, children interpret disjunction as inclusive-or *in spite of* their adult models rather than because of them.

The situation, then, is that children growing up in diverse linguistic environments converge on a single interpretation of disjunction in language—the inclusive-or interpretation—even though adult speakers, for different reasons and to varying extents, obscure the fact that this is the right interpretation.[18] Thus we have strong evidence for a linguistic and conceptual universal among young language learners. Moreover, this is a universal of precisely the type that can't be explained in terms of empiricist general-purpose learning. Given the input that children are faced with, a general-purpose learning mechanism should settle on the exclusive-or interpretation (at least initially).[19]

[17] Notice that even if this *weren't* the case—if instead, adult speakers of Chinese and Japanese weren't using inclusive-or—that this argument would still be strong (in fact, arguably even stronger). This is because it would be even better evidence for inclusive-or being innate and associated with language if children learning languages where the adults were using exclusive-or nevertheless themselves interpreted these languages as using inclusive-or.

[18] The inclusive-or interpretation has also been found in children speaking Turkish (Geçkin et al. 2017), Catalan (Pagliarini et al. 2021), French and Hungarian (Pagliarini et al. 2022), and to some extent among Italian children, though for Italian-speaking children, the situation is somewhat more complicated (Pagliarini, Crain, Guasti 2018; Pagliarini, Lungo, et al. 2022).

[19] If adult speakers of Chinese and Japanese don't appear to accept the entailments that go with understanding disjunction as inclusive-or, why have linguists argued that this is not because they don't understand disjunction as inclusive-or? The reasoning here is complex, but we will sketch the core idea. (It should be kept in mind, however, that, as we noted in footnote 17, our case is even stronger if linguists are wrong about this.) The reasoning relies on the background hypothesis that Universal Grammar has a parameter for whether disjunction is a *positive polarity* item (which by definition must take scope over negation) (Crain and Khlentzos 2010). The idea of the scope of a term comes down to how much of the sentence the term applies to, so when one term has scope over another, the first applies to more of the sentence, and includes that part of the sentence that the second applies to. An example of a positive polarity item interacting with negation in an English sentence will help to clarify what this means. In English, "some" is a positive polarity item. *Sue did not touch some of the toys* means that there are *some* toys that Sue did not touch, not that it is *not* true that

Earlier in this chapter, we noted that the argument from universality faces a major practical challenge stemming from the current paucity of cross-cultural data, especially the paucity of data for small-scale societies and other non-WEIRD populations. We have argued that a strong case can nonetheless be made for universality in some of the domains where we do have cross-cultural data. The examples we have looked at to illustrate this have been geometrical concepts (comparing Mundurucú adults and children to Western adults and children), FALSE BELIEF (comparing Yasawans, Shuar, Salar, and Western children), DEATH (comparing Shuar children to US children), and certain logical concepts (comparing Chinese-, Japanese-, and English-speaking children). While further examples could be discussed, these should suffice to illustrate the argument from universality. Of course, as we remarked earlier, it is certainly possible for a psychological trait to be universal without being innate. The question, though, isn't whether this is possible. Rationalists agree with empiricists that universality doesn't necessarily argue for a rationalist account. The real issue is, and always has been, what the best type of explanation is of a universal trait's origin.[20]

We have argued that at least in these particular instances of universality, there are good reasons to think that the best explanation is, in fact, a rationalist one. In each case, there are key facts that suggest that the universal isn't the product of general-purpose learning and that it depends upon certain innate concepts or rationalist learning mechanisms. For example, with the conceptualization of death, there was the fact that Western urban children and Shuar children live in conditions with substantially different experiences of death. The Shuar children routinely witness and even participate in the killing of animals and face the very

there are some toys that Sue did touch. Even though "some" occurs later in the sentence than "not", it is interpreted as applying more widely in the sentence than "not" does. Crain and Klentzos claim that, just as "some" is a positive polarity item in English, disjunction (inclusive-or) is a positive polarity item in Chinese and Japanese. This means that the Chinese or Japanese equivalent of *Alice didn't eat eggs or cereal* gets interpreted as something like *Either Alice didn't eat eggs or Alice didn't eat cereal* (with the "or" applying more widely than the "not", even though "or" occurs later in the sentence than "not"). This explains why it doesn't entail that *Alice didn't eat eggs* and that *Alice didn't eat cereal*. But why think that disjunction is a positive polarity item in these languages? Part of the evidence for this is that, if disjunction is a positive polarity item, then the conjunctive entailment should go through when the negation occurs in a higher clause than disjunction (because the scope of the disjunction will be limited to the clause it is in). And it does. For example, from *Jen didn't say Ted ordered pasta or sushi* (in Chinese and Japanese) it follows that *Jen didn't say Ted ordered pasta* and *Jen didn't say Ted ordered sushi*. As we noted earlier, however, nothing in our argument turns on this being the correct interpretation of adults.

[20] It is also worth repeating that concept nativism allows for traits that are not universal if only because of the complications mentioned earlier—conditional and disjunctive universals, competence masking, innate traits being modified or overridden, rationalist learning that produces context-sensitive outcomes, and the possibility of abnormal development in degraded environments. What this means is that, just as empiricist accounts need to be considered in cases of universality, rationalist accounts need to be considered in cases of variability.

real threat of being killed by animals themselves. In contrast, the Western urban children (from Berlin) have little experience of death, an often highly sanitized understanding of where animal products come from (meat, leather, etc.), and are unlikely to face any danger from nearby animals. And yet the Western children seem to follow the same developmental trajectory as the Shuar children regarding their inferences about death versus sleep. This striking fact is inexplicable if children's understanding of death forms on the basis of general-purpose learning, for example, learning that combines personal observations of dead and dying animals with adult testimony. But this fact is to be expected if children possess an innate domain-specific living/dead discrimination system. The other examples we have discussed (involving geometrical concepts, mental state concepts, and logical concepts) follow much the same pattern. In each case, the crucial fact isn't just universality. It is an instance of universality that is best explained by a rationalist acquisition base.

In sum, although there is a need for far more research across cultures, and cross-cultural research on a wider range of conceptual capacities, there is already enough evidence to say that the argument from universality makes an important contribution to the case for concept nativism.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0011

12
The Argument from Initial Representational Access

Our fourth argument for concept nativism is based on the idea that learning concepts in certain conceptual clusters requires an initial representational foothold in the conceptual domain—in order to learn concepts in the cluster, the learner must already have some representations pertaining to the conceptual domain or a semantically nearby domain. The form of argument here is closely related to the argument that we gave in Chapter 5 against Locke's attempt to explain the origins of all general representations. That argument showed that if a learner only has access to non-general representations (representations of particular entities *as particulars*), they will not be able to learn general representations through a process of abstraction. We then argued that this argument was not limited to abstraction but similarly applied to essentially any learning process with only these resources. The core problem was that learning mechanisms with access to only non-general representational resources could not bridge the gap to learning a fundamentally different type of representation—general representations. In order to learn general representations, the learner must already have some general representations. The argument from initial representational access argues that for concepts in certain conceptual clusters, there is a similar representational gap that can't be crossed without already possessing an initial representational foothold in the conceptual domain at issue.

This type of argument has a number of precursors in the historical rationalism-empiricism debate. One is Kant's claim that there are certain ways of representing the world that are so basic that they constitute preconditions for cognition and couldn't possibly be the outcome of a learning process (Kant 1781/1998). Most famously Kant argued that space is the form of outer sense (i.e., objects must be seen as occupying some position in space) and that time is the form of inner sense (i.e., experiences of sensations and feelings must be temporally ordered). In a similar spirit, but with an interest in accounting for the variety of ideas that are expressed in natural language, Leibniz proposed there must be an "alphabet of thought" that forms the basis for all human concepts and that includes "only those that which are truly necessary for defining all the others" (Wierzbicka 2001, p. 233). Leibniz's argument for a rationalist account of this alphabet was that understanding a concept by understanding a definition only goes so far. At some point, we reach a set of abstract concepts where any attempted definition

invariably leads to a vicious circle. His conclusion was that these concepts are innate and inherently understood—that is, understood without needing to grasp a definition—and that these are the elements from which all other concepts are fashioned.

These and other rationalist arguments in the history of philosophy give expression to the core idea behind the argument from initial representational access, that empiricist theorizing about the origins of concepts hasn't fully recognized the extent of the challenge that empiricism faces. The arguments point to the existence of what might be called *foundational* conceptual clusters—clusters containing sets of interrelated concepts where some of the concepts in question can be acquired if others are already possessed, but where the full set of concepts cannot be learned solely on the basis of concepts and representational resources from outside the domain. Instead, at least some concepts in the conceptual cluster must be innate or acquired via rationalist learning mechanisms.[1] Prime candidates include such concepts as those involved in the representation of basic logical and numerical content, space and time, agency, causality, modality, and the sorts of categories that make up the primary metaphysical distinctions embedded in common-sense thinking and human language, for example, representations for events, objects, substances/stuffs, individuals, properties, and kinds.

In order to spell out the argument from initial representational access more explicitly, we want to distinguish two related problems that are at the heart of the argument. The first—*the why problem*—is the problem of explaining *why* an empiricist learner who does not possess any representations in the domain in question would be prompted to formulate concepts from this domain. The second—*the how problem*—is the problem of explaining *how* an empiricist learner would be able to formulate these concepts given her prior ways of representing the world.[2]

The difficulty associated with the first problem is that the experiences an empiricist learner is limited to wouldn't push her in the right direction. This point can be hard to appreciate because any theorist considering these issues will already *have* the concepts in question, making it difficult to see them as anything other than a natural outcome of normal human experience. To fully comprehend

[1] The term *foundational conceptual cluster* is just intended to provide a readily available way of referring to conceptual clusters where the argument from initial representational access is taken to apply. Whether a given conceptual cluster is a foundational conceptual cluster, and indeed whether there are any such conceptual clusters, is not a matter of stipulation, and would need to be argued for in any given case. We will also refer to *foundational concepts* (concepts in foundational conceptual clusters) and *foundational domains* (content domains that foundational conceptual clusters are directed at or about).

[2] We are using the term *empiricist learner* as shorthand for learners as empiricist models conceive of them, in particular, learners who are constrained by the sort of acquisition base that is postulated by empiricist theories of concept acquisition (see Chapter 2).

the force of the why problem, you have to make a concerted effort to imagine *not* having these concepts or any representations with related content. You have to imagine that all you have to begin with is the empiricist acquisition base, with its general-purpose mechanisms and its minimal stock of sensorimotor representations.[3] Given this austere starting point, what sorts of experiences would prompt the new way of thinking?

Take the concept BELIEF. To acquire this concept requires being able to see the world in terms of hidden causes with representational properties that might account for people's behaviour. Belief attributions depend on inferences that are grounded in external cues of various kinds. For someone who already knows about mental states, some of these cues will readily lead to thoughts about the other person's inner life. But what about a learner who has none of this to begin with—no prior representation of beliefs or belief-like states, no understanding that people's behaviour is caused by hidden inner states, and so on? All this learner has to go on is people's behaviour itself, represented purely in terms of low-level perceptual representations of things like their facial expressions, their posture, and their movements.[4] As Leslie (1987) has remarked, "it is hard to see how perceptual evidence could ever force an adult, let alone a young child, to invent the idea of unobservable mental states" (p. 422).

Granted, learners in this situation may find themselves in a social world in which it is difficult to predict what others will do. They may even grasp that they are in this situation and look for variables that will help them to better understand people's behaviour. But given that, by assumption, they are unable to represent mental states as such and instead can only represent properties and relations that are perceivable, the needed variables would be exceedingly peculiar. It's one thing to have a sense that your current ways of representing the world are inadequate, and quite another to have an inkling about a wholly new way of conceptualizing the situation. What in this learner's experience would ever lead her to begin to think about the matter in terms of *unobservable causes*, better yet unobservable causes with *representational* properties? Notice, as well, that the fact that an empiricist learner herself has such states surely isn't enough. Many animals have beliefs (or other representational mental states) too but no inkling that

[3] As we develop the argument from initial representational access, we will often just refer to these representations as *perceptual representations*. However, it should be kept in mind that these should be understood to be low-level perceptual representations (e.g., ones that generally pick out sensory properties and not higher-level properties like object kinds) and that empiricists make use of motor representations as well.

[4] People also talk about mental states and use special constructions to do so (e.g., words that take sentential objects). But in order to make the case that this could assist an empiricist learner in acquiring a conception of mental states as such, one must address the question as to how such a learner would ever be in a position to interpret these words properly. Moreover, as we saw in Chapter 9, children are competent with mental state concepts at ages well before their abilities with these parts of language are in place (another example of arguments for concept nativism interacting).

BELIEF (or MENTAL REPRESENTATION) is critical to accounting for the actions of other agents.[5]

The *how problem* takes this problem one step further. It asks how a motivated empiricist learner could even formulate the target concepts to begin with. Sticking to BELIEF for the moment, the task might be tractable if children innately possess some basic elements of common-sense psychology, for example, if they innately have the ability to represent a few types of mental states and had innate knowledge of some paradigmatic circumstances in which they occur. This might give children the basic tools needed to posit that there are unobservable representational states and provide the scaffolding necessary for learning more about such mental states and others like them. But without any prior ability to represent and reason about mental states at all, there is an enormous gap between where children begin and where they end up.[6] How exactly would an empiricist learner bridge this gap?

Much the same point applies to other foundational conceptual clusters outside of folk psychology. Of course, there needn't be a deep puzzle about how new concepts are learned when a learner can draw upon directly related representational abilities. If you already know about kin and about properties and relations that are central to kinship categories (e.g., male/female, parent/offspring), then acquiring a new kinship concept, like GRANDMOTHER, needn't be all that mysterious. The same goes for acquiring a new financial concept, like ANNUITY, assuming you are already familiar with contracts, money, and insurance companies. This isn't to say there aren't important questions about how these concepts are acquired even then. But most theorists would agree that such new concepts can be formulated by using older concepts in new ways and by incorporating a bit of new information. In the simplest of cases, this may take the form of a definition (e.g., *x is y's grandmother just in case x is a female parent of one of y's parents*). But the how problem becomes pressing when this process depends on access to concepts that cannot themselves be acquired in this way or otherwise acquired

[5] Some empiricists might respond to Leslie by claiming that empiricist learners can access their own beliefs through introspection. But while introspection may tell us something about our own experience, it's not clear that beliefs as such are introspectable (Carruthers 2011). Moreover, much the same problem arises with respect to introspection in any case. To the extent that we are able to understand ourselves as possessing beliefs through introspection, the question arises as to how this is achieved. If we don't already have a way of representing beliefs as such, it isn't clear how introspection would give us a way of representing the contents of introspection *as* beliefs. But if we already possess the ability to represent beliefs as such, then introspection wouldn't explain where BELIEF comes from. The capacity to introspect the contents of introspection as beliefs seems to presuppose the very representational ability whose origins it is being asked to explain.

[6] It may be that a learner doesn't necessarily require concepts of other representational mental states in order to acquire belief, and that this concept can be acquired on the basis of closely related concepts, such as the concept of an unobservable cause and the concept of a representation. But this just means that these concepts, particularly REPRESENTATION and BELIEF, are part of a tight cluster of interconnected concepts, and the problem is much the same for acquiring REPRESENTATION as it is for BELIEF.

through the sorts of domain-general learning mechanisms available to an empiricist learner.

Taken together the why problem and the how problem form the backbone of the argument from initial representational access. For an important selection of conceptual domains, these problems mean that it is all but impossible for an empiricist learner to acquire concepts in the domain in question.[7] But since humans do reliably acquire representations in these conceptual domains, this means that we aren't empiricist learners. The concepts must be innate, and so acquired non-psychologically, or else acquired via a rationalist learning mechanism on the basis of other concepts or representational resources in the relevant domain. As we will show, this argument poses a serious challenge for any thoroughgoing empiricist account of concept acquisition.

In order to appreciate the full strength of the argument from initial representational access, it will be useful to examine some of the main ways that empiricists have responded to the challenges that the why problem and the how problem present. We will focus on one particularly important type of response—and a cornerstone of much empiricist theorizing—which addresses the how problem with a two-pronged *reduce-or-eliminate* strategy.

Reductive semantic programmes of one kind or another have always been central to empiricist models of psychological and linguistic representation and continue to be important in contemporary theorizing.[8] These can take different forms, but, in the ideal case, they are guided by the assumption that all concepts are ultimately composed from a stock of semantic primitives that originate in perceptual systems and are restricted to low-level perceptual content. As a result, it can be held that no matter how far removed from perception any concept may appear to be, its content is exhausted by things that can be perceived. Locke provides an early and especially clear example of this view:

> if we warily observe the Originals of our Notions, that even *the most abstruse Ideas*, how remove soever they may seem from Sense, or from many operation of our own Minds, are yet only such, as the Understanding frames to it self, by repeating and joining together *Ideas*, that it had either from Objects of Sense, or from its own operations about them: So that those even large *and abstract* Ideas *are derived from Sensation, or Reflection*, being no other than what the Mind, by the ordinary use of its own Faculties, employed about Ideas, received from

[7] In particular, we have in mind here an empiricist learner possessing an austere initial representational starting point in relation to the representational domain (i.e., one with no innate representations and no innate domain-specific learning mechanisms pertaining to the domain in question).

[8] To be clear, we are not claiming that empiricists are required to adopt a programme of this sort, only that empiricists have frequently done so in an attempt to deal with the sorts of problems that are highlighted by the argument from initial representational access.

> Objects of Sense, or from the Operations it observes in it self about them, may, and does attain unto. (Locke 1690/1975, p. 166)

This type of semantic view sets the stage for the extensive use of the Acquisition by Composition model of conceptual development (the ABC model; see Chapter 5). This model claims that learning a concept is a matter of composing it from its semantic constituents.[9] Notice that if all concepts are ultimately composed of the same stock of low-level perceptual representations, then acquiring any complex concept is largely a matter of settling on the right combination of perceptual representations. These combinations may come about in a number of different ways. Some may result from processes that track perceived correlations (e.g., seeing that a certain shape regularly appears with a certain colour), some may be mediated by language (e.g., associating perceptual properties that are jointly singled out by a given phrase), while others may be randomly or creatively formed combinations that prove to be useful or interesting (e.g., combinations that take place in imagination). For a theorist who relies on the ABC model to support an overall empiricist approach to conceptual development, it doesn't matter which of these specific processes is behind the acquisition of any given concept. What does matter is that every concept is either a low-level perceptual representation (which is simply produced by the activation of a perceptual system without learning) or a complex representation (acquired by composition in a way that reflects the correct semantic reductive analysis of the concept). Either way, there would be no need to postulate a richer stock of innate representations. The innate representations would be confined to the semantic primitives that go hand in hand with an empiricist version of the ABC model.

But what if a concept can't be acquired in this way because it lacks the needed semantic structure? What if it isn't a perceptual semantic primitive itself but also isn't exclusively composed of perceptual primitives? This is where the eliminativist strategy kicks in. Many empiricists would say that this situation is reason enough to conclude that the concept in question is spurious. The classic source for this move is Hume's highly influential empiricism about concept acquisition, which epitomizes how the eliminativist strategy can be used in tandem with the reductive strategy. (We should note that although we are presenting some of these strategies and ideas in terms of their classic philosophical sources, the core moves

[9] Although we ourselves argue that the ABC model is not mandatory and suggest other ways a new concept can be learned (see especially Chapter 25), we don't think that these alternatives provide empiricists with a way of circumventing the argument from initial representational access. For the sorts of foundational concepts at issue, the why problem remains pressing—learners would still need the innate representational abilities that would position them to recognize the need to push into the new representational domain. And even if this problem could be overcome, alternative acquisition mechanisms for the sorts of foundational concepts at issue are likely to require considerable innate domain-specific resources.

are ones that—as we will see shortly—are directly relevant to, and very much alive in, contemporary philosophy and cognitive science.)

For Hume, all complex ideas are decomposable into simple ideas that are copies of simple impressions.[10] This psychological commitment is at once the basis of his critique of philosophical views he deemed problematic and also the basis for his positive proposals regarding the true content of contentious philosophical terms. On the critical side, Hume claimed to demonstrate that, for certain philosophical terms, a rigorous analysis of their previous use reveals that they don't live up to the semantic reductive constraint that all concepts must abide by. On the positive side, he proposed alternative definitions that could provide these terms with legitimate content, backing them with ideas that he took to satisfy his reductive constraint.

The paradigmatic example of these empiricist themes is Hume's treatment of the conceptualization of causality, particularly the supposition that a cause possesses a force or power in virtue of which it brings about its effect, or, as Hume put it, that there is a *necessary connection* linking cause and effect (Hume 1748/1975). Hume argued that NECESSARY CONNECTION can't be grounded in perception. When we see one billiard ball crash into another, we can't literally perceive the necessity linking the first ball's impact with the second's motion, or perceive any hidden force or power acting on the second ball. All we can see is the contact and the subsequent motion. Thus, for Hume, the idea of a necessary connection between objects is problematic. It is an idea for which there is no corresponding impression.[11]

At the same time, Hume held that there is a positive, alternative theory that does conform to the semantic reductive constraint in which ideas inherit their content from impressions. On this approach, we first experience Xs being repeatedly followed by Ys, which gives rise to a psychological disposition in which thoughts of Ys follow thoughts of Xs. This disposition, in turn, gives rise to a *feeling of expectation* that Ys follow Xs. It is this feeling that is the impression for the idea that there is a necessary connection between Xs and Ys. For Hume, then, there is nothing more to this idea, and as a result he concluded that, when properly analysed, the idea doesn't concern anything involving a necessary

[10] For Hume, *ideas* are the representations involved in reasoning, while *impressions* are the representations involved in sensation and affect. His famous copy principle grounds the content of all ideas in experience by requiring that simple ideas inherit their content from corresponding simple impressions: "All our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent" (1739/1978, I.i.1).

[11] We are simplifying Hume's argument somewhat. He also considers the possibility that the idea of necessary connection traces back to impressions of internal causes and their effects, such as an impression of the conscious intention to move an arm (cause) leading the motion of the arm (effect) (Hume 1748/1975, section VII, part 1). Hume claims that this proposal fares no better, since we can't experience the power of the will to bring about the arm's motion. Rather, we form impressions of the will's intentions but have to learn through experience that these are followed by arm movements and the like.

connection between objects or a hidden power or force (there is no such idea). Instead, it is just the idea of a feeling of expectation that is mistakenly projected onto external objects.

In short, Hume's reductive analysis is meant to show that there needn't be any obstacle to an empiricist learning model for an idea that we refer to as "necessary connection" provided that we are willing to accept what many would regard as a fairly radical change in the content we assign to this idea. On Hume's account, it is an idea or concept whose content is exhausted by the internal feeling of expectation that Ys follow Xs, a feeling that is acquired by regularly observing many Ys following Xs.

Many empiricists have employed much the same overall reduce-or-eliminate strategy, often underestimating the difficulty of achieving successful reductions.[12] We will just consider one further example here, one which vividly illustrates just how far empiricists have been willing to push this strategy. The example is Carl Hempel's (1935/1980) treatment of ordinary psychological concepts and his claim that these are perfectly consistent with a behaviourist psychology.

Hempel proceeded by assuming that an analysis of a psychological statement should be one that preserves its meaning yet doesn't use any other psychological language, and by assuming that the meaning of a statement is to be given in terms of the objective physical conditions that could be used to verify the statement. So if physical conditions of verification could be provided for a statement like *Paul has a toothache*, this would at once articulate its meaning and demonstrate that ordinary ways of talking and thinking about mental states such as pain are consistent with physicalist scruples. On the other hand, if the needed physical verification conditions couldn't be provided, then, for Hempel, this would show that it is really a pseudo-statement, "a sequence of words correctly constructed from the point of view of grammar, but without content" (p. 17).

Hempel's (1935/1980) analysis of the statement *Paul has a toothache* included five proposed verification conditions (p. 17; list renumbered from original):

[12] Further philosophical examples include Locke's discussion of the sensory basis for concepts of space, time, and infinity (Locke 1690/1975), and numerous reductive claims in the writing of the logical empiricists (see, e.g., Carnap 1932/1959; Ayer 1946/1952). This strategy is not restricted to philosophy, however. Much the same reduce-or-eliminate strategy drives a substantial body of empiricist work in cognitive science. For example, in psychology, semantic reductive programmes have been coupled with empiricist theories of concept acquisition that are committed to some form of lexical conceptual structure—often either the classical theory of concepts (in which lexical concepts encode definitions of the items that fall under them) or the prototype theory (in which lexical concepts encode statistical properties of the items that fall under them) (Laurence and Margolis 1999). Similar views are also prevalent in the embodied cognition literature, especially among those who claim that higher cognition is a form of sensorimotor simulation. Theorists who hold this perspective often take it as an important challenge to show that even highly abstract concepts (e.g., TRUTH and OUGHT) are grounded in sensorimotor representations (see, e.g., Lakoff and Johnson 1999; Boroditsky and Prinz 2008; Glenberg 2015). For discussion of embodied cognition in relation to the rationalism-empiricism debate about the origin of concepts, see Chapter 22.

1. Paul weeps and makes gestures of such and such kinds.
2. At the question "What is the matter?", Paul utters the words "I have a toothache".
3. Closer examination reveals a decayed tooth with exposed pulp.
4. Paul's blood pressure, digestive processes, the speed of his reactions, show such and such changes.
5. Such and such processes occur in Paul's central nervous system.

But notice that this analysis is inadequate precisely because it tries to get by without mentalistic concepts. This becomes clear when one considers the qualifications needed to make some of his claims have any degree of plausibility. It simply isn't true that a person with a toothache will utter "I have a toothache" to the question in 2 unless that person *understands* English. Even then, he needn't utter those exact words. He might instead use different words that *express a similar thought* (e.g., "my tooth hurts" or "I have a lot of pain in my tooth"). And he needn't even use words that express any of these thoughts if he doesn't *want* to respond sincerely or doesn't *want* to explain what is going on with him at that moment. The problem Hempel faces is that, while one might try to amend his proposal by inserting qualifications along these lines, doing so would incorporate mentalistic elements into the analysis—in which case his analysis wouldn't be fully reductive after all.

Hempel's problem here, it turns out, is a recurring problem for empiricists. Despite the long history of attempts to analyse philosophically significant terms or concepts in accordance with empiricist strictures, the analyses never seem to work (Fodor 1981). Nor is it acceptable to simply eliminate all the recalcitrant concepts. People really do get toothaches. This is something we ought to be able to think and talk about.

The repeated failure of the reduce-or-eliminate strategy highlights how powerful the challenge posed by the argument from initial representational access can be. In the remainder of the chapter, we will examine a sample of relevant representational domains in more detail as illustrations of the argument and how it contributes to the case for a rationalist treatment of concept acquisition. While our primary interest is in the contemporary interdisciplinary rationalism-empiricism debate, we need to begin by taking a closer look at Hume's analysis of how causation is conceptualized, as it widely recognized that Hume's discussion of these issues forms the essential theoretical backdrop for understanding contemporary debates in cognitive science concerning how causation is conceptualized. Many of the most important views are framed directly as responses to Hume's analysis, and Hume's discussion still provides the clearest account available of the theoretical options open to empiricists.

Hume's account requires that there is a feeling of expectation, that this feeling traces back to a psychological disposition that is grounded in numerous

experiences of Ys following Xs, and that there is nothing more to the idea that X causes Y—no further representational content to the effect that X has a power that makes Y happen. As a result, Hume's account suffers from a number of difficulties. One is that his account doesn't do justice to the fact that people are prone to adopt causal theories in the absence of reliable correlations, where the causal belief itself leads to the perception of illusory correlations. For example, a belief in the power of an amulet may lead its owner to "see" the many times it protected him from harm. In such cases, it is often the belief that Xs cause Ys that make us think we have seen a strong correlation between Xs and Ys, not the actual experience of Ys reliably following Xs.

Hume's account also can't accommodate cases where people represent that X causes Y even though they haven't had any experiences of Xs or Ys in the past. It is not uncommon for an observer to assign a causal interpretation for what is for them a novel type of event. An inexperienced chemistry student might make the mistake of trying to clean a piece of potassium metal by dropping it in a bucket of water, which would result in an explosion. Clearly there is no expectation of an explosion, and the student may not have ever seen potassium metal in the past or have experienced a similar sequence of events (e.g., an explosion after immersing another metal in water). Yet in the course of witnessing the explosion, it would be perfectly natural—perhaps inevitable—to think not only that the potassium metal's contacting the water caused the explosion but that there is something about the metal (or better yet, the metal and the water), that is responsible for or necessitates this violent reaction.

This same example also illustrates a related and important limitation of Hume's empiricist learning model. It has difficulty accounting for the fact that people have no difficulty representing causes that *defy* expectation (Fair 1979). In all likelihood, the chemistry student's past experience tells her that water extinguishes fires and can be used to prevent combustion. So, if anything, she should suppose that placing potassium metal in water is the last thing in the world that would lead to an explosion. Nonetheless, it only takes this one instance to prompt the thought that this supposition is incorrect and that there is something about potassium metal and water that makes combining them so volatile.

Notice, too, that if people conceptualize causes and effects principally in terms of their statistical properties, and not in terms of the powers of the entities involved, then the information they ought to want when trying to figure out the cause of an outcome should be information about the conditions that covary with the outcome. But, in fact, what people seem to ask for instead is information about whether one or another mechanism could account for the outcome (Ahn et al. 1995). For example, when participants in one experiment were told that *John was not afraid of Dave's raccoon on this occasion*, to determine the cause of John's mindset, they asked things like whether the raccoon was in a cage (and hence couldn't attack Dave), as opposed to asking about what features of the

situation had or had not covaried with John's being afraid in the past. Information that is directed towards causal mechanisms is sought even for events that are described with nonsense sentences. When told *the feb mimbled the wug*, instead of asking about whether febs mimble a lot, or whether wugs often get mimbled, they asked things like whether the feb was mad at the wug (Ahn et al. 1995).

All of these considerations argue for a richer and more abstract understanding of CAUSE than Hume allows for in order to do justice to the psychological facts, one that represents causes as imbued with powers that necessitate their effects.[13] Leonard Talmy (1988, 2000) has developed an elegant *force dynamics theory* that elucidates this concept by showing how human language captures various types of interactions in terms of the represented forces, positions, and tendencies of the entities involved.[14]

At the core of this account is a mental model in which there are two force-exerting entities. The first, the *agonist*, has an intrinsic force tendency to move (or act) or not, while the second, the *antagonist*, may or may not oppose the agonist. What happens according to this mental model depends on the forces of the entities (whether they are for movement or rest), their comparative strength, and whether they are concordant or opposed. The various combinations of these dimensions capture a number of related outcomes, not just the direct causing of motion, but also such related everyday causal notions as *enabling* of motion, *preventing* motion, and *allowing* something to move or be moved. For example, the sentence *the ball's hitting it made the lamp topple over* depicts an instance of direct causation in which the agonist (the lamp) has an intrinsic tendency to not move that is overcome by the greater force of the antagonist (the ball). In contrast, *the plug's coming loose let the water flow from the tank* depicts an instance of allowing in which the agonist (the water) has an intrinsic tendency to move that is no longer impeded by the stronger opposing force of the antagonist (the plug).

One of the advantages of Talmy's force dynamics account is that it teases apart types of causally relevant events that people distinguish in everyday thinking that Humean (and other) models of the psychology of causation lump together. For example, people readily distinguish causing from enabling, and this distinction falls right out of the force dynamics model's set of options.[15] But for Hume, these

[13] We should emphasize that our concern here is with the psychological question of how people represent causation, not with metaphysical questions about the nature of causation.

[14] For overviews of research on the force dynamics approach to the representation of causation, see Pinker (2007) and Wolff (2017). See also Wolff (2007) for a computational model that builds on Talmy's insights.

[15] Wolff (2007) illustrates this point by contrasting *a cold wind caused him to close the window* and *a crank enabled him to close the window*. Although the concepts CAUSE and ENABLE are similar in content and are used in related ways, it would be extremely odd to say *a cold wind enabled him to close the window* or *a crank caused him to close the window*. As we note in the text, the inability to distinguish these related concepts and the different roles they play in human psychology is a problem for Hume's account. Wolff points out that it is a problem for other accounts of the psychology of causation that shun forces as well. Psychological accounts based on the representation of counterfactual conditionals incorrectly predict that the crank *is* taken to be a cause, since the window wouldn't have

ideas are conflated, as the statistical relationship between an enabling condition and its effect is on a par with the statistical relationship between a cause and its effect.

Many theorists have supposed that the problem of explaining how CAUSE is acquired presents a major stumbling block for any account, like Talmy's, that adopts an understanding of CAUSE in terms of the representation of forces. The difficulty such theorists see for these accounts is essentially the same one that Hume identified—that forces can't be perceived. (The thought is that if forces can't be perceived, then any theory that understands the concept CAUSE in terms of hidden forces will be unable to explain how the concept CAUSE is acquired.)

But whether forces can be perceived or not is relevant *only* given empiricist assumptions like Hume's. If a concept isn't required to have a semantic structure that can be analysed entirely in perceptual terms, FORCE might well be part of the acquisition base (as suggested by Leslie 1994; see also Baillargeon et al. 2009). To apply FORCE in a given situation, a learner doesn't have to literally sense force. It just needs to be the case that the learner's inferences are guided by a set of principles—perhaps an innate set of principles—that identify perceivable spatial-temporal cues that correlate with the possession, influence, and transmission of force. Principles involving such cues could occasion the activation of FORCE without FORCE itself being analysable in spatial-temporal terms. An innate schema of this type could include representations of the most basic categories of the force dynamics models (including FORCE, AGONIST, ANTAGONIST, RESULTANT, and CONCORDANCE) and a certain number of principles for interpreting the dynamics of an event. This type of rationalist theory provides a natural way of developing the richer notion of causation (in terms of forces) that we have seen is needed.

The suggestion here is that what is innate is a *skeletal* force dynamics model, not a completely filled in model for all of the many types of interactions that adults distinguish and understand, and certainly not one that spells out in advance all of the specific mechanical properties of each type of physical object. As we emphasized earlier, rationalist accounts typically involve learning mechanisms—learning mechanisms with a distinctively rationalist character—and thus typically provide *starting points* for development as opposed to full-fledged mature systems. In this case, the proposal is that there is an innate special-purpose mechanism that structures and facilitates learning about the behaviour of interacting entities by attending to relationships among what the mechanism takes to be their forces.

Is there an empiricist alternative for how FORCE is acquired, an account where forces are perceived after all, pace Hume, and where FORCE is constructed from more basic perceptual representations? A number of theorists have claimed that

closed if the crank hadn't been present. And the arrows that connect the variables in a Bayesian network don't distinguish CAUSE from ENABLE, or PREVENT from DESPITE (Sloman 2005; Wolff 2007).

there is, going back as far as Reid, who speculated that the idea of force is "derived from our voluntary exertions" and that "if we were not conscious of such exertion, we should have no conception at all of a cause, or of active power" (Reid 1788/2011, p. 278). Similar views have been held by Piaget (1930/2013) and more recently by White (2009, 2012) and Wolff and Shepard (2013). What these views have in common is the idea that children can learn about forces even if they can't literally see forces because they can still experience forces when they act on objects (e.g., pushing a ball) or are themselves acted upon by objects (e.g., being hit by a ball). The critical experiences include the sensation of undertaking intentional motor activity, the haptic sensations of pressure on the skin, kinaesthetic sensations associated with changes in the joints and the relative positions of one's body parts, and experiences pertaining to one's sense of balance.

For example, White suggests that children's haptic experience of force as they act on objects leads to the formation of stored representations that combine visual features of these object interactions with haptic features. Subsequently children notice a resemblance between these stored representations and object interactions that they see but aren't participating in themselves. This, in turn, leads to a visual impression that is "a kind of generalization of the proprioceptive impressions of force" (White 2009, p. 580). Take the example of a simple *launching event* in which there are two dots, A and B: B is initially stationary, A moves towards B until they make contact, and this is immediately followed by B moving in the same direction A had been heading in, and A staying where it was at the point of contact. As long as there is no spatial gap between A and B before B moves, and no delay between the time when the contact occurs and when B's motion occurs, this arrangement gives the distinct impression that A causes B's motion (Michotte 1946/1963). White's explanation of this impression is that it is a generalization of the proprioceptive representation obtained by pushing or kicking an inert object. The resemblance between the motions of A and B and, say, the infant's stored representation of pushing a toy results in similar visual-haptic representations being associated with seeing A move as it does.

The simplicity of this explanation is no doubt alluring. It suggests that empiricists really have no difficulty providing an account that avoids innate representational abilities that are specifically geared towards the representation of causation. But it is important to see that the explanation illicitly trades on a richer interpretation of its combinations of visual and haptic representations than its empiricist restrictions permit. It is certainly true that haptic, kinaesthetic, and vestibular sensations may be active when an agent judges a force to be present, even for forces that have nothing to do with the agent herself. The problematic move is the implicit assumption that the representation of force *reduces* to representations of these experiences (combined with stored sensorimotor representations) and hence that FORCE could be acquired by an empiricist learner.

The issue for all accounts of this sort is exactly the same one that Hume called attention to. Consider what happens when an infant kicks a toy and has the haptic and kinaesthetic sensations that occur as her foot makes contact with the toy and the toy comes to move. According to White, the infant sees the consequence of this action and forms a complex haptic-visual representation that is supposed to constitute conceptualizing the force she exerts on the toy. But why is this conceptualized as *imparting force* rather than the formation of what is merely a co-occurrence (or when it occurs repeatedly, a correlation) of haptic and visual sensations? Notice that it is certainly possible to have a sequence of sensations—a sensation in the foot at the moment of contact followed by seeing the toy fall over—and to represent this sequence of events merely as a sequence. But clearly this is completely different than representing a force dislodging the toy?[16]

Or consider what happens when the toy is pushed towards the infant and nudges her foot. What makes this more than a sequence in which the infant merely sees the toy's motion and feels a sensation of pressure in her foot and sensations in the joints of her leg? On what grounds can an empiricist like White say that the infant represents this in terms of the object *possessing a force* that impinges upon her? To be sure, Hume would say that the infant can't sense these forces, and doesn't in fact ever come to represent them. For Hume, all the infant feels is the sequence of sensations (and after numerous repetitions, she will be aware of an expectation linking these sensations). Given his empiricist assumptions, Hume is right about this restriction: this is all that an empiricist is licensed to assume at this point. What we are left with is a substantial gap between the representation of these feelings and the representation of something as abstract as a force. Hume rightly sees that this gap means that if empiricism is to be

[16] There are other difficulties with these proposals apart from why or how an empiricist learner would ever come up with the idea of force (much less do so reliably). Another concerns how a learner is supposed to generalize the sense of force that she initially acquires through her own interactions with objects to the interactions in which she isn't a participant. For example, when a learner first sees the simple launching event, why interpret this as resembling the stored representation of the learner herself pushing or kicking an inert object. Object A isn't anything like the learner. It doesn't *look* like her, or any relevant part of her (a hand or foot). It may be true that A moves towards B and then B subsequently moves, just as the learner's limb moves towards the toy and then the toy subsequently moves. But there is no reason that an unbiased empiricist learner should attend to this highly abstract commonality in the patterns of movement in the two cases, as opposed to the innumerable *dissimilarities* between the cases. Moreover, the exertion of force is not actually correlated with any simple pattern of movement. Large heavy objects *won't* move if kicked, fragile objects may *break* (rather than move) if kicked, and so on. The rationalist, of course, can readily explain why a learner would see similarities in these perceptually dissimilar events. The innate force dynamics model tells the learner in advance that the objects in physical interactions play different roles—agonist and antagonist—and the launching event exhibits spatial-temporal features that the model takes to be a paradigmatic instance of a mechanical event. Much the same type of account will explain why the learner would see causal actions that she herself is involved in in causal terms. In both cases, the innate force dynamics model guides the learner's interpretation. Interestingly, this sort of model further suggests that rather than being the part of the explanation of the origins of our concept CAUSE, the feeling of exertion that accompanies action comes to be interpreted as indicative of the deployment of force as a *consequence* of the fact that such feelings are associated with events that are already represented in causal terms.

maintained, the only real option may be eliminativism about any abstract concept involving notions like force.

To put this in the terms used earlier, White's account doesn't really face up to either the why problem or the how problem. Why would a learner who by hypothesis is entirely blind to the possibility of the existence of forces in the world come to see these correlations between different types of sensations as indicative of the existence of a force? And even if the learner could somehow see the need to represent these correlations in a new way, how would her non-causal representations allow her to formulate a distinctly causal interpretation of her experience, one that employs a representation like FORCE? There is nothing in White's account that warrants attributing to children truly causal forms of representation.[17]

Of course, the failure of this empiricist alternative doesn't mean that infants must be conceptualizing the world and their actions in purely Humean terms, representing only sensations and not forces. Infants can still be conceptualizing the world in terms of forces, just as White supposes. We just have to allow that the concept FORCE is something more than the associated sensations and that it is instead grounded in characteristically rationalist psychological structures associated with a force dynamics model that interprets these sorts of contingencies in terms of the forces involved. This model allows an infant to interpret herself as the antagonist when she kicks the ball, and as the agonist when she is hit by the ball. In short, adopting a rationalist perspective here explains why it is that infants are able to interpret certain types of sequences of sensations in terms of forces.[18]

Adults clearly have concepts pertaining to forces and the domain of causation. And there is little doubt that the psychological structures that are connected with the work we have been discussing support a rationalist account of the development of these concepts. But as with a number of the examples that we have discussed in other chapters, there is a further question about the representations involved in the interpretation of simple launching events and similar situations. When an infant sees a launching event and interprets the two objects as occupying different roles (agonist and antagonist) are these representations themselves conceptual as well, or are they nonconceptual?

As with the case of faces and biological motion (discussed in Chapters 8 and 10), this depends on how the conceptual/nonconceptual distinction is being drawn.

[17] A related empiricist proposal claims that a theory of causality can be quickly "bootstrapped" by a general-purpose inductive learning mechanism and input from several simple perceptual analysers, including an analyser that tracks "a feeling of self-efficacy" (Goodman et al. 2011, p. 112). However, the representation of causation that this model produces isn't a force dynamics representation. Instead, its focus is on capturing the dependence relations between events, which it does using causal Bayes nets (see Pearl 2000 and Sprites et al. 2001 for the mathematical formalism that this approach builds on; see also Gopnik et al. 2004 for related work on how causal Bayes nets might be learned in development). Accordingly, this proposal does not provide a model of how our ordinary concept of causation can be learned via domain-general learning processes either.

[18] For further discussion of empiricist accounts of the origin of the concept CAUSATION, see Chapter 21.

On some of the accounts of this distinction that were outlined in Chapter 6, the representation of a launching event will turn out to be conceptual (e.g., since it wouldn't satisfy Fodor's picture principle, the representation of a launching event would be discursive—not iconic—on Fodor's account and hence conceptual). On the other hand, on other ways of drawing the conceptual/nonconceptual distinction outlined in Chapter 6, the representation of a launching event won't turn out to be conceptual (e.g., since it wouldn't satisfy the Generality Constraint, accounts that require concepts to satisfy this constraint would hold that it is nonconceptual). The key point for our purposes, however, is that even if the representation of a launching event is considered nonconceptual, the domain-specific mechanism for analysing force dynamics (which we have argued is part of the acquisition base) almost certainly plays a major role in the acquisition of representations that are considered to be concepts by all accounts. This includes the concepts that are expressed by such words as "cause", "enable", and "prevent".[19] Accordingly, these uncontroversially conceptual representations provide further examples of concepts whose origins are best explained in terms of some form of rationalist account.

Let's turn now to another example. We mentioned earlier that logic is a good candidate for the argument from representational access. It is important to keep in mind that logical inference and thoughts with logical content aren't confined to the world of logicians and people with technical training in formal methods. Ordinary cognitive life traffics in thoughts involving negation (*not*), conjunction (*and*), disjunction (*or*), conditionals (*if…then…*), and biconditionals (*if and only if*). And while people aren't perfectly rational, they are at least minimally logical in that they can readily grasp at least some of the logical implications of some of their thoughts.

So where do logical concepts come from? As is widely known, some logical concepts can be defined, but these definitions make use of *other* logical concepts. It turns out that different sets of fundamental logical concepts would serve equally well for defining the others. For example, disjunction can be defined in terms of negation and conjunction (making it possible in principle to learn OR on the basis of NOT and AND), or alternatively conjunction can be defined in terms of negation and disjunction (making it possible in principle to learn AND on the basis of NOT and OR).[20] But this approach can only go so far. It shows us how some logical

[19] Or to put this in other words, in the local rationalism-empiricism debate concerning PREVENT and closely related concepts, if the concepts themselves aren't innate, they will still trace back to a local acquisition base that includes an innate mechanism for representing force dynamics.

[20] Using the symbol "~" to represent negation (meaning "it is not the case that"), we can see that (p or q) is equivalent to ~ (~p and ~q). This is because if it is not the case that (it is not the case that p *and* it is not the case that q), then it must either be the case that p or the case that q (or both). Likewise, (p and q) is equivalent to ~(~p or ~q). This is because if it is not the case that (either p is not the case *or* q is not the case), then it must be the case that both p and q. In this way, just negation and either disjunction or conjunction are sufficient for deriving the full standard set of basic logical

concepts can be defined in terms of others, but it still leaves us with the question of how the more basic logical concepts—the ones that enter into these definitions—are acquired.

Although our focus is on psychology—how logical concepts are acquired—there is an illuminating parallel in Quine's (1936/1976) discussion of truth by convention. In the twentieth century, one major account of logical truth tried to explain the existence of necessary logical truths in a way that doesn't lead to any deep metaphysical commitments. The thought was that logical truth might derive from linguistic conventions governing logical expressions, conventions that guarantee the truth of certain statements that employ logical terminology. According to this approach, there needn't be any language-independent logical facts, or any language-independent necessities. For example, the fact that *if* (p *and* q) *then* p is, and must be, true doesn't tell us about some mysterious feature of the world; it is simply a consequence of the conventions that language users have adopted for the connectives "and" and "if…then…". Attractive as this account is, it cannot be right. As Quine (1936/1976) pointed out, there are an infinite number of truths that instantiate this type of schema. As a result, the truths instantiating the schema can't be listed individually. So establishing their truth by some form of conventional stipulation would, at the very least, require conventions that help themselves to enough logic to express the needed generality.[21] For example, we might adopt a convention that *uses* terms like "all" and "if…then…" that stipulates that it is true of *all* sentences that *if* they are instances of this schema, *then* they are true. But if some logical concepts (e.g., ALL, IF…THEN) are required to make the relevant conventional stipulations, then logical concepts cannot all be the product of conventional stipulations.

Put in psychological terms, Quine's insight is that for a system of representation as powerful as the human conceptual system, the contents of logical concepts cannot be derived from purely non-logical resources. A system of representation that already has access to some logical concepts can be used to formulate more logical concepts, but logical concepts cannot be created out of entirely non-logical resources. A mind that doesn't possess at least some innate logical abilities wouldn't be able to break into the circle of logical concepts even though logical concepts can generally be defined (i.e., defined in terms of other logical concepts). These considerations do not tell us which specific logical concepts are likely to be the innate ones. But this level of specificity is not required by the

connectives (negation, conjunction, disjunction, the conditional (if, then), and the biconditional (if and only if)).

[21] A similar observation is made by Lewis Carroll in his dialogue featuring Achilles and the Tortoise (Carroll 1895). Achilles heroically tries to provide an explicit justification for an instance of the inference rule modus ponens (if *p* then *q*, *p*, so *q*), but is thwarted by the Tortoise, who points out that the justification itself depends on accepting another instance of modus ponens (which incorporates the first instance in its antecedent), leading Achilles to produce a further justification for *that* inference, and so on, ad infinitum.

argument—one can have good grounds for holding that some concepts in a given domain are innate (or acquired via characteristically rationalist learning mechanisms) without knowing which specific concepts this is true of.[22]

We have previously highlighted some interactions among the different arguments for concept nativism, which show how these arguments can work together. Here we want to highlight the interaction between the argument from initial representational access and the argument from early conceptual development, illustrating this interaction with the example of logical concepts.

There is evidence that children as young as 12 months of age already reason in accordance with logical principles (Cesana-Arlotti et al. 2018; Cesana-Arlotti, Kovács, and Téglás 2020). For example, Cesana-Arlotti et al. (2018) showed infants brief videos in which there is an ambiguity concerning the identity of an object which could be resolved by disjunctive syllogism—the logical inference *A or B, not A, therefore B*. In one video, two objects with distinct features yet identical tops are initially fully visible before becoming hidden behind a curtain. A cup then scoops up just one of the objects in such a way that only its top portion is visible, so that it is unclear which of the two objects is the one in the cup. The curtain then lowers to reveal one of the objects (say, A). At this point, if the infants can do logical reasoning, they are in a position to infer the identity of the object in the cup using the disjunctive syllogism. They should thus form the expectation that B is in the cup and be surprised if B isn't there. This is exactly what happens. Infants look significantly longer when the object in the cup is revealed to be A rather than B. What's more, like adults, infants also display a tell-tale cue indicating that they are employing a logical inference. Just at the point in the video when the disjunctive syllogism can be used to determine the identity of the object in the

[22] Prinz (2002) has argued that logical cognition can be fully accounted for without positing *any* innate logical concepts. However, his view involves another instance of the reduce-or-eliminate strategy that we encountered earlier, where he embraces the eliminativist option. Essentially the argument rests on the claim that we don't need any *innate* logical concepts because we do not need *any* logical concepts or representations of any kind (innate or otherwise) in order to explain logical cognition. According to Prinz, we only need what might be called *non-logical representations*—representations that do not employ any logical concepts at all—and logical processes that operate over these non-logical representations. Certainly, we agree that if there are no concepts in a given domain, then we don't need to posit any innate concepts in that domain. But this seems to us rather extreme lengths to go to in order to deny the existence of innate logical representations. Even if it were true that people do not employ logical concepts in their everyday thinking—and instead rely solely on logical operations over non-logical representations—it seems clear that it is possible for humans to acquire logical concepts such as OR and NOT. Presumably such acquisition would, on Prinz's view, depend on the prior possession of the abilities to employ logical processes on non-logical representations. But then the pressing question is: Where do these abilities to employ logical processes come from? Are they innate or learned? And once we ask this question, it should be clear that a variant of the psychologized version of Quine's argument would apply in this case as well—leading to the conclusion that while some of these abilities may be learned by relying on others, they could not all be learned in this way. Some logical abilities would have to be innate. Thus, even if Prinz's eliminativism regarding logical representations were true of most everyday cognition, a rationalist account of the origins of logical concepts would still be warranted. (However, for evidence that it is *not* true of most everyday cognition, see, e.g., Papeo, Hochmann, and Battelli et al. 2016.)

cup, their gaze shifts to this object and their pupils dilate, in the same way as happens when adults make a logical inference in such circumstances. Taken together, these results suggest that 12-month-olds can perform elementary logical inferences and have the representational resources to support such inferences.[23]

The interaction that we want to call attention to is how the problem of initial representational access—particularly the aspect of the argument that we have called the *why problem*—strengthens the argument from early development. Applied to the domain of logic, the why problem is the problem of explaining why a learner who has no logical representations—who has only non-logical representations—would ever start to think in terms of logical representations. Such a thinker might entertain complex perceptual representations and simple concepts. But they would not be able to entertain any complex representations expressing such things as conjunction (*A* and *B*), disjunction (*A* or *B*), or negation (*not-A*). Why would such a thinker ever even come to focus on, for example, an object not having a given property? And if previously they were only capable of representing that something *does* have a given property, not of representing that something *doesn't* have the property, what representational resources could they use to guide them to being able to represent negation as such? Why wouldn't a thinker in this position just stick to thinking in completely non-logical terms?

How does this problem make the argument from early development stronger? The short answer is that the harder the learning problem, the longer the time period that we should expect an argument from early development to be applicable.[24] And acquisition problems where the argument from initial representational access apply are paradigmatically hard problems for an empiricist learner. This is because the learning problem in such cases isn't just one of having sufficient time to get enough input to detect a pattern in the data. It is one of coming to see the data in completely foreign terms that one doesn't have the resources to represent as such.

It may be thought that the why problem for acquiring logical representations could be addressed if we think of the learning process as being aided by culture, natural language, and guidance from those who already have such representations. One problem with this suggestion is that this would mean conceding that the problem of early development *has* been strengthen by the problem of initial

[23] The simplest account of these results is that infants have explicit representations of the logical concepts OR and NOT. But as Cesana-Arlotti et al. note, another possibility is that infants set up a model with two options, and when one of these options is ruled out, they adopt the other. This would be a way of implementing a disjunctive inference without employing an explicit symbol for disjunction. Both of these types of accounts support a rationalist account of the origins of logical concepts, however. On one account, explicit logical concepts are likely to be innate. On the other, such concepts are presumably grounded in innate abilities that mimic explicit logical inferences.

[24] It should be noted that the hardness of a learning problem isn't just a function of the problem itself, but also what the learner can bring to bear on the problem. In particular, what's hard for an empiricist learner needn't be at all hard for a rationalist learner because they have very different cognitive resources to draw upon.

representational access—at least to the point where children possessed the relevant portions of natural language so that this type of social scaffolding could guide their learning. In this case, evidence that infants possessed such representations already by the age of 12 months would be powerful evidence for a rationalist account of the origins of these representations. This suggestion also faces a second problem. This is the problem of how children could make use of these resources even if they were available. Suppose that a child acquired elements of natural language but still thought in completely non-logical terms. What type of guidance could be given through language? The problem is that the children have to be able to *understand* the guidance, and if they are not able to entertain thoughts involving logical concepts, they will also not be able to understand *sentences* involving corresponding contents.[25]

So far, we have looked at two cases: causal concepts and basic logical concepts. Human beings also naturally form thoughts with logically *modal content*, reasoning about what may or may not be possible, impossible, necessary, or contingent. This is another place where the argument from initial representational access very likely applies. How could a mind that had no modal representational abilities acquire these from scratch? Imagine going about your business and taking in the various things around you. You might see that unsupported objects drop or that the addition of two objects to a collection of two yields four, but all this tells you is about what *has* happened in your experience, not what *could* happen or what *must* happen. Likewise, you might represent the actions that you perform, but all this tells you about is what you *did* and what happened afterwards, not what *might* have happened had you acted differently. Nonetheless, people are quite capable of thinking about what must happen and about what might have happened. Counterfactual reasoning based on remembered action is particularly common in the context of negative assessments about an outcome. These prompt the thought *if only I had done this differently* and reflection on ways to prevent similar bad outcomes in the future.[26]

The question is whether empiricist models of concept acquisition can explain the acquisition of these abstract representational abilities given the representational restrictions they impose on the acquisition base—initially an empiricist learner has no modal representational abilities whatsoever to draw upon. But if a learner were only capable of thinking about what actually happens, why would she begin to think about what did not happen but could have happened? And more importantly how would she gain the conceptual resources required for such a thought to

[25] The problem of using language to bridge such representational gaps is often underestimated; see, for example, the discussion in Chapter 5 of McDowell's remarks on the role of language in conceptual development.

[26] See Téglás et al. (2007), Téglás et al. (2011), and Téglás et al. (2015) for work suggesting that infants represent possibilities. Here too the argument from early development complements the argument from initial representational access, just as we saw with logical concepts.

be possible? Suppose that such a learner entertains a thought that is false. If she possesses concepts of truth and falsity, she could presumably come to represent the state of affairs as false. But why—and how—would she go on to represent it as *possible* as well? Likewise, it is easy to see how a learner who possessed concepts corresponding to integer value quantities, but no modal notions like possible and necessary, might learn that adding one item to another has regularly resulted in a collection of two. But why, and how, would she come to think that 1 + 1 *must* be 2—that it is impossible for 1 + 1 to not equal 2?[27]

Children from an early age are keen to engage in pretend play in which they are highly motivated to suspend beliefs about how things actually are and to act out scenarios based on suppositions about other possibilities (e.g., treating a banana like a telephone while clearly understanding that it is still a banana; Leslie 1987). But how could they do this without a cognitive architecture that has design features that motivate imaginative thinking and that support the elaboration of possibilities that the thinker full well recognizes aren't real? More generally, how do children come to be able to entertain thoughts about what's possible, what's necessary, and what might have been?

Nichols and Stich (2003) offer a plausible sketch of what a core aspect of this architecture might look like. As they see it, an essential part of being able to represent different possibilities is having a model of the world that can be updated and that is effectively offline so that it doesn't interfere with what one actually believes the world to be like or directly guide action. This leads them to posit the existence of what they call a *possible worlds box*, a psychological mechanism that houses representations which are in the same code as beliefs and which interact with whatever mechanisms elaborate the consequences of holding a belief but that have a different functional role than beliefs.[28] The contents of the possible worlds box are a (partial) model of the world—some of which may correspond to how the world is, but some of which does not. Nichols and Stich don't say whether they take there to be a rationalist account of the origins of their possible worlds box. But such an account could provide the beginnings of what would be needed here. It is just the beginnings of what would be needed because, as described, the possible worlds box could be used for different purposes, not just for modal reasoning. This is in fact what Nichols and Stich intend—for example, they suggest that it is also involved in reasoning about another agent's view of the world when you understand it to be different from your own. This

[27] Of course, there is a further question of how an empiricist learner could come to possess representations of integer values (see below).

[28] This figurative talk of a box in the head is ultimately be understood in terms of the functional roles of the representations in question. An alternative way of thinking of it is that these representations are tagged with a marker to indicate their distinctive role in thought (different from ordinary beliefs). And representations that are inferred from such tagged representations would be similarly tagged.

means that, on its own, it doesn't differentiate this type of content from content that is explicitly about what is merely possible. So this system would need to be supplemented, likely with further rationalist resources, in order to support distinctively modal representations and reasoning. It nonetheless offers a start. What this suggests is that the argument from initial representational access gives grounds for holding that the possible worlds box—or something much like it—is either innate or acquired via a rationalist learning mechanism and that it involves articulation (in the sense of Chapter 2) in relation to a larger arrangement of psychological structures that supports modal reasoning.[29]

We will round out our discussion by considering two further candidates for domains where the argument from initial representational access seems promising—number and time. Let's start with the representation of numerical content. In a discussion that resembles Leslie's remarks about the challenge of learning about mental states (quoted above), Stanislas Dehaene (1997) puts what is essentially the argument from initial representational access in the form of an analogy. "[I]t seems impossible for an organism that ignores all about numbers to learn to recognize them. It is as if one asked a black-and-white TV to learn about colors!" (pp. 61–62).

Suppose you already have a representational foothold in the domain of numerical quantity—maybe not a grasp of complex arithmetic operations or even an appreciation of exact integer values, but at least the ability to represent approximate numerical quantity, as described in Chapters 8 and 10. Then learning further numerical concepts (e.g., integer concepts) may still be a very difficult problem, but this representational foothold would at least allow you to get input to your learning process that has numerical content. On the other hand, if you had no way at all of representing numerical quantity as such, then numerical phenomena could be staring you in the face and you would look right past them. In comparing a collection of four apples with a collection of eight, you might notice that the latter has more food, that it is heavier, that it blocks more light, and so on. But none of this gets you any closer to actually representing that it is *numerically* larger than the other.

[29] This isn't to deny that modal knowledge, imagination, and counterfactual reasoning take time to develop. To think about a possibility that conflicts with your current understanding of the world requires executive control to inhibit real-world knowledge from interfering with what is being imagined. It also requires working memory and attention to switch between the understanding of how things actually are to what is being imagined. As inhibition, attention, and working memory improve with age, so should the ability to entertain and reason about various possibilities. And people's understanding of modal matters may also undergo genuine development, adding additional principles governing inferences, or becoming more sophisticated in other ways over time. But that is not at all the same thing as saying that the very idea that things could be other than they are could reliably develop solely on the basis of representations with no modal content and without an innate basis for modal reasoning.

Some theorists have claimed that the representation of approximate numerical quantity isn't innate and that children can learn to represent approximate numerical quantity. But how exactly? The problem we see with these accounts is that although they are supposed to explain the acquisition of approximate numerical quantity without presupposing any prior numerical content, they offer no insight about why or how a learner is supposed to make this leap given the initial types of representations that the model is restricted to. In a recent critique of so-called *number sense* theories (theories that follow Dehaene (1997) in supposing that there is an innate system for approximate numerical quantity), Leibovich et al. (2017) suggest that learning about approximate numerical quantity is a drawn-out process in which children tease apart numerical quantity from non-numerical continuous magnitudes after noting how they correlate.[30] But what they don't explain is how a learner could detect these correlations between number and continuous magnitudes without being able to independently represent the two things that are supposed to be represented as being correlated: number, on the one hand, and continuous magnitudes, on the other. Rather than explaining where the representation of approximate numerical quantity comes from, their model tacitly *presupposes* that children have a certain amount of numerical representation to begin with (Margolis 2017).

Our final example involves the acquisition of temporal concepts. This turns out to be an interesting example because some of the ways that people think and talk about time strongly resemble the ways they think and talk about space. Notice that just as a friend can be *at* the corner, she can arrive *at* 5:00pm, and that just as she can plant tomatoes *between* the oak trees and the lake, she can walk the dog *between* lunch and dinner. We also often speak of time in such a way that the future is *in front* of us and the past *behind*. This fits with two common metaphors (Lakoff and Johnson 1999). In one, time is moving while the observer stays put (*the holidays are coming up*), while in the other the observer is moving across a fixed landscape in which locations count as different moments (*we passed the deadline*). These striking facts have been thought to illustrate that what may seem to be fully abstract concepts actually turn out to be representations that are grounded in more concrete forms of representation and ultimately in sensorimotor simulations (Lakoff and Johnson 1999). If this is true, then perhaps

[30] See also Mix et al. (2016), who argue that perceptual learning theory can explain how children first become able to differentiate approximate numerical quantity from non-numerical magnitude. We certainly agree that perceptual learning can explain how learners discover features of their environment and form some types of new perceptual concepts, but *why* would children who are incapable of seeing the world in numerical terms focus their learning on numerical quantity as such? And *how* could they come to represent numerical quantities to themselves on the basis of only their prior non-numerical representational abilities?

temporal concepts are constructed out of spatial representations and there is no need for any innate special-purpose mechanisms for representing temporal content.

We'd suggest, on the contrary, that as Kant recognized more than two centuries ago, temporal representation is too basic to be derived from non-temporal forms of representation. At best, the spatial metaphors that suffuse temporal cognition show that certain aspects of temporal cognition depend upon spatial representation, not that temporal concepts are exhausted by their spatial content.

To see why, notice that spatial metaphors are commonplace. They pop up when we think and talk about not only time (*the meeting was moved from Thursday to Friday*), but also possession (*the money went to Smith*) and the ascription of attributes (*Jones went from elated to depressed*), with corresponding inferences across these different domains (Gruber 1965; Jackendoff 1983).[31] Yet clearly these different spatial metaphors have different content. To put it bluntly, one is about time, one is about possession, and one is about a person's attributes—and none are literally about space. What seems to be happening here is that the same abstract rules of inference are being indexed to different domains (time, possession, etc.) and that the domain in which these inferences are understood to take place makes an essential contribution to what the speaker is thinking and talking about. So it's only because speakers have more basic representations of time, possession, and so on that these distinct spatial metaphors can even be devised and used to convey different types of non-spatial content. This point is entirely general and not at all peculiar to space-time metaphors. A common way of modelling the cognitive processes involved in metaphor is as a mapping from a base domain to a target domain, in which inferential patterns from the base are projected into the target (Gentner 2003). However, for this to work, there has to be at least some representation of the target domain over which the mapping is defined, otherwise the metaphor creation process lacks the materials it needs to even get started. It is only because speakers have more basic representations of time, possession, and so on that these distinct spatial metaphors can be devised and used to convey non-spatial content.

Returning to the spatial metaphors for time,[32] it is also worth pointing out that while time is represented as having affinities with space, there are important differences between the two and that these show that there must be more to temporal representation than the representation of space. Galton (2011) notes that

[31] For example, just as the fact that *the meeting is on Friday* can be inferred from *the meeting was moved from Thursday to Friday*, so the fact that *the money is Smith's* can be inferred from *the money went to Smith*.

[32] Note that spatial metaphors aren't the only ones used in talking and thinking about time. Financial metaphors are also common—one can, for example, *spend* time, *save* time, *invest* time, or *waste* time. This no more means that the content of temporal concepts must be grounded in financial representations than the spatial metaphors mean that the content of temporal concepts must be grounded in spatial representations.

the core and unique feature of time is *transience*, an attribute that is hard to describe without lapsing into circularity but that can be conveyed through examples, such as the sayings *here today, gone tomorrow* and *you only live once*. The crucial thing to keep in mind is that space as such lacks transience. When transience enters into a spatial metaphor, this is indirectly through space's relation to motion, as in *time flies* and *we put those troubles behind us*. Moreover, what makes motion suitable for representing time is the fact that motion involves change—*change in general* rather than *change in spatial position*. Notice that there are other ways to represent time via change without mention of space. This can be done with metaphors of production/creation and consumption/destruction, as in *to make time for a friend* and *it ate up all of his time*. What all this suggests is that transience can't be explicated spatially. Instead, it has a sui generis character. According to the argument from initial representational access, this is exactly what makes it such a good candidate for an innate form of representation. You need an initial representation of time to get further, richer ways of representing time. The conceptualization of time that is familiar from spatial metaphors requires at least some temporal representation to begin with.

To recap, in this chapter we have discussed the conceptual domains associated with the representation of causation, mental states, logic, modality, number, and time. In doing so, we have argued that concepts in these domains trace back to characteristically rationalist psychological structures in the acquisition base—they are either innate or acquired via rationalist learning mechanisms. The crux of the argument in each case is that acquiring concepts in that domains would be all but impossible without a prior representational foothold in that domain. We could present further examples of the argument from initial representational access. But the examples we have discussed illustrate how the argument simultaneously poses a significant challenge for empiricist theories of conceptual development and offers much support for a rationalist approach. Empiricist attempts to evade the argument have frequently taken the form of the reduce-or-eliminate strategy, where reductive analyses have been based on the ABC model of conceptual development. But as we have argued, these ways of trying to evade the argument fail to come to grips with the real challenge this argument poses. Fundamentally the argument is about the need to have a representational foothold in certain domains to acquire new concepts connected to these domains. Such a foothold remains elusive in the absence of a rationalist account of the acquisition base.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0012

13
The Argument from Neural Wiring

Our fifth argument for concept nativism is based on neurological evidence. In principle many types of neurological evidence might be brought to bear on the rationalism-empiricism debate regarding the origins of concepts. The *argument from neural wiring*, however, focuses on reasoning based on one particular type of neurological phenomenon, namely, what is called *neural plasticity*. Neural plasticity refers to the ways in which the brain changes in response to experience and action and to trauma to the body. Our focus on neural plasticity as an argument for concept nativism may seem surprising, since plasticity is typically taken to be an argument for *empiricism*, not rationalism. Many theorists have supposed that the brain's plasticity undermines a rationalist view of cognitive and conceptual development on the grounds that plasticity suggests that there are few constraints on the structure and function of cortical areas. But what if the brain's plasticity isn't so open-ended and instead takes the form of *constrained plasticity*, in which the brain is highly plastic but its plasticity is nonetheless limited in ways that suggest that characteristically rationalist psychological structures play a substantial role in conceptual development? In that case, the neurological evidence would point in the opposite direction, providing support *for* concept nativism, not against it. The argument from neural wiring says that this is exactly the situation we are in. Neural plasticity is highly constrained in ways that are best explained within a rationalist framework.[1]

As a first step to seeing how plasticity might support a rationalist account of conceptual development, it is important to recognize that all theorists, rationalists and empiricists alike, agree that the brain is plastic in certain ways. If it weren't—if the brain couldn't undergo changes in response to experience—learning wouldn't be possible. But just because the brain is plastic in this sense doesn't mean that its structure is determined by sensory input in the domain-general way that is often assumed (Pinker 2002; Marcus 2004).

In fact, there is striking evidence that much of the structure of the brain, including the segregation of cortical layers, is not driven by sensory input. In a landmark investigation, Verhage et al. (2000) examined neurological development in a group of mutant (or "knockout") mice whose brains were unable to release any neurotransmitters, which are required for synaptic transmission, and

[1] See Laurence and Margolis (2015) for related discussion, which addresses additional points concerning the relationship between neural plasticity and concept nativism.

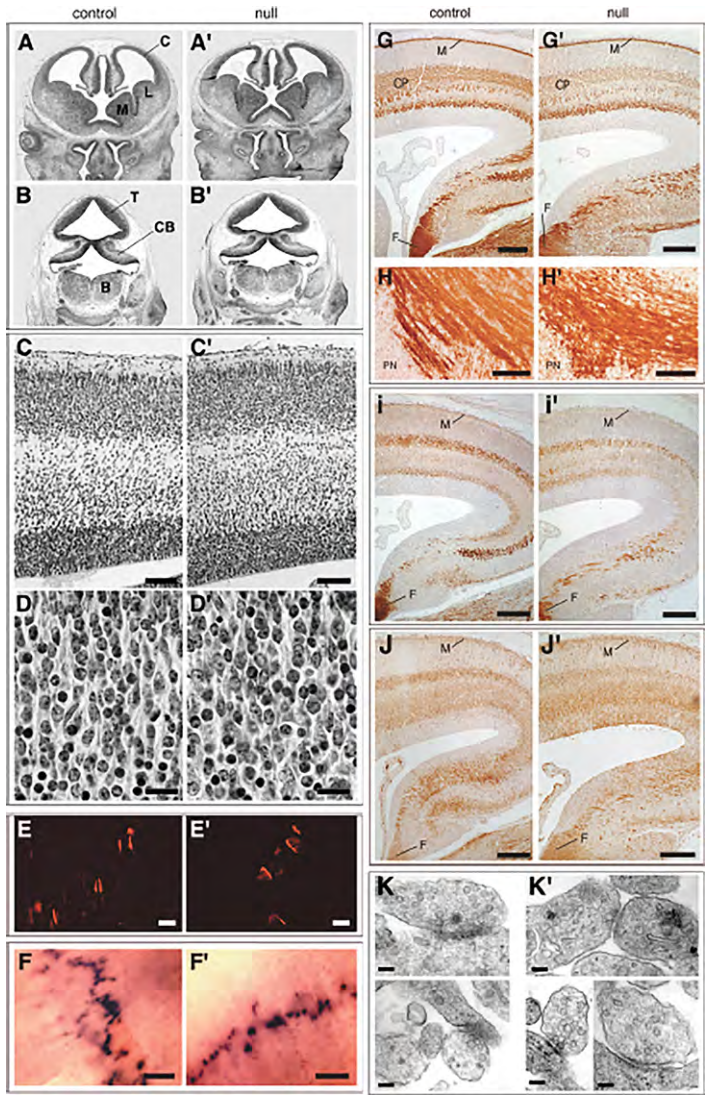

Figure 13.1 Neurological development in mutant mice that were genetically engineered to eliminate synaptic transmission (null) and in normal mice (control). (From Verhage et al. 2000. Reproduced with permission.)

thus were deprived of all synaptic transmission.[2] Under these circumstances, *no experience-driven neural development could have occurred*. Yet these mutant mice and control littermates were found to have brains that were remarkably similar

[2] Specifically, the expression of the *munc18-1* gene was suppressed in the mutants, resulting in the absence of neurotransmitters.

up until birth (see Figure 13.1). (After birth, of course, the mutant mice died. Without functioning synaptic communication, the brain cannot support even the most basic life functions, such as breathing.) This degree of similarity shows that many features of even the fine-grained structure of the brain can develop without any sensory input or feedback. Experience-driven neural activity plays more of a role in fine-tuning and maintaining this structure, rather than in establishing the overall organization itself (Marcus 2004).

Putting aside questions about the overall structure of the brain, what about other forms of neural plasticity? Many who are opposed to concept nativism point to evidence for dramatic instances of functional plasticity associated with sensory deprivation holding that these functional changes in the brain refute concept nativism.[3] These functional changes aren't supposed to be instances of the everyday type of plasticity that all theorists agree to. Nor are they supposed to be instances of constrained plasticity either, since they typically involve cortical tissue taking on new or unusual functions (e.g., sections of visual cortex being used to analyse auditory input). But does this functional plasticity support an empiricist account of conceptual development? Our answer is no. These functional changes aren't as novel as they may first appear. On the contrary, they are often highly constrained in ways that are indicative of the rationalist approach to concept acquisition.

Let's begin by considering congenital sensory deficits. In such cases, cortical areas that usually process sensory information from an impaired modality end up taking input from another sensory modality—an organization (or reorganization) known as *crossmodal plasticity*. An empiricist who supposes that the cortex has a high degree of equipotentiality—where any portion of the cortex is equally well suited to perform any cognitive function—would predict that the resulting functions of these cortical areas would differ from the normal case, reflecting the difference in the input. But work with congenitally blind individuals shows that, contrary to this prediction, many downstream components of the visual cortex and related brain areas have the same functional specificity in the congenitally blind as in sighted individuals.

Take the representation of faces in the visual cortex. In normal development, the fusiform face area (FFA) (in the fusiform gyrus) responds selectively to faces. There has been much speculation about why this particular patch of the brain is such a reliable locus of face-specific processing. One proposal has been that the FFA isn't destined to represent faces but that it takes on this role because it

[3] Empiricists also point to evidence from so-called rewiring experiments where an animal is surgically altered so that sensory input from one sensory modality feeds into a different part of the brain—for example, retinal projections being rerouted so that the neural signals that would normally go to the primary visual cortex are fed to the primary auditory cortex via the auditory thalamus (Sharma et al. 2000). See Pinker (2002) and Laurence and Margolis (2015) for discussion of why this work doesn't argue against rationalism.

receives much of its input from an area in the visual cortex that is responsive to the foveal region of the retina. Given that faces are typically targets of attention and so appear in this region of the retina, the FFA's developmental history might include a disproportionate number of facial stimuli (see, e.g., Hasson et al. 2002). The problem with this and related proposals, however, is that the visual cortex of congenitally blind individuals has also been found to have the same face-specific response (Murty et al. 2020).[4] In this work, sighted and congenitally blind individuals were both asked to haptically explore three-dimensional printed faces and other types of objects. Even though the blind participants had never *seen* a face—and so had no visual input of faces from the foveal region of the retina—face selectivity was found in their visual cortex in much the same location as in the sighted participants.[5]

A similar pattern of the same functional specificity in sighted and blind individuals occurs with the representation of the spatial location of objects in the visual cortex. In one study, researchers presented early blind participants with auditory stimuli that varied in terms of sound type (different piano chords) and spatial location (Renier et al. 2010).[6] Their brain activity was measured during two behavioural conditions, an identification condition (in which they had to determine whether sequentially presented stimuli were of the same sound type) and a location condition (in which they had to determine whether they had the same location). What was found was that a part of the visual cortex associated with the representation of *visual* spatial location in sighted participants (the anterior part of the right middle occipital gyrus, or MOG) was differentially active for the *auditory* spatial localization task relative to the auditory

[4] In addition, there is also a question of whether ordinary visual experience could be the driving force behind the functional specificity of the FFA given how early in development the ventral visual stream exhibits much of the differentiation found in adults. fMRI data indicate that at as early as 2 months of age, the FFA selectively responds to faces, the PPA (the parahippocampal place area) selectively responds to scenes, and the EBA (extrastriate body area) selectively responds to bodies (Kosakowski et al. 2022).

[5] Other researchers have explored the role of experience in the formation of face-selective cortical areas using a deprivation experiment similar to the Sugita (2008) study discussed in Chapter 10. Macaque monkeys were raised without exposure to any faces and later tested for whether they developed the normal cortical patches that selectively respond to faces, bodies, and scenes (Arcaro et al. 2017). The finding was that the response to bodies and scenes was the same as in control participants but that the face-deprived monkeys appeared to lack the usual face-selective response in the superior temporal sulcus. This led these researchers to conclude that "seeing faces is necessary for face domain formation" (the title of their paper). In contrast, face-deprived chicks have been found to have neurons that selectively respond to schematic face-like stimuli as opposed to other configurations of the same components, including upside down schematic faces (Kobylkov et al. 2024). There are interesting questions about this apparent discrepancy and about how best to interpret the results with macaques. But regardless, the main point worth stressing is that, as noted in the text, studies directly focused on face-selective cortical activity in the human brain strongly support the view that visual experience is *not* necessary for face selectivity to develop in the FFA, as this appears even in congenitally blind humans.

[6] Early blind participants include those who are blind from birth and those who develop blindness early in childhood. In this experiment, the early blind participants were blind either from birth or by the second year of life.

identification task. An analogous tactile task was also run with the same participants. In this case, their fingertips were given different types of stimulation (for the identification task) or there was stimulation to different fingers (for the spatial location task). Once again, the right anterior MOG was differentially active for the spatial localization task relative to the identification task.[7] The upshot of this study is that, while the MOG is clearly plastic—in the early blind it comes to subserve *auditory* and *tactile* spatial localization abilities that it does not subserve in sighted individuals—the plasticity it exhibits is a form of *constrained* plasticity. The MOG continues to carry out the function of spatial localization in the early blind, just with different types of sensory input.

Further neuroimaging studies have revealed the same pattern of constrained plasticity in other neural regions with associated spatial functions. For example, other research has examined the activity in the posterior parietal cortex (PPC), which is involved in the representation of space for purposes of guiding action (Lingnau et al. 2014). The PPC normally takes its sensory input primarily from vision and exhibits a pronounced gradient—with posterior subregions recruited more heavily for (visual) guidance of reaching and grasping, and anterior subregions more for planning and execution of motor action. Using a reaching task that wasn't visually guided—one that was proprioceptively guided instead—the researchers were able to show that the same pattern of functional differentiation occurs in congenitally blind participants. They compared the brain activity of congenitally blind individuals and blindfolded sighted individuals who were required to touch and grasp objects in different specified locations. Notice if the brain's plasticity takes the form of constrained plasticity, then sighted and blind study participants should have similar activation in the anterior portions of the PPC since both would be equally engaged in the planning and execution of motor action. But constrained plasticity predicts that there should also be significantly greater activation in the posterior portions of the PPC in the blind participants since their PPC would have adapted to make use of *non-visual* sources of information about spatial location in order to guide their reaching and grasping. This is exactly what the researchers found, leading them to conclude that "neural plasticity acts within a relatively rigid framework of predetermined functional specialization" (Lingnau et al. 2014, p. 547).

One of the major features of the visual cortex is the functional division corresponding to two broad networks of interrelated neural regions. The ventral visual stream (often referred to as the *what pathway*) represents object properties, and is involved in object recognition; the dorsal visual stream (often referred to as the *where pathway*) represents object location and the spatial relations between

[7] Blindfolded sighted control participants doing these auditory and tactile tasks did not show the same activation of the MOG. However, MOG activation in sighted participants did occur in a comparable visual task.

objects, and is involved in object-directed action. The results we have reviewed indicate that the dorsal visual stream continues to exist in early blind and congenitally blind participants and that its component subregions engage in the same functional processing for object location despite profound changes in sensory input (auditory or proprioceptive vs. visual).[8]

Further investigations of the dorsal visual stream fill out this picture by showing that it is not just the representation of spatial location that is preserved. For example, researchers have also examined activity in the dorsal occipito-temporal cortex in congenitally blind adults, focusing on a region of interest encompassing the hMT+ complex, which normally represents the direction of visual motion (Wolbers et al. 2011). To determine whether this region retains the same function when deprived of its usual (visual) input, brain activity was measured while congenitally blind participants heard leftward and rightward broadband noise signals, as well as static control stimuli. The region of interest examined was specifically involved in motion detection in congenitally blind participants even though the sensory input in this case was auditory, not visual. Once again, we have an impressive instance of plasticity (the fact that the dorsal visual pathway is co-opted for auditory processing), but the plasticity is constrained, preserving the normal functional specificity of a dorsal pathway subregion.[9]

These and related studies showing preserved functional specificity in the visual cortex support what is known as the *metamodal hypothesis* regarding the brain's functional organization (Pascual-Leone and Hamilton 2001). According to this hypothesis, much of the brain is composed of distinct computational systems whose functions are established independently of their sensory input. These systems are capable of processing sensory information from differing modalities but settle on a given modality when the input from that modality provides the best fit for the computations carried out—thus giving the appearance of modality specificity. On this view, it is a misnomer to speak of the "visual cortex", the "auditory cortex", etc. Rather, each of these broad areas are composed of neural systems that engage in computations that create a preference for a given modality, but the computations performed aren't inherently about visual or auditory content, and so, when the preferred input is unavailable, the brain switches to the next best fit.

There is now a considerable amount of evidence in support of the metamodal hypothesis. For our purposes, though, what matters is the implication for rationalism. Notice that to the extent that the brain is organized in this way, we have grounds to suppose that the functional specificity associated with particular regions of the brain is not always a product of general-purpose learning. The

[8] For related evidence that visual experience isn't needed to establish the ventral/dorsal large-scale organization of the visual system, see Striem-Amit et al. (2012).

[9] Other work shows that the same sort of constrained plasticity exists when portions of the *auditory* cortex are given visual inputs in cases of congenital deafness (e.g., Bola et al. 2017).

reason why the hMT+ computes direction of motion, for example, can't be because this is required by its visual input; it performs the same function in the complete absence of visual input in the congenitally blind. Rather, the most plausible explanation of its functional specificity is that this brain region is initially organized for computing direction of motion, and this results in its selecting visual input when visual input is available because visual input is optimal for the computations it performs.[10]

The evidence that we have just reviewed regarding spatial representation argues for a variety of characteristically rationalist psychological structures pertaining to spatial position, direction, and motion being part of the acquisition base. Arguably, these will include both representations and processing mechanisms. Whether the representations involved here are taken to be conceptual will depend on how the conceptual/nonconceptual distinction is drawn. But regardless of whether they themselves are deemed conceptual, these representations (and the other characteristically rationalist psychological structures that this and related work argues for) almost certainly play a major role in the acquisition of a variety of representations that are uncontroversially conceptual, pertaining to

[10] We should emphasize that we are not claiming that the function of every brain area is fixed independently of experience, or that many brain regions exhibit no plasticity. Nor are we claiming that brain areas always support a single function (even individual neurons in simple organisms have been shown to be capable of performing multiple functions; see, e.g., Li et al. 2014), or that the primary function of a brain area is unaffected by sensory input. All theorists who adopt a broadly materialist approach to the mind suppose that psychological changes are realized by changes in the brain, and so every brain region will exhibit substantial neural plasticity. Concept nativists embrace cognitive development and change just as empiricists do, taking individuals to undergo psychological change throughout their lifetimes—some of which is supported by characteristically rationalist learning mechanisms, some by characteristically empiricist learning mechanisms, and some by a mix of characteristically rationalist and characteristically empiricist learning mechanisms. We highlight these basic points because critics with preconceptions about rationalism and empiricism that are at odds with the views as we have characterized them might easily overlook them. For example, in a critical commentary on Laurence and Margolis (2015), Plebe and Mazzone (2016) seem to misunderstand our view by overlooking that we wholeheartedly accept these sorts of environmental influences on the brain. They repeatedly point out that the environment is important to brain development, suggesting that this truism—which we of course endorse—somehow undermines our concept nativism. But no one on either side of the rationalism-empiricism debate denies that experience has an important impact on the brain. We should also note that their critique doesn't actually attempt to address any of the positive evidence we cite in support of the argument from neural wiring. Instead, they present a general argument against the possibility of rationalist accounts of conceptual development based on the fact that there are gross similarities in neural tissue across the cortex, arguing that this shows that cortical tissue is functionally unconstrained. However, the existence of gross similarities in cortical tissue is perfectly compatible with functional specialization. They also argue that the timing of brain development is inconsistent with concept nativism, claiming that "the basic development of connectivity [in the brain] takes place early, before the period infants acquire most of their concept[s]" (p. 3904). But this just begs the question against concept nativism. Concept nativists provide evidence that many concepts are acquired substantially earlier than empiricists suppose (see the argument from early development in Chapters 8 and 9). And where concepts are acquired later in development, they may still be acquired on the basis of rationalist learning mechanisms. In this case, what matters to concept nativism isn't the timing of when a concept is acquired. The critical question is whether the learning mechanism traces back to characteristically rationalist psychological structures in the acquisition base (see Chapter 2).

position, direction, and related concepts, such as PATH, MOTION, LOCATION, TOWARDS, AWAY, LEFT, RIGHT, FRONT, BACK, ABOVE, and BELOW.

In the context of a view like ours, in which many other concepts in many other domains are either innate or acquired via rationalist learning mechanisms, these concepts add to the case for concept nativism by broadening the range of concepts it covers. At the same time, however, it's worth noting that these concepts are among the types of concepts that an empiricist who accepts some innate concepts might posit, as well.[11] And, as we will see in Chapter 21, Jean Mandler argues for a moderate form of empiricism that does just this, developing an account that aims to explain the origins of all concepts based on a small stock of innate spatial concepts which (sometimes combined with bodily feelings) work in conjunction with a particular type of domain-general learning mechanism.

Spatial concepts are not the only type of concepts where an argument of the type we have been discussing can be given, however. Other work of this type, focusing primarily on constrained plasticity in the visual cortex, makes a case for a rationalist account of concepts in other conceptual domains. For example, Mahon et al. (2009) examined the ventral visual stream's representation of living versus non-living kinds. It is well known that the ventral visual stream exhibits neural specialization for these differing categories, with the representation of artefacts (e.g., tools and houses) in medial regions and the representation of living animate things (e.g., animals and faces) in lateral regions. A common empiricist theory in neuropsychology is that this medial-to-lateral organization stems from the differing visual features associated with these categories (e.g., Rogers et al. 2005). The basic idea behind this proposal is that general-purpose learning establishes differentiated neural-representational systems for artefacts and living kinds because experience with exemplars from these two general categories produces visual experiences with different characteristic features. One way to test this supposition is to compare the brain activation in sighted and congenitally blind participants using a common task that is known to generate ventral visual stream activation in sighted participants. This is exactly what Mahon et al. did, asking sighted and congenitally blind participants to make size judgements upon hearing words for artefacts and animals.

Now if representations of living and non-living kinds were organized as they are in the ventral visual stream of sighted participants because of a response to some measure of *visual similarity*, it would be deeply surprising to find the same fine-grained functional differentiation along the medial-to-lateral axis among the congenitally blind. But this is just what was found. Thus it is highly unlikely that

[11] Recall from Chapter 2 that empiricism is compatible with the acquisition base containing a very limited number of innate concepts or other characteristically rationalist psychological structures (particularly when such concepts are relatively concrete/non-abstract, as these concepts are) in addition to the domain-general learning mechanisms that form the core of any empiricist account.

the general organization of these areas is driven by visual differences between artefacts and living kinds that general-purpose perceptual learning manages to pick up on. As Mahon et al. note, the similar pattern of activation in sighted and congenitally blind participants suggests "that the organization of the ventral stream innately anticipates the different types of computations that must be carried out over objects from different conceptual domains" (Mahon et al. 2009, p. 403). Similar results regarding the representation of action verb meanings (e.g., "run") (Bedny et al. 2012), and tools (Mahon et al. 2010) suggest that the brain regions underlying a variety of evolutionarily important conceptual domains exhibit constrained plasticity in development.[12]

So far we have focused primarily on examples of representations in the visual cortex. And it is certainly noteworthy that the functional/representational organization of the visual cortex is retained notwithstanding the complete lack of visual input. But just as important are instances of higher cognitive *amodal* neural systems retaining their functional specificity despite a lack of visual input. After all, these systems still depend on sensory information, so the information they draw upon will differ in dramatic ways when visual information is not available.

Consider the impact of blindness on the development of mentalizing abilities. Blind individuals lack access to many of the perceptual cues that are typically associated with other people's mental states, including, for example, their facial expressions, direction of gaze, and body posture. Blind individuals also can't rely on first-person experience to understand other people's visual perception of events. However, despite these radical differences, the location of the neural substrates for mentalizing in early blind individuals (including congenitally blind individuals) is the same as for sighted individuals (Bedny et al. 2009).[13] Notice how unexpected this is on the assumption that mentalizing is acquired largely on the basis of general-purpose processes that are, in the first instance, driven solely by low-level perceptual cues. Why would the same cortical areas end up with the same peculiar functions given such grossly different access to the evidence for mental activity? In contrast, this constancy in function is naturally explained on the hypothesis that the functions realized by these cortical areas and their

[12] See also Chapter 22 for discussion of the embodied cognition research programme, where we discuss related work on the representation of hands, feet, tools (and actions involving hands vs. feet) in congenitally blind individuals and individuals who were born without upper limbs.

[13] Bedny et al. (2009) offer a qualification to these conclusions, citing work that suggests that an understanding of false belief develops at a later age in blind individuals than sighted individuals, perhaps as late as 8 years old (e.g., Peterson et al. 2000). However, this work is based on traditional false-belief tasks rather than non-traditional false-belief tasks (see Chapter 9 for discussion of this distinction). Since sighted children have been shown to pass non-traditional tasks at much younger ages than traditional tasks, it would be very interesting and revealing to determine if blind infants can pass a non-visual, non-verbal non-traditional false-belief task at a comparable age to sighted infants. As far as we know, though, all non-traditional false-belief tasks that have been run on infants to date have been visually based tasks, so new tasks would need to be designed in order to test this possibility.

high-level patterns of connectivity are largely fixed independently of perceptual input, reflecting a psychological capacity that traces back to a rationalist local acquisition base involving innate characteristically rationalist psychological structures.

All of the cases of constrained plasticity that we have considered to this point have involved *preserved function* in the face of dramatic variation in sensory input. In such cases, one might expect dramatic changes in function if brain plasticity were not highly constrained. We turn now to a second type of case of constrained plasticity, namely, cases involving cognitive and conceptual impairments which are due to focal brain damage or genetic disorders/anomalies.[14] Such impairments argue for constrained plasticity because they show that the brain is *insufficiently plastic* to compensate for certain difficulties, despite ample opportunity for neural reorganization based on exposure to relevant features of the environment.

Notice, for instance, that empiricists who hold that the brain's plasticity is unconstrained should predict that early focal damage to a brain region that is normally associated with a specific representational capacity—the FFA's representation of faces, for example—shouldn't be permanently debilitating. On such empiricist views, this damage should be compensated for by the reorganization of neural functions in other brain regions given the continued exposure to relevant external stimuli. The visual system's regular exposure to faces after damage to the FFA should be enough for a system for representing and recognizing familiar faces to develop somewhere else in the brain, especially given how important face recognition is to everyday life. So it should be surprising to an empiricist who holds that the brain's plasticity is unconstrained if the inability to represent faces were to persist. Moreover, it should be all the more surprising if an individual with early focal damage were unable to recover the ability to represent faces but ended up being perfectly capable of representing and discriminating other types of comparably complex objects.[15] In general, then, cases in which the brain is

[14] While we will be using terms like "genetic disorder" and "genetic anomaly" in conformity with the widespread use of these terms in cognitive science, it should be noted that they are problematic for several reasons. One is that, although genetic factors can profoundly affect development, it should nonetheless be borne in mind that no traits are solely genetic—all phenotypic traits depend on both genetic and environmental factors (as explained in Chapter 3). The other is that "disorder" in this context, and also in such terms as "autism spectrum disorder", has a negative connotation. However, many would argue that atypical conditions such as autism spectrum disorder (also referred to as *autism spectrum condition* or *autism*) shouldn't be thought of in this way and that they should instead be understood to simply be part of the larger landscape of normal human variation.

[15] Of course, neurological disorders typically don't affect a single functional system in isolation (e.g., the FFA's representation of faces) but rather involve a variety of co-occurring deficits. For example, a stroke may result in damage to functionally distinct yet physiologically related brain areas that are equally dependent on the impeded blood flow. Nonetheless, cognitive deficits are sometimes quite specific. Prosopagnosia, a deficit in the ability to recognize faces, may be accompanied by other forms of agnosia, but can also occur as a selective deficit in which the impairment is specific to faces. Such individuals may be unable to recognize faces or discriminate them from other faces while at the

unable to compensate for impairments of this kind—especially when there is considerable exposure to relevant environmental stimuli—constitute another type of argument against empiricism.

How might an empiricist respond to this objection? A common response is to claim that what might appear to be a category-specific deficit deriving from an impairment to an innate domain-specific mechanism actually traces back to more general deficits of one kind or another. Consider, for example, deficits that appear to be specific to the representation of living kinds in semantic memory (memory related to general knowledge) (Capitani et al. 2003).[16] This includes patients with significant impairments for representing animals (elephant, duck, etc.) in contrast with artefacts (pen, key, etc.). A standard empiricist explanation of such cases holds that semantic memory isn't organized in terms of a categorical distinction between the living and the non-living (or animals versus artefacts), but instead is organized in terms of the properties that exemplars of particular categories possess. Different types of properties are taken to figure more prominently in the representation of categories of living versus non-living kinds. In one influential account, *visual properties* are taken to be more prominent for living kinds, and *functional properties* for non-living kinds. If this account were correct, then focal damage to the neural substrate for the representation of visual properties would disproportionately affect living kinds, while damage to the representation of functional properties would disproportionately affect non-living kinds (Warringon and McCarthy 1983; Farah and McClelland 1991). A standard rationalist approach, in contrast, maintains that semantic memory *is* organized in terms of a categorical distinction between living and non-living kinds. On this view, these innate categories are subserved by dedicated neural circuits, as are a number of other category types with particular evolutionary significance, such as animals, tools, faces, and food (Caramazza and Shelton 1998; Mahon and Caramazza 2009).

As our interest is in the rationalism-empiricism debate, a particularly important type of case to consider in evaluating these two different types of explanation is one in which a category-specific deficit in early development results from neural damage or a genetic anomaly.[17] Farah and Rabinowitz (2003) documented the case of Adam, who sustained brain damage when he was just 1 day old. At age 16,

same time have no difficulty recognizing complex objects that aren't faces or discriminating them from comparably similar objects (Busigny, Graf et al. 2010; Busigny, Joubert et al. 2010; Rezlescu et al. 2014). This sort of specificity regarding a representational deficit can persist in spite of many years of exposure to relevant stimuli (faces in the case of prosopagnosia) and a very strong vested interest in the subject domain.

[16] For more on semantic memory, see Chapter 19.

[17] The strategy we are employing here is to use findings regarding atypical development to support a view about neurotypical development. We should note that Annette Karmiloff-Smith and other neuroconstructivists have questioned this approach and its ability to support rationalist accounts. In Chapter 20, we explain their objections to this approach and show that they are unfounded.

Adam was tested for his knowledge of living and non-living kinds, and a significant difference between the two was found. Adam had a severe impairment for knowledge regarding living kinds (responding to testing at chance levels), yet his performance was normal or near normal regarding non-living kinds. His difficulty with living kinds was also comprehensive in that it affected visual and non-visual properties alike, while his knowledge of non-living kinds (both visual and non-visual) was spared. Consequently, Adam's psychological profile doesn't fit well with the empiricist explanation of category-specific deficits in terms of selective damage to the representation of a given type of property (in this case, to visual properties).

What's more, Adam's case speaks directly to the limitations on neural plasticity in cognitive development. Despite the fact that the neural damage occurred very early in development, and despite the fact that Adam had years of experience in infancy and childhood in which other aspects of his psychological development proceeded normally, his brain was unable to compensate for the damage it had sustained in terms of representing living kinds.[18] As Farah and Rabinowitz put it:

> phrased in terms of Adam's surviving brain tissue, despite its adequacy for acquiring semantic memory about nonliving things, it could not take over the function of semantic memory for living things. This implies that prior to any experience with living and nonliving things, we are destined to represent our knowledge of living and nonliving things with distinct neural substrates. (Farah and Rabinowitz 2003, p. 408)

In a related study, Farah et al. (2000) examined a different specific representational deficit in the same individual, namely, Adam's difficulty with faces. At the age of 16, Adam had the classic profile of prosopagnosia—lesions in occipitotemporal cortex (bilaterally), with a severe impairment in the ability to recognize faces relative to good, though not perfect, object recognition abilities. As with the living/non-living distinction, this uneven cognitive profile raises the question of why other neural tissue was unable to compensate for the damaged neural tissue—a striking lack of plasticity—especially given the obvious importance of face recognition in daily life (Pinker 2002).

Now there are a number of possible explanations for why the representation of faces might be impaired, just as there are different possible explanations for the selective impairment to the representation of living (or non-living) kinds.

[18] A related type of case occurs with the inability to recognize voices—an impairment known as *phonagnosia*. In one study of this impairment, Garrido et al. (2009) report an instance of developmental phonagnosia in which the 60-year-old individual had a long history of selective impairment for recognizing voices that she was unable to overcome, despite having normal hearing abilities and despite performing well on tests for speech perception, vocal emotions, music, environmental sounds, and faces.

Duchaine et al. (2006) addressed this issue by examining another patient, Edward, who suffered from developmental prosopagnosia. This study took advantage of the opportunity to test on a single individual all of the alternatives to the rationalist domain-specific explanations that have appeared in the face-perception literature. Among the empiricist explanations considered were the possibility that Edward suffered from a general difficulty regarding the representation of individuals within a category, a general difficulty with holistic processing, a general difficulty with configural processing (i.e., representing the spacing between features), and a general difficulty in acquiring expertise for object categories. For example, the configural processing explanation was evaluated by having Edward make same-different judgements for photographs of faces and houses that had been digitally altered. The distance between the eyes or windows was changed, or these features themselves were replaced with similar features in the same relative spacing. In this case, Edward's performance was normal for detecting changes to houses, but dramatically worse (three standard deviations below the mean) for detecting commensurate changes to faces. Likewise, the expertise hypothesis was evaluated using corresponding face- and body-matching tests, in which the goal was to identify which of two rotated faces or headless bodies matched a target. Here, too, Edward had great difficulty with faces, but his performance with bodies was normal—in fact, he scored at the high end of the normal range for body recognition. These and the results from Duchaine et al.'s other tests indicate that Edward's difficulty is genuinely face specific, and consequently that there are face-specific developmental mechanisms that may be selectively impaired.

Edward's impairment (unlike Adam's) is most likely the result of a genetic anomaly. Though not all genetic anomalies that result in representational deficits are as focused as prosopagnosia—most result in uneven but often predictable profiles of spared conceptual abilities and impairments—they can still provide an excellent source of evidence regarding the limits on the brain's plasticity. For example, individuals with Williams syndrome, which involves a rare genetic anomaly (Schubert 2009), exhibit severe deficits in certain types of reorientation tasks which rely on geometrical representation[19] but have relatively spared face recognition abilities (Bellugi et al. 2000; Lakusta et al. 2010), and they have intact biological motion representation in spite of other types of motion representation deficits (Jordan et al. 2002; Reiss et al. 2005). These patterns are highly unexpected if we assume that the brain's plasticity is relatively unconstrained and that the development of these abilities is driven by sensory experience and domain-general learning.

[19] We discussed this briefly in Chapter 1. Recall that in a typical reorientation task, participants in a rectangular room are shown the hiding place for an object and are gently spun around until they become disoriented. They are then asked to locate the object.

One particularly well-studied and illuminating case is the impairment to mentalizing abilities found in individuals with autism spectrum disorder (ASD).[20] In a groundbreaking early investigation, Baron-Cohen et al. (1985) examined three groups of children on a traditional false-belief task—neurotypical preschool children, children with Down syndrome, and high-functioning children with ASD. The false-belief task they used was the Sally-Anne task mentioned earlier, in Chapter 9, in which a protagonist (Sally) places a toy in her basket only to have it moved (by Anne) to another location (a box) while she is away from the scene. When Sally returns, participants are asked where she will look for her toy. Neurotypical children and Down syndrome children both answered correctly, saying that she will look in the basket (where Sally should falsely think that it is), while children with ASD overwhelmingly gave the incorrect response, saying that she will look in the box (where it actually is).

Subsequent work has shown that this failure is specific to the understanding of belief and is not part of a general difficulty with understanding representation (Leslie and Thaiss 1992). In this study, neurotypical preschool children were compared with high-functioning children and adolescents with ASD—this time using both false-belief tasks and structurally similar tasks with photographs and maps. (In a *false-photograph task*, a Polaroid photo is taken of an object in one location, only to have the object moved before the photo is developed. Then the question asked is where the object will be in the photograph.) The result was that the participants with ASD performed rather well on the false-photograph and false-map tasks, despite poor performance on false-belief tasks. By contrast, children who didn't have ASD found the false-photograph and false-map tasks more difficult than the false-belief task.

Further studies have shown that children with ASD not only have difficulties with traditional false-belief tasks, but they are also unable to anticipate an actor's actions when presented with evidence of the actor's false belief in a non-traditional spontaneous-response task (Senju et al. 2010). This is not due to a general inability to understand action, as they correctly predict an agent's actions when the agent doesn't have a false belief, and are able to correctly attribute goals to an agent even when the agent fails to achieve his goal (Carpenter et al. 2001). Further, this sort of impairment persists into adulthood. Adults with ASD who can correctly answer explicit questions about what an agent with false beliefs will do nonetheless fail to spontaneously anticipate that an agent will act the same way in a live situation (Senju et al. 2009). This suggests that they are solving

[20] There is evidence that numerous different genetic anomalies are associated with ASD (see, e.g., Huguet et al. 2013). ASD may also not be a single unified condition, but rather a collection of related conditions with overlapping symptoms. For this reason, some researchers now speak of autism spectrum *disorders* (or ASDs) instead of autism spectrum disorder.

traditional false-belief tasks using consciously formulated rules that substitute for an intuitive understanding of the source of action. Other studies suggest a similar conclusion. Neurotypical adults modulate their behaviour when they are observed because of the potential effect on their social reputation (e.g., giving more money to a charity in the presence of others than when alone). In contrast, high-functioning adults with ASD don't modulate their behaviour in this way (Izuma et al. 2011). Likewise, neurotypical adults in large-scale Western societies take into account the absence of negative intentions when formulating a moral judgement pertaining to someone who accidentally causes a negative outcome. Here, too, high-functioning adults with ASD in the same societies behave differently, treating cases with and without negative intentions in the same way (Moran et al. 2011).[21]

Thus, a convergence of evidence suggests that ASD is associated with a selective representational impairment, one that affects the formation and use of certain mental state concepts but not other concepts of comparable difficulty. And just as with Adam's impairments with living kinds and faces, this impairment persists despite ample exposure to relevant stimuli; the brain appears to be insufficiently plastic to overcome these core deficits associated with ASD.

We have seen that a variety of neurological data support the idea that neural plasticity is not as open-ended as concept nativism's critics often suppose. Significant aspects of neural development supporting the conceptual system is substantially unaffected despite dramatic differences in sensory input (e.g., differences due to congenital sensory deprivation). And in other cases, the brain is insufficiently plastic to compensate for focal damage or genetic anomaly, resulting in lifelong conceptual impairments even when the disruption occurs very early in development and is followed by years of experience of stimuli from the conceptual domain that a brain with unconstrained plasticity might use to overcome the deficit.

We conclude that, while there is certainly a great deal of evidence for plasticity in the brain, neural plasticity is constrained in a way that is best explained by an overall rationalist framework. The brain is not comprised of an equipotential network that is sculpted into differentiated functional units through general-purpose learning. Much of its development is grounded in a differentiated and complex arrangement of distinct neural-representational systems that are specialized for processing specific types of information. We have illustrated this underappreciated fact with a number of examples that show how the argument from neural wiring can play an important role in the case for concept nativism. These examples point to an assortment of characteristically rationalist psychological

[21] The right temporo-parietal junction (rTJP), which is known to be a critical mentalizing brain area (Koster-Hale and Saxe 2013), is particularly involved in modulating moral judgements according to whether a harm is accidental or intentional (Buckholtz et al. 2008; Young and Saxe 2009). Interestingly, the normal spatially distinct responses within the rTJP for accidental versus intentional harms is absent in adults with ASD (Koster-Hale et al. 2013).

structures in the acquisition base that make a significant contribution to the acquisition of concepts in the domains of space and motion, faces and individuals, living and non-living kinds, tools, action categories, and types of psychological states (especially propositional attitudes). Since the very idea of the brain's plasticity is routinely assumed to support an empiricist view of conceptual development, this is a striking outcome. We expect that as further research along these lines continues—research that is sensitive to the possibilities of constrained plasticity—it will only reinforce the case for concept nativism.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0013

14
The Argument from Prepared Learning

Our sixth argument for concept nativism—the *argument from prepared learning*—turns on the insight that conclusions about the character of the acquisition base can sometimes be drawn by attending to the relative ease or difficulty of learning across different conceptual domains. The aim of the argument is to identify patterns of this kind that indicate that the mind is innately prepared to specifically learn about some types of things but not others. The first step in any instance of this argument is to identify a pattern of development where facts pertaining to the relative ease of learning suggest the workings of a domain-specific rationalist learning mechanism. For example, this might be a case where learning in one conceptual domain occurs extremely rapidly (this might involve exposure to as little as a single instance in some cases), while comparable learning involving other conceptual domains may require exposure to many instances over an extended period of time. The second step in the argument addresses where this domain-specific mechanism comes from. Empiricists aren't necessarily opposed to domain-specific learning mechanisms. As was explained in Chapter 2, their primary opposition is to domain-specific learning mechanisms that aren't acquired solely through the operation of more fundamental domain-general processes. To the extent that this sort of domain-general learning can be ruled out or deemed unlikely, we may conclude that the domain-specific mechanism in question is a rationalist domain-specific learning mechanism. The fact that there would then be a rationalist account for the concepts it draws upon and the concepts whose acquisition it supports would thereby lend support to concept nativism.

While the argument from prepared learning has a venerable history in cognitive science and was particularly prominent in the 1960s and 1970s, it has received far less attention since then. Some of the case studies that we will discuss are based on research that is not typically seen as being linked to this earlier work, and these case studies draw on research that hasn't been explicitly formulated in terms of the argument from prepared learning. Although this form of argument has been neglected and arguably has played a relatively minor role in recent developments in the rationalism-empiricism debate, we think it is an important one and that it has a great deal of untapped potential. One of our aims here is to call more attention to this form of argument and to encourage researchers to further explore what it might tell us about the contents of the acquisition base.

In Chapter 4, we discussed one important and relevant line of research that factored into early discussions of prepared learning: Garcia and Koelling's (1966)

investigation of how rats learn which foods are harmful and should be avoided. Garcia and Koelling found that rats rapidly learn to associate their nausea with a particular correlated taste but fail to associate it with equally salient correlated sights or sounds even though they are perfectly capable of learning to associate such visual and auditory stimuli with electric shocks. As we saw in Chapter 4, this suggests that rats' avoidance of certain foods isn't driven by general-purpose learning. If the learning were grounded solely in domain-general associative learning, then in principle any discriminable stimulus that is comparably well correlated should be equally associable with nausea. But this is clearly not the case. The underlying learning mechanism is prepared to link specific types of events: *nausea* and antecedent *taste experience*. Garcia and Koelling put the point using the language of reinforcement learning, remarking that different "reinforcers are not equally effective for all classes of discriminable stimuli", and speculating that "evolution favored mechanisms which associate gustatory and olfactory cues with internal discomfort" (p. 124).

Not long after the appearance of Garcia and Koelling's (and other related) findings, Martin Seligman presented a landmark discussion of the argument from prepared learning in the context of a sweeping critique of the assumption that all learning is governed by the same domain-general principles. Seligman pointed out that the learning theorists at the time weren't paying sufficient attention to data from their own research demonstrating that not "all events are equally associable" and that learned associations fall along "a continuum of preparedness" (Seligman 1970, p. 406). For example, it was known that pigeons easily learn to peck a lighted key when the provision of food is associated with the key being illuminated, and that this is true even when the pecking itself is completely superfluous to obtaining the reward (i.e., food is automatically provided following the key being illuminated). For Seligman, the fact that the pecking-food association is ineffectual but learned anyway was a clear sign that pigeons are prepared to associate the action of pecking with obtaining a food reward. In contrast, although cats can be trained to press a lever to open a door to escape an enclosure, it was known that they have great difficulty learning to open a door by licking or scratching themselves, even when this behaviour is perfectly correlated with the door opening—an instance of what Seligman described as their being *counterprepared* to learn an association (in this case, the association between licking/scratching themselves and the door opening).

Like Garcia and Koelling, Seligman turned to broad evolutionary considerations to explain why cats and other animals would be prepared (or counterprepared) for different types of learning. He pointed out that, just as each species' sensory systems may be more or less suited to the conditions of a learning experiment, the same holds for its learning mechanisms, which likewise have "a long specialized evolutionary history" (p. 407). For example, cats, as a species, would have had an evolutionary history in which licking their own bodies wouldn't have

had any effect on nearby obstructions, but one in which pawing obstructions would sometimes allow them to escape. Thus it shouldn't be all that surprising that cats have trouble learning that licking their body causes a door to open.

In the examples that follow, we too will mention broad evolutionary considerations in connection with the argument from prepared learning, but these don't directly factor into the structure of the argument. All that matters to whether the argument from prepared learning applies is whether facts about the relative ease or difficulty of learning something counts as evidence for a domain-specific rationalist learning mechanism. Still, thinking about evolution can often be a useful heuristic for identifying potential instances of prepared learning. The starting point when using this strategy is to ask about the types of learning mechanisms that would have been favoured by natural selection given the adaptive problems our ancestors faced and the likely fitness consequences of competing domain-specific and domain-general solutions.[1] If a good argument can be made that a domain-specific solution might have outperformed domain-general alternatives, we can then directly investigate whether such a mechanism exists using the logic of the argument from prepared learning—by looking for the characteristic pattern of relative ease of learning in a given domain compared to other domains. Further work on food aversion illustrates how even very basic facts concerning an adaptive problem can guide productive research on prepared learning. For example, as we noted in Chapter 4, Wilcoxon et al. (1971) reasoned that a taste-based aversion system, rather than a vision-based aversion system, makes sense for rats (nocturnal foragers with poor vision). However, the opposite is true for birds that rely heavily on vision for foraging. This led them to hypothesize that such birds would differ from rats in being predisposed to rapidly associate a *visual* cue with illness, which they went on to demonstrate for Bobwhite quail using an experimental protocol similar to Garcia and Koelling's.

Two further aspects of the argument from prepared learning should be mentioned before we turn to the case studies that will occupy the bulk of this chapter. First, like the other six arguments for concept nativism in Part II, the argument from prepared learning takes the form of an argument to the best explanation. In all of the examples that we will discuss, the claim is that facts pertaining to the relative ease of learning in a given domain compared to other domains are best accounted for if we postulate the existence of a domain-specific rationalist learning mechanism, not that these facts provide a deductive proof that there is such a mechanism.

Second, in many instances, there is a question about exactly what domain-specific rationalist learning mechanism is involved in the pattern of prepared

[1] Adaptive problems are enduring conditions in evolutionary history that presented opportunities or obstacles bearing on an individual's evolutionary fitness (Tooby and Cosmides 2005).

learning. Given the current state of the data, we can often see that some type of domain-specific rationalist learning mechanism underlies the pattern even though further research is clearly needed to clarify the character of the rationalist learning mechanism in more detail.

With these preliminaries out of the way, let's delve further into the argument from prepared learning by examining a few of the areas where the argument from prepared learning is promising regarding the human mind and where there is reason to suppose that it can add to the case for concept nativism. We will focus on domains involving the representation of food, animals, purpose, and emotions. We will begin with the representation of food.

Food is an especially interesting focal point for accounts of human learning because humans are food generalists who have to learn what to eat in the face of many potential dangers, including toxins, pathogens, and diets that provide insufficient calories or that are missing crucial nutrients (Rozin 1990). Much of this learning also has to take place relatively early in life, with food selection being absolutely critical to survival. It is somewhat surprising, then, that the representation of food has received little attention in the rationalism-empiricism debate.

Some of this neglect may be because of the fact that children undoubtedly benefit from social learning. They are guided to eat suitable things by their caretakers and by observing what is eaten by others in their community. Since social learning is used to learn about such a wide range of domains (clothing, religious rituals, gender roles, technology, etc.), it is often assumed to be antithetical to domain-specific learning. But much the same is true of associative learning, and we have already seen that associative learning is fully compatible with rationalism, and that not all associative learning is domain general—in some instances it is constrained by domain-specific rationalist learning mechanisms. So, the fact that social learning is involved in learning about food does not argue against adopting a rationalist perspective in this domain.

One area where there is reason to expect to find a food-related rationalist learning mechanism is the representation of meat. Compared to other foods in ancestral times, meat would have been a highly concentrated source of nutrients, including protein and fat, and there is evidence that meat played a role in human evolution as an essential component of the human diet prior to the advent of agriculture (Bunn 2007). At the same time, however, meat would have been a particularly dangerous food source. Unlike many plants, which advertise their toxins to warn off being eaten, pathogens present in meat are often undetectable. The meat of a dead animal, no longer protected by the animal's immune system, is also ripe for the rapid proliferation of pathogens. Taken together, these two factors—the high nutritional value of meat and the large risk associated with contacting and ingesting meat—may have etched into our minds a deep ambivalence towards meat,

making meat a desirable food but also one that readily elicits misgivings (Fessler and Navarrete 2003).[2]

Given the dangers associated with meat, one possibility is that we possess domain-specific rationalist learning mechanisms that regulate our social learning about meat and that heighten meat's potential as an object of disgust, influencing learned food selection and learned food aversion (Tybur et al. 2016). These mechanisms ought to create more resistance to eating new types of meat than other foods and make it is easier to learn to avoid meat. Fessler and Navarrete (2003) report a range of findings that support these predictions. These include the finding that meat accounts for more acquired food aversions than any other food category (Mattes 1991), that reasoning is ineffective at overcoming neophobia towards meat but not neophobia directed at non-meat foods (Martins et al. 1997), and that, in many cultures, animal products are prototypical elicitors of disgust (Fessler and Navarrete 2003).

This work suggests that there is a differential pattern of learning associated with meat which is potentially tied to a rationalist learning mechanism. Experimental work confirming this is quite limited at present, in part because of constraints on the sorts of direct experimentation that can be done in the investigation of acquired food aversions. Experimenters can't freely explore people's learned aversions based on eating different types of foods that have been surreptitiously contaminated. Still, there are ways to test for aspects of a learned aversion in a laboratory setting. For example, Tybur et al. (2016) showed participants images of different meat and non-meat foods along with images involving cues for pathogens. They were then asked about such things as how nutritious it was and how tasty. The association of pathogenic cues had a significant negative effect on their judgements of how much they desired to eat the meat, but didn't affect how much they desired to eat the other foods (and there was no effect on judgements about objective features of either meat or non-meat foods). The differential pattern of learning here—readily developing a reduced desire to eat meat when visually linked to pathogen cues, but not readily developing a comparably reduced desire to eat other food similarly linked to pathogen cues—points to a potential argument from prepared learning concerning representations of meat and the domain of meat.

For such an argument, however, we need evidence that this differential pattern is one that is grounded in a rationalist domain-specific learning mechanism. In principle, the evidence in Tybur et al. (2016) might be explainable in terms of domain-general learning, and Tybur et al. didn't directly rule out this possibility. One reason for thinking that this pattern shouldn't be explained in such terms, however, comes from related evidence from non-human animals. For example,

[2] Of course, this evolutionary reasoning isn't enough to show that a rationalist account of MEAT and related concepts is true—nor is it meant to—just that it is worth exploring. As we explained in Chapter 4, evolutionary hypotheses bearing on the contents of the acquisition base are just that, hypotheses. They must be independently tested and assessed relative to the same evidential standards as hypotheses that aren't generated on the basis of evolutionary thinking.

Fessler and Navarrete (2003) observe that laboratory rats with controlled feeding histories are more easily conditioned to develop aversions to food sources that are high in protein than they are to food sources that are high in carbohydrates. It's unlikely that these laboratory animals had prior experiences that could have led to this predisposition being acquired through domain-general learning (e.g., that they all had more incidents of illness following the ingestion of high amounts of protein than incidents of illness following the ingestion of high amounts of carbohydrates).[3] While more evidence would be needed to firmly establish that humans have a predisposition for differential learning patterns associated with meat that are best explained in rationalist terms, the corroborating evidence from animals provides initial support for this view.[4]

Another promising type of argument from prepared learning in the domain of food is motivated by a different sort of learning problem in human evolution. Ancestral children had to be receptive to new foods to learn what to eat and what things to even conceptualize as food. For younger children, the danger of ingesting a lethal or debilitating substance would have been mitigated by the oversight of their parents and caretakers, who would have had substantial control over what was eaten, and who could also have been observed as reliable social models for what to eat. So there would be little evolutionary pressure for very young children to be constrained regarding food selection—they could be open to any foods their caregivers provided them with. But as a growing child became more independent and more mobile, lacking constraints about what to eat and experimenting with eating new things would have meant a substantially increased risk of illness and death. Moreover, the benefits of being open to new kinds of foods would have declined on the assumption that our ancestors often lived their whole lives in the same environment and hence that a child could be exposed to all of the foods that were typically eaten in their community in one or two full cycles of the seasons. The problem, then, was to be sufficiently motivated to eat new things at just the right time—when individuals had the most to learn about what to eat and were relatively insulated from the risks of being indiscriminate.

Cashdan (1994, 1998) has proposed that this situation presented the right conditions for a food-specific rationalist learning mechanism to emerge that operates within a *sensitive period*—a period in which learning specifically within this domain is dramatically facilitated relative to other times—which tapers off when children are about 2 years old. This suggests the possibility of an argument from prepared learning in this domain. This argument would turn on a pattern of accepting new items as things to be eaten that differs systematically within versus outside of this sensitive

[3] Fessler and Navarrete (2003) mention a number of other suggestive findings with animals, including that chimpanzees and baboons hunt but don't scavenge (treating a found carcass as not being a source of food) and that olive baboons rely heavily on social cues regarding whether to eat a novel type of meat but don't use such cues in relation to novel vegetables.

[4] Such a predisposition would also help to explain why meat is the subject of more taboos across cultures than other types of food. See Fessler and Navarrete (2003) for details.

period (with new items being accepted quite easily during the sensitive period but not outside the period), and where the same sort of differential pattern doesn't occur for social learning regarding other content domains (for example, regarding the acceptance of new games or new social norms).

Cashdan discusses a number of findings that support this argument. One is that, in contrast with older children, children below the age of 2 happily put just about anything in their mouth, including objects that older children and adults recognize as non-food (e.g., paper) or consider offensive (e.g., imitation/fake dog faeces) (Rozin et al. 1986). Another is that, in her own research with middle class American children, Cashdan (1994) has found that children's willingness to try new foods peaks around 2 years old and then significantly declines.[5] This peak corresponds with the age at which the number of incidents of accidental poisoning among children in the US peaks, with the subsequent decline likely indicating the benefits associated with older children being more discriminating about what they eat. Cashdan also found that the children in her study who started on solid food at a later age—and thus typically had fewer opportunities to sample new foods in the hypothesized sensitive period—ended up with a narrower diet as older children.

This work supports the differential pattern of learning in the food domain within and outside of the hypothesized sensitive period but without directly comparing the food domain with other domains. Although there is evidence that other domains (like those that we mentioned a moment ago) don't follow the same pattern with the same type of sensitive period, it would be desirable to have more systematic evidence and more direct comparisons with the food domain. A more fully developed version of this argument would greatly benefit from having, for example, a detailed direct comparison of children's learning in the food domain with comparable types of learning in other domains in a large-scale longitudinal cross-cultural study. But the initial findings reported by Cashdan suggest that this is a potentially fruitful domain for pursing such an argument.

An interesting feature of this version of the argument from prepared learning is that the pattern of prepared learning in this case is *diachronic*. It involves a differential pattern of learning that may be best explained in rationalist terms but one that has to do with learning being easier in some periods than in others, where this pattern of relative ease is specific to a particular domain. This highlights the fact that the argument from prepared learning is compatible with a range of ways in which there can be a differential pattern of relative ease or difficulty of learning across different conceptual domains.

Another domain that is a promising candidate for the existence of prepared learning is the domain of animals. One aspect of particular interest has to do with

[5] A notable feature of childhood eating patterns is the persistent resistance to new foods by older children regardless of how much parents try to get them to eat these foods.

how people learn about dangerous animals. Here, too, basic evolutionary considerations are suggestive. In ancestral times, there would have been substantial fitness consequences regarding knowledge about environmental dangers and about dangerous animals in particular. Moreover, learning that relied largely on individual experimentation—for example, approaching new animals to see how they behave—would have been very costly compared to learning that capitalizes on other people's knowledge and previous experience. Thus there would have been a distinct advantage to there being a rationalist learning mechanism for socially mediated learning about dangerous animals.

We have already seen that there may be rationalist learning mechanisms for social learning in the food domain (ones that guide learning about what types of items to consider as food and what is safe to eat) which shape children's food preferences. In addition, there are reasons to suppose that socially mediated learning should not simply be a matter of blindly accepting whatever socially transmitted information one is exposed to. Some sources are more knowledgeable or trustworthy than others, so it would be natural to expect that learners would have biases that helped them determine how to weigh testimony and information from different sources (Henrich and Boyd 1998; Harris 2012; Kline 2015; Harris et al. 2018; Henrich and Gil-White 2001). And since not all information is of equal importance, it would also make sense for learners to have domain-specific *content biases* that motivate them to attend to some particular types of information more than others and that trigger immediate retention of this information without the need for repeated exposure or social feedback.

Reasoning along these lines, Barrett and Broesch (2012) proposed that children have an innate content bias for attending to and retaining socially transmitted information about dangerous animals. According to their proposal, children should respond differently when told that a new animal is dangerous (or not) than when told other facts about an animal. They should be more likely to remember the danger-related information than these other facts even after just a single presentation and even after a significant period of time has elapsed.

To test this proposal, Barrett and Broesch worked with children from two radically different communities, one in which there is little or no direct exposure to the dangers that animals pose (4- to 6-year-old children in an urban environment in the United States) and one in which there is a very real possibility of exposure to dangerous animals from a young age (4- to 11-year-old children in small traditional Shuar villages in the Amazonian region of Ecuador).[6] Children in both communities were presented with pictures of a number of unfamiliar animals

[6] The age range for the Shuar children was broader because the community population is smaller and suitable participants are harder to come by. The researchers dealt with this problem by testing all available children in one Shuar village who were willing to participate (forty-four children in total) and by placing the older Shuar children in the Shuar control group—a conservative arrangement, as it worked against the tested hypothesis.

and told the animal's name, whether it eats plants or animals, and whether it is dangerous or not. Crucially they were told each piece of information just once. Not much later they were tested to see what they knew about the animals when shown photographs of them again, and they were tested a second time after a week had passed. In both cases, they were compared to control groups who hadn't been told anything about the animals and so were making judgements solely on the basis of looking at the photographs (establishing the baseline for the sorts of information children could derive simply from the photographs).

The result was a significant advantage for the experimental groups (who had been told the information about the animals) over the control groups (who hadn't been told this information), but only when it came to reporting which animals are dangerous. In fact, they did markedly better than the control groups when asked whether the animals were dangerous even after a week had elapsed. But, even though the experimental group had been told the information and the control group hadn't, they showed no advantage at any time of testing for providing the animal's names or knowing what they ate. What's more, the overall pattern for the children from the different communities was quite similar despite the vast difference in their day-to-day level of risk of being harmed by an animal, suggesting that the advantage for remembering the danger information derives not just from a content bias but from an innate content bias that has been shaped over evolutionary time rather than one that is a product of individual learning. The pattern of learning here, grounded in the operation of a domain-specific rationalist learning mechanism, points towards another case of prepared learning.[7]

Given just the results from this one study, one thing that remains unclear is the precise content of this domain-specific rationalist learning mechanism. It could be a mechanism that is specifically sensitive to the domain of dangerous animals, or the domain might instead be that of dangerous living things (including toxic plants, for example), or the domain might encompass any type of dangerous thing whatsoever. Though these possibilities differ from one another in terms of the degree of alignment between the posited domain-specific resources and the target learning domain, all three possibilities involve some form of domain-specific rationalist learning mechanism and hence support a rationalist account (see Chapter 2). Nonetheless, this is an area where further work needs to be done to clarify the details of the rationalist learning mechanisms involved. As a first step, Barrett, Peterson, and Frankenhuis (2016) have started to examine how readily children learn information about danger compared to other types of information for different types of entities, including animals, artefacts, and

[7] While this work is with children, the logic of the argument in the text is not that of an argument from early development. Interestingly, though, this work does support another form of argument for rationalism beyond the argument from prepared learning we are discussing, namely, the argument from universality—since the cultures involved provide a relatively stringent cross-cultural test.

food—again, comparing Shuar children and urban American children. This study found that children in both communities show a distinct advantage for learning about dangerous animals compared to other types of danger or other facts about these kinds.[8] This is a remarkable pattern given that urban children generally face more threats from artefacts (e.g., automobiles or electrical appliances) than from dangerous animals. It suggests that there are characteristically rationalist psychological structures in the acquisition base that are responsible for there being a content bias specific to learning about dangerous animals.[9]

A related proposal that offers a particularly nice illustration of the logic involved in the argument from prepared learning is the *animate monitoring hypothesis* (New et al. 2007). According to this hypothesis, the human mind has an innate category-specific mechanism of attention that monitors the environment for animals (not just dangerous animals) regardless of one's goals so that current action can be rapidly responsive to such information. This proposal (along with closely related proposals for domain-specific rationalist learning mechanisms targeting this domain) contrasts with the view that attention is domain general, that is, that there are no category-specific mechanisms in attention. It also contrasts with the view that any category-specific effects that might be found in attention are learned via domain-general processes that have benefitted from an abundance of experience with instances of the privileged category.

Like the proposal regarding the acquisition of knowledge about dangerous animals, the motivation for looking into the animate monitoring hypothesis is rooted in general evolutionary considerations. As New et al. note, during the course of human evolution, rapidly detecting and monitoring the presence of non-human

[8] A related study in Fiji looked at older children's and adults' ability to recall a range of animal properties (Broesch et al 2014). Interestingly, it was found that the content bias for whether a type of animal poses a threat diminishes with age and that older Fijian children and adults are generally good at single-trial learning for a variety of types of animal-related information. Broesch et al. (2014) point out that this is perfectly consistent with the operation of a domain-specific content bias for dangerous animals in younger children, as there is a greater need for younger children to specifically keep track of this particular type of information in light of their less developed memory and attention capacities and their lack of experience and knowledge about the animals they are likely to face.

[9] Another type of argument from prepared learning regarding the representation of animals has been thought to show that there are innate perceptual templates that aid in detecting and responding to a few particular types of animals that constituted recurring threats in ancestral times—snakes and spiders being the two most commonly hypothesized examples. Although we take this general proposal to have some plausibility, much of the experimental work that has been claimed to support prepared learning for snakes and spiders hasn't used the type of stimuli that would be needed to exclude important competing hypotheses. For instance, a widely cited study by Öhman et al. (2001) contrasted snakes and spiders with mushrooms and flowers and found an advantage for detecting the former. But as snakes and spiders are both animals, it is unclear whether this result speaks to prepared learning for representing snakes and spiders as such or for representing animals more generally (or alternatively for representing threatening entities regardless of whether they are animals). However, for some encouraging more recent work on snakes and spiders which does include some of the needed contrasting stimuli, see New and German (2015); Gomes et al. (2017); and Van Strien and Isbell (2017).

animals would clearly have been relevant to human fitness in many ways—for example, an animal could be prey, a predator, a general threat (e.g., if startled), or a source of information about hidden dangers and resources (e.g., lairs and nests). And other humans, who of course are also animals, would obviously bear on an individual's fitness as well. Moreover, animals' locations and status would have also been subject to rapid changes compared to other aspects of the environment, making the general class of animals, including human animals, a particularly important time-sensitive category for our ancestors to be aware of. New et al. take this to suggest that there would have been selection pressure for adaptions specific to detecting and monitoring animals. We will see in just a moment that there are a number of findings that support the animate monitoring hypothesis. This work can be seen as feeding into an argument from prepared learning in light of the fact that the findings argue for there being a rationalist learning mechanism of some type that makes learning about some kinds of things (animals) easier than learning about other kinds of things (e.g., vehicles). In this case, though, the learning in question isn't, in the first instance, about dangers associated with animals. It is about whether an animal is present in the immediate environment and about where such animals are located.

How might one test the animate monitoring hypothesis? New et al. used a *change detection task*. In this type of experiment, participants are instructed to indicate if they see a change when shown a scene depicted in two nearly identical photographs that are presented in succession repeatedly, one after the other. Importantly, the only difference between the two photographs is that one is missing a single object that appears in the other (e.g., a scene of a farm that has a wheelbarrow in one photo but not in the other). As anyone who has tried a "spot the difference" challenge in a newspaper's puzzles page will know, it can be surprisingly difficult to detect these sorts of differences even when the two photographs are presented side by side and even when the object appearing in one photograph but not the other takes up a substantial portion of the image. You can look at the two photos over and over again and still not see the difference. Yet once it is pointed out, it usually seems utterly obvious.[10]

The key question for New et al. was whether people would be better at detecting changes regarding animals (including humans) compared to changes regarding other types of entities. The other categories tested were plants, small- and medium-sized manipulatable artefacts, large humanmade structures (the sort that can be used as a fixed landmark), and vehicles. The result was a clear advantage for animals. For instance, partcipants were far quicker and more accurate at

[10] This difficulty is part of a larger pattern in which people can fail to notice even major changes that happen right before their eyes and that would seem impossible to miss, for example, not noticing that the person you are talking to has been replaced by an entirely different person following a brief distraction (Simons and Levin 1998).

detecting changes involving elephants or pigeons than they were at detecting changes involving large grain silos or coffee mugs, even when the animal's presence in a scene was objectively more obscure (e.g., an elephant in the distance against a like-coloured background versus a distinctively coloured coffee mug in the foreground of a scene).

Of course, there is always a question about whether such a response is truly driven by the representation of an animal as an animal, as opposed to the representation of changes in low-level perceptual features or other incidental aspects of the stimuli. New et al. did a number of things to address this concern. They took into account independent assessments of how interesting the various target objects were to establish that the effect wasn't merely a result of the animals being the more interesting visual stimuli. They also used inverted versions of the photos, which hampers categorization while preserving their low-level features, and found that this eliminated the advantage for detecting changes vis-à-vis animals. For these reasons, New et al. concluded that their data argue for the predicted effect, an advantage that is specific to the representation of animals and that points to the existence of a domain-specific mechanism for detecting and monitoring the presence of animals in the immediate environment.

To show that this effect stems from an *innate* category-specific mechanism (or more generally, from a rationalist learning mechanism) requires that we also rule out the competing hypothesis that it is owing to a more fundamental general-purpose learning mechanism that experience has calibrated for the detection of animals. Prior to completing this task, the participants in New et al.'s study—university students in California—would certainly have had encounters with pets and common urban and suburban animals (and, of course, copious experience with other humans). However, two aspects of the data argue against domain-general learning. One is that the study participants would have had a comparable amount of experience with vehicles, including a history in which vehicles (e.g., cars and buses) would have constituted more of a threat in fitness terms than non-human animals. Yet they were considerably faster and more accurate at detecting non-human animals. The other crucial finding is that they showed no advantage for detecting humans compared to non-human animals, despite living in an environment in which there are far more encounters with humans than non-human animals.

Our interest in this work is in how it illustrates the logic of the argument from prepared learning. Suppose for the moment that New et al. are right that they have identified category-specific effects in attention that are specific to the representation of animals. This would support an argument from prepared learning in the following way. First, there would be a pattern of relative ease of learning about one domain compared to others. Measured both in terms of speed and accuracy, the detection of changes involving an animal surpasses the detection of changes involving other types of entities. Second, there would be reason to believe that

this isn't the product of domain-general learning. If it were, then there ought to be an advantage for detecting changes to humans over non-human animals, and there would be no reason to expect an advantage for detecting non-human animals over vehicles—but neither of these are supported by the data. Rather, New et al. appear to have located a category-specific effect that is the product of a rationalist learning mechanism that is particularly geared towards animals as such. Like the rationalist learning mechanism for biological motion detection mentioned in Chapter 10, this mechanism, if it exists, would be part of a collection of rationalist learning mechanisms that contribute to the identification of animals and to the organization of incoming information about animals, supporting the acquisition of various types of concepts and knowledge related to animals, including concepts for different types of animals.[11]

As things stand, the evidence for an innate animal-specific attentional mechanism is mixed. While some subsequent research has questioned the animate monitoring hypothesis, much other research (including work using other experimental paradigms) suggests that there may indeed be a distinctive attentional advantage for identifying animals over other types of things. For example, Hagen and Laeng (2016) used a change detection task with the original stimuli from New et al. (2007) and also new stimuli that were created by replacing the animals in New et al.'s images with comparably sized artefacts. The study participants, in this case, were just as good at detecting the artefacts (in the new stimuli) as they were at detecting the animals (in the original stimuli). This raises the question of whether there was something about the placement of the animals in the original stimuli that attracted participants' attention, as opposed to the target animals being categorized *as* animals. However, the New et al. (2007) study has since been replicated using entirely new stimuli with different animal placements (Altman et al. 2016).

What's more, other researchers have found that the subtle presence of an animal in a photo interferes with the detection of an inanimate target, as would be expected on the assumption that an animate monitoring system directs attention

[11] The animate monitoring hypothesis is concerned with a category-specific effect on attention regarding the representation of animals. A closely related view is that there may also be *memory* advantages associated with representations of animals. A variety of work supports this related hypothesis too. For example, there is evidence for enhanced memory for simple shapes with animacy cues over simple shapes without animacy cues (van Buren and Scholl 2017). Another study, which examined both threatening and non-threatening animal stimuli, found a memory advantage for animates that is independent of any memory advantage for threat (Leding 2019). A memory advantage for animates over inanimates has even been found to extend to remembering words for animates versus inanimates, over short and long intervals, and independently of whether learning was incidental or directed (Félix et al. 2019). This memory advantage for animates can be viewed as part of a broader *adaptive memory hypothesis*, according to which human memory is innately prepared to facilitate the recall of fitness relevant information (Nairne and Pandeirada 2008; Nairne 2022). However, there may also be a specific link between the memory advantage for animates and the attentional advantage for animates, as it has been suggested that the memory advantage stems from a richer encoding process due to heightened attention to animate stimuli (Meinhardt et al. 2019).

to any visible animals (Altman et al. 2016). A comparable category-specific effect on attention has also been found when unexpected objects are flashed on a screen while participants are concentrating on a task (e.g., searching for a given word). Consistent with the animate monitoring hypothesis, unexpected animals are more likely to be noticed than unexpected artefacts (Calvillo and Hawkins 2016). And other work shows that animals are identified *as animals* remarkably quickly—in just over a tenth of a second. In fact, people can identify animals as animals faster than they can identify what type of animal it is they are seeing. Surprisingly, you can see that something is animal before you can see that it is a dog (as opposed to a bird) (Wu et al. 2015). People are also able to see that an animal is present before recognizing what type of scene they are looking at, one that takes place in a natural environment (where animals are more common) versus one that takes place in a humanmade environment (Crouzet et al. 2012). This last finding is significant in that it argues against the possibility that judgements about the presence of an animal are made so quickly because they are based on global scene statistics rather than the direct recognition that the item in the scene is an animal.

Further work will need to be done to fully test the animate monitoring hypothesis. But for our purposes, what matters most is how it illustrates the potential of the argument from prepared learning to contribute to the case for concept nativism. The critical point here isn't that attention for animals appears early in life (as in the argument from early development), that it is universal (as in the argument from universality), or that it is rooted in neural systems that are functionally constrained regarding what they can represent and process (as in the argument from neural wiring). It is that there appears to be a distinctive pattern in which a specific type of learning is significantly easier for one conceptual domain (animals) than other kinds of conceptual domains (e.g., artefacts), a pattern that can't be explained purely in terms of empiricist domain-general learning mechanisms.

So far, we have been working with a characterization of the argument from prepared learning that focuses on how readily something can be learned. A related phenomenon, which was also noted in early discussions of prepared learning, concerns how readily something can be *unlearned*, or how resistant it is to correction in the face of counterevidence. Ease of learning and resistance to unlearning aren't exactly two sides of the same coin. Just because a certain type of learning is achieved rapidly doesn't mean that its products should become fixed features of cognition. From an evolutionary perspective, a predisposition to learn something may arise as a solution to an adaptive problem; resistance to unlearning depends additionally on the nature of the environmental contingency that a mechanism evolved to respond to. For example, while some domain-specific learning mechanisms may have been selected to respond to features that are subject to short-term environmental variation (e.g., pollen locations for foraging bees), others may have been selected to respond to features that are either not

subject to environmental variation or where environmental variation is not likely to occur in an individual's lifetime (e.g., the centre of rotation in the night sky). It makes sense that the former should be easier to update (Rozin and Kalat 1971).

An important variation on the argument from prepared learning looks at cases that are likely to involve the latter type of mechanism, where it is relatively difficult to *unlearn* something in the face of counterevidence compared to unlearning in other domains. Just as it can be easier or harder to learn some things according to how well they match the mind's domain-specific rationalist learning mechanisms, it may be easier or harder to abandon some ways of thinking depending on how much they depend on these kinds of mechanisms. An example of this from early discussions of the argument from prepared learning is the phenomenon of filial imprinting in birds (Rozin and Kalat 1971). Newborn geese, for example, rapidly form a strong social bond with the first object that they encounter that meets certain conditions, and subsequently follow this object around. Once this learned social preference is formed, it is extremely resistant to unlearning. As Rozin and Kalat note, it makes sense that the learning involved in imprinting would have this pattern of ease of learning and difficulty of unlearning. "In the case of imprinting, we have great resistance to extinction [unlearning] in a case where clearly the environment will not vary (i.e., the species will not change), and the proper imprinting object is almost certain to be present at the time of imprinting" (Rozin and Kalat 1971, p. 480).

Our next example of the argument from prepared learning focuses on a case like this, where information in one domain is especially difficult to unlearn in the face of counterevidence. The example concerns the persistence of teleological thinking in adult cognition.

By *teleological thinking*, we mean the disposition to explain objects, properties, and events in terms of the purpose that they are taken to have. For example, it is only common-sense to suppose that a predatory wasp's stinger is for injecting venom into its prey. There presumably is a naturalistic causal story about how any particular wasp's stinger comes to grow at the tip of its abdomen, but such explanations seem intuitively insufficient to account for *why* wasps have stingers in the first place. For that, one wants to know what *purpose* the stinger serves, or what the stinger is *for*. Likewise, many people view certain significant events in their lives not as chance happenings or as merely the consequence of their own decisions, other people's actions, or other ordinary preceding causes. They view these as having a larger purpose, taking the world to possess a form of design that orchestrated the situation to culminate in this outcome ("it was meant to be"). Interestingly, children seem to be especially avid teleological thinkers. Unlike educated adults, who often restrict their overt teleological explanations to artefacts and to parts and properties of living things, studies in the United States have found that children happily embrace teleological explanations for clouds and rocks and other non-living natural phenomena from preschool until about 9 to 10

years of age. A child might say, for example, that clouds are *for raining* or that prehistoric rocks were pointy "so that animals wouldn't sit on them and smash them" (Kelemen 1999a, 1999b).

What happens later with older children and adults? Is teleological thinking readily relinquished with greater maturity, exposure to more education, or the adoption of a broadly scientific outlook on the natural world? One indication that people have trouble letting go of teleological thinking is how hard it is to form an accurate understanding of Darwinian natural selection. The core components that make up this theory aren't actually that complicated. They don't require working with an abstruse mathematical formalism, for example. What's required is appreciating how the appearance of design can be explained so long as three factors are in place: variation in a population, an environment that favours individuals with some of these variants over others, and a biological mechanism whereby these successful variants are passed on to their owners' progeny. The key to this account is that, after many generations, these factors can lead to an accumulation of changes resulting in a remarkable fit between a species and its environment, including the development of intricate biological structures that exhibit the marks of good design even though there was no one guiding the process.

As simple as the outline of this account is, people have tremendous difficulty grasping its non-teleological character. All too often, students (including biology and medical students) recount natural selection as being responsive to an animal's wants and needs. That is, they describe it as holding that organisms acquire a given trait (e.g., a longer neck or greater speed) because they need it for some purpose (to reach food or to evade predators) and that these beneficial new traits are passed on to their progeny—a misunderstanding that persists even following courses that are specifically tailored to address such common errors (Evans 2002).

What's more, in a study that examined the gains in students' understanding of the theory of evolution in a course on evolutionary medicine (which emphasizes the practical value of evolutionary thinking for understanding health and disease), teleological thinking was found to be the main impediment to gains in understanding natural selection. Students' *acceptance* of the theory of evolution was influenced by other factors, such as their parents' attitudes and religious beliefs, but these factors did not impact on their *understanding* of natural selection. So teleological thinking impedes understanding of natural selection, even in a context where such understanding is of practical value to students, and this negative influence is independent of the students' acceptance or lack of acceptance of the theory of natural selection (Barnes et al. 2017).

Further evidence that teleological thinking is hard to give up comes from a population that is professionally opposed to teleological explanations of natural phenomena—physical scientists at high-ranking research universities (Kelemen et al. 2013). It may be that many students taking science courses haven't given much thought to why teleological explanations in science are problematic, but

physical scientists are invested in an explanatory framework that shuns teleological explanations. It is a normative standard in the field that its explanations should be restricted to non-teleological physical-causal mechanisms. Still, despite this commitment and routine exposure to purely physical-causal accounts of natural phenomena, these scientists exhibit teleological thinking when asked to assess different explanations under conditions that are cognitively taxing (e.g., when asked to quickly evaluate a claim like *The Earth has an ozone layer in order to protect it from UV light*).[12] Kelemen et al. conclude from this work that "A broad teleological tendency…appears to be a robust, resilient, and developmentally enduring feature of the human mind that…gets masked rather than replaced, even in those whose scientific expertise and explicit metaphysical commitments seem most likely to counteract it" (p. 1081).

All of this work on teleological thinking provides a preliminary foundation for another instance of the argument from prepared learning, but as we noted above, in this case, the argument is focused more on the relative difficulty of *unlearning* rather than on the relative ease of *learning*. Children and adults alike are all drawn to inappropriately teleological explanations. Children give teleological explanations where their parents would explicitly reject them. Even professional scientists, who explicitly shun teleological explanations of natural phenomena, find themselves employing these sorts of explanations when under cognitive load, suggesting that the psychological mechanisms that underpin teleological thinking continue to function in the background and require cognitive effort to resist. This pattern is specific to the domain of teleological explanations. And there is reason to suppose that it is grounded in rationalist learning mechanisms, rather than acquired via purely domain-general learning mechanisms. The main domain-general alternative is that teleological explanations are instilled by parents and the larger culture, particularly in communities that promote creationist religious thought, and that teleological explanation becomes ingrained through repeated exposure and use. However, this seems unlikely given that the same overall pattern in childhood appears in Britain, which is a far more secular Western society than the United States (Kelemen 2003), and in China, a non-Western society with a recent history of institutionally enforced atheism (Schachner et al. 2017).

At this point in the research, however, it is hard to say exactly which innate psychological structures make teleological explanations so resistant to elimination. Among the options are the possibility that the acquisition base includes a schema for explaining different types of phenomena in terms of a "design stance"

[12] In related research with adults who self-identify as atheists or as being non-religious (and therefore should not see the world in terms of purpose stemming from God's design), teleological thinking also occurred when their cognition was taxed (Järnefelt et al. 2015). A belief in the world as a type of self-sustaining cosmic organism proved to be a factor, but teleological tendencies persisted even when these Gaia beliefs were controlled for.

(Keil 1992), that there is an innate domain-specific module (or a module grounded in characteristically rationalist psychological structures) for interpreting the purpose of a personal experience (Bering 2002), or that the representation of purpose in the natural world builds on the mentalizing faculty's capacity to represent the intentions of an artefact's creator (Kelemen 2011). All of these proposals fall on the rationalist side of the rationalism-empiricism spectrum.[13] They either postulate innate concepts and rationalist learning mechanisms that are specifically geared towards representing natural entities as having a purpose, or rely on rationalist learning mechanisms that use related forms of conceptualization (e.g., AGENT, GOAL, and INTENTION). For present purposes it isn't important to settle which of these is the right account. What matters is just seeing that something along these lines may be needed to explain the resilience of teleological thinking, and that resistance to eliminating teleological thinking illustrates another way in which arguments from prepared learning can be developed.[14]

We have illustrated the argument from prepared learning using examples involving the representation of food, danger, animals, and purpose. We will close out our discussion by briefly touching on one further example, this time from the domain of emotions. Emotions are often thought of in terms of the feelings that are associated with them. But at least as significant is their functional role in human psychology. Emotions have characteristic elicitors, involve distinctive physiological changes, and have specific downstream effects on cognition and action. And crucially for current purposes, emotions also serve important communicative functions that are linked to their characteristic facial and bodily expressions. Fear, for example, has the communicative function of signalling a possible threat, which can be detected in an expression in which typically the eyes are widened and the mouth is slightly open (among other things). We will

[13] In referring to the last of these as a rationalist theory, we are assuming a rationalist account of core mentalizing capacities. We discussed some of the evidence and arguments for this view in Chapters 9, 11, and 13. See also Chapter 21 for more on the representation of goals. It should also be noted that any account along these lines would also need to explain why acquiring mentalizing abilities would reliably lead to such excessive teleological thinking, and such explanations may well require further rationalist constraints or articulation in addition to a domain-specific rationalist learning mechanism for core mentalizing abilities.

[14] Another resilient pattern of thought that is at odds with the relevant science (or so this book argues) is…empiricism itself. The ubiquity of empiricist tendencies once led the eminent psychologist Lila Gleitman to quip that "empiricism is innate" (Wang and Feigenson 2019). Although we don't want to make too much of this, Wang and Feigenson have followed up on Gleitman's lighthearted remark by testing whether people tend to offer empiricist or rationalist explanations of psychological capacities that are fairly well understood in cognitive science, including capacities where the evidence strongly supports the existence of domain-specific rationalist learning mechanisms (e.g., representation of approximate numerical quantities). The results were substantially skewed towards empiricism, a pattern that held for all of the populations they sampled: children in the United States, adults in both the United States and in India, and academics in the United States (albeit with a reduction in empiricist answers among researchers in the cognitive sciences). Wang and Feigenson don't exactly claim that empiricism is innate, but they do suggest that there may be something about the nature of the mind itself that encourages an empiricist outlook.

argue that an argument from prepared learning can be made around the communicative function of emotions. In developing this argument, we'll focus on representations associated with the communicative function of pride.

Much contemporary work on the emotions has omitted pride from the short list of so-called basic emotions, which have been thought to be associated with pancultural facial expressions (Ekman 1992).[15] But emotional expressions aren't confined to the face—not for pride and not for other emotions (Keltner, Sauter et al. 2019). The prototypical pride expression includes a posture where the chest is expanded, the head is tilted back slightly, there is a small smile (just in the mouth, not a full smile that includes the eyes), and arms are akimbo or outstretched above the head.[16] This prototypical pride expression is recognized as easily as other paradigmatic basic emotions (Tracy and Robins 2008a), and its recognition isn't unique to WEIRD populations (Tracy and Robins 2008b).[17] For example, tribespeople in Burkina Faso living in a preliterate and culturally isolated society recognize pride displays in photographs of both African and American participants, providing strong evidence that the pride display is not acquired via cross-cultural transmission.

According to recent proposals that offer an adaptationist perspective, pride and the capacity to recognize pride serve a number of functions related to achievement (Tracy et al. 2010; Tracy et al. 2013; Sznycer et al. 2017; Sznycer et al. 2018).[18] Pride functions to motivate people to develop skills and accomplishments that will lead them to be more valued by others, to signal their success, and to recalibrate their expectations regarding the amount of consideration they deserve. Likewise, the capacity to recognize pride allows onlookers to identify individuals who should be accorded prestige, who it would be desirable to form alliances with, and who provide valuable models to be copied. On this general approach, the capacity to detect pride originates in a domain-specific rationalist learning mechanism for learning about people's earned success and for capitalizing on this

[15] There has always been some controversy about which emotions should be counted as basic emotions, about what sorts of tests would establish that a particular type of emotion is a basic emotion, and even about the very idea that there are basic emotions (see, e.g., Crivelli and Fridlund 2019). For discussion of these controversies and a defence of basic emotions, see Keltner, Tracy et al. (2019), which builds on the account originally given in Ekman (1992).

[16] It is worth emphasizing that while the prototypical expression of pride includes all these elements, specific instances of pride may deviate from this prototype in various way, much as specific instances of fear or sadness may lack some of the prototypical features associated with the expression of these emotions.

[17] As noted earlier, WEIRD is an acronym for Western, educated, industrial, rich, and democratic (Henrich et al. 2010). See Chapter 11 for discussion of how the study of WEIRD and non-WEIRD populations figure into the argument from universality.

[18] Tracy et al. (2010) actually distinguish between what they take to be two facets of pride, *authentic pride* and *hubristic pride*, where only authentic pride is taken to be related to earned achievement. Whether pride exhibits this two-faceted structure is open to debate (see, e.g., Holbrook et al. 2014). However, for present purposes this only affects whether the functions noted in the text pertain to *pride* or just to one of its two facets.

Figure 14.1 Spontaneous pride display following athletic success in sighted (left) and congenitally blind (right) athletes. (From Tracy and Matsumoto 2008. Photos by Bob Willingham. Reproduced with permission.)

knowledge in future social interactions through innately articulated connections (in the sense of Chapter 2) with related systems.

Now there are a number of reasons for supposing that an account along these general lines is correct (including an argument from universality along the lines that we have just noted). For purposes of illustrating the argument for prepared learning, however, a study by Tracy and Matsumoto (2008) particularly stands out. This research examined the spontaneous response to success and failure among both sighted and blind elite athletes from over thirty nations. Not only did both sighted and blind athletes produce the same pride displays after successful performance, but this held even for *congenitally blind* athletes, who would never have seen the expression before (see Figure 14.1).

Consider for a moment what an empiricist model for learning this display might look like. Since congenitally blind learners can't see others' heads tilting back, the small smile, and so on, they would have to be told by their parents, peers, or teachers how to hold themselves, or would have to be guided through direct manipulation (e.g., having the position of their arms adjusted). However, it's very unlikely that this manner of learning would reliably lead to their producing even one or two of the components of this pride display much less the full display. Even less likely is it that it would reliably lead to the same full display occurring across the many cultures examined in this study.

We take it that the most plausible explanation of why congenitally blind people across the globe spontaneously produce the same pride display in a moment of high achievement is that it is regulated by a domain-specific rationalist learning mechanism for signalling this kind of success. Of course, a signalling mechanism isn't the same thing as a mechanism for categorizing these signals. But we think it's reasonable to suppose that a domain-specific rationalist signalling mechanism that is stable over evolutionary time should go hand in hand with a domain-specific rationalist learning mechanism for interpreting these signals. In short, this work suggests that the human mind is so prepared to produce the multifaceted pride display following a moment of earned success that people do not even need to perceive the display or rely on explicit instruction to know when and how to do this themselves. And this, in turn, suggests that the mind is equally prepared for representing pride and using associated concepts, such as PRESTIGE, and regulating appropriate responses to these social signals. As a consequence, there should be domain-specific rationalist learning mechanisms for the production and interpretation of pride, and likewise for other basic emotions.[19]

This chapter has been about the argument from prepared learning. The general logic of this argument has to do with cases where there is a differential pattern of learning (e.g., learning in one domain is easier than in others) and where this pattern is unlikely to be accounted for by domain-general learning mechanisms and is instead best explained in rationalist terms. We have seen that asymmetric patterns of learning are suggestive of a rationalist basis for concepts like FOOD, MEAT, ANIMAL, DANGER, PRIDE, and PRESTIGE. We have also seen that there is an interesting variant on the argument from prepared learning that concerns cases where specific types of psychological structures are difficult to unlearn or are resistant to correction, like the teleological thinking that often distorts people's understanding of natural selection. This suggests that concepts like PURPOSE and DESIGN trace back to rationalist psychological structures in the acquisition base as well. The argument from prepared learning is one of the least appreciated arguments among our seven arguments for concepts nativism and consequently its implications for different conceptual domains have hardly begun to be explored. We hope that the examples we have given to illustrate the logic of the argument encourage researchers to revitalize this form of argument, as it has the potential to make a much more powerful contribution to the overall case for concept nativism than it has thus far.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0014

[19] We should note that the claim that there are domain-specific rationalist learning mechanisms for the identification and representation of emotions does not entail that these mechanisms or representations (e.g., PRESTIGE) are operative or present at birth, only that they trace back to characteristically rationalist psychological structures in the acquisition base (see Chapter 2).

15
The Argument from Cognitive and Behavioural Quirks

We now turn to the last of our seven arguments for concept nativism—the *argument from cognitive and behavioural quirks*. "Cognitive and behavioural quirks" is our term for surprising or unexpected facts about people's minds that are especially puzzling if it is assumed that concept learning is wholly governed by domain-general learning mechanisms but that come to make sense when the relevant concepts are taken to be innate or acquired via rationalist learning mechanisms. The argument for concept nativism turns on the fact that the best and most satisfying account of these quirky phenomena is one in which they can be seen to trace back to characteristically rationalist psychological structures, typically structures that are specific to the conceptual domain in question. Since quirks are unexpected or even mysterious on empiricist accounts—empiricist accounts don't predict them and they aren't equipped to explain them—the argument provides a distinctive type of inference to the best explanation for concept nativism. Rationalist accounts not only provide satisfying explanations of such quirks but often predict further quirky phenomena, which are corroborated by further research.

Quite often the quirky phenomena that feed into the argument are only uncovered as a product of rationalist theorizing or in the context of a rationalist research programme. This was certainly the case in early rationalist research in linguistics, where (as we noted in Chapter 1) Chomsky pointed to a wealth of puzzling linguistic patterns that had been previously overlooked because of the then prevailing domain-general perspective. Consider, for instance, the following example (Chomsky 1965):

(1) Mary expected John to leave.
(2) John's leaving was expected.

Given this pair of sentences, it would be natural to suppose that (3) and (4) should equally be possible:

(3) Mary persuaded John to leave.
(4) John's leaving was persuaded.

But (4) is clearly unacceptable in English. Likewise, if we extend the obvious pattern in (5) and (6)

(5) John ate an apple.
(6) John ate.

to (7) and (8), then (8) would mean something entirely different from what it in fact means (Chomsky 1986).

(7) John is too stubborn to talk to Bill.
(8) John is too stubborn to talk to.

Since (6) means (roughly) that John ate something or other, by analogy, (8) should mean that John is too stubborn for John to talk to someone or other, when what it actually means is that John is too stubborn for anyone else to talk to John.

Rationalist research following in the tradition of Chomsky's work in linguistics has successfully identified and explained a multitude of quirky patterns of this sort. But while rationalist theorizing is often the reason why such quirky phenomena are discovered in the first place, it doesn't particularly matter how they come to be known. All that matters is that they need to be explained and that they remain deeply puzzling if cognitive development is restricted to domain-general mechanisms.

Earlier, in Chapter 1, we briefly highlighted some cognitive and behavioural quirks connected to the phenomenon of spatial reorientation. We will begin our discussion here by revisiting these findings and showing how they form the basis for an argument from cognitive and behavioural quirks for a rationalist account of the origins of geometrical concepts. First, a reminder of the theoretical context. Recall that young children, adults under conditions of cognitive load, and many types of animals exhibit a similar pattern when searching for an object that had previously been seen being hidden in a corner of a rectangular enclosure. After becoming disoriented, they search for the object in the correct corner but also search equally in the geometrically equivalent opposite corner of the enclosure. They do this even if the enclosure has prominent featural cues that could serve to disambiguate between these two corners, such as a brightly coloured wall. Evidentially, reorientation under these conditions is governed solely by a representation of the geometry of the enclosure.

As we noted in Chapter 1, Spelke and Lee (2012), adopted a rationalist and evolutionary perspective on the problem of reorientation. This led them to reason that a well-designed system for navigation would focus on representing geometrical properties of the environment's extended surface layout rather than featural cues or other landmark information. The thinking here is that, in many natural environments, attending to the geometry of the contours of an

environment is the best way to re-establish one's bearing. The objects and markings that occur within the area may be more subject to change over short time periods or less distinctive, and attending to their many details would impose a high processing cost. In contrast, the geometry of the contours of an environment is likely to provide a reliable, stable source of information which can be represented more economically.

From this perspective, reorientation should be governed in the first instance by a rationalist learning mechanism for representing geometrical properties for purposes of navigation, and it would be predicted that individuals should reorient themselves making use of the geometrical properties of an enclosure while ignoring even extremely prominent featural cues. Based on this prediction, Spelke and Lee discovered a set of highly unexpected results. They found that 4-year-old children rely on the geometrical arrangement of a set of walls even when these "walls" are a mere *2 centimetres* high, but fail to exploit a salient arrangement of tall, dark, freestanding pillars (180 centimetres tall) or a salient floor marking producing the same geometrical pattern as the boundary imposed by the walls (Lee and Spelke 2008, 2011) (see Figure 1.2).

Pretheoretically there is no reason to suppose that the geometrical properties associated with the rectangular shape that is formed by very short walls would be more salient than ones associated with the rectangular shape formed by four tall pillars or by a large coloured patch on the floor. And if children are supposed to learn how to reorient themselves using just general-purpose learning mechanisms, as empiricists suppose, it is all the more puzzling why they should make use of the tiny walls but not the landmarks (tall pillars) or featural cues (floor markings). After all, modern children live in highly manufactured environments and spend a great deal of time indoors, where featural cues and artificial landmarks are often highly reliable navigational signals. However, when seen from the rationalist perspective that motivated Spelke and Lee's investigation, this highly surprising pattern becomes intelligible. Given the supposition that there is an innate domain-specific system for reorientation of the kind they postulate, it makes sense that participants would ignore landmark and featural cues and focus on geometrical information—even rather subtle geometrical information—regarding the area's extended surface layout. This is because the proposed mechanism for navigation isn't a general-purpose system that can make use of any type of salient information. It is a system that is built to respond to only particular types of information—namely, information about the geometry of the contours of an environment.

In the rest of this chapter, we will further illustrate the argument from cognitive and behavioural quirks by examining its applicability to three additional content domains—route selection, social categorization, and physical reasoning.

Our first example provides a further instance of the value of evolutionary thinking for predicting and discovering cognitive and behavioural quirks. In this

case, the evolutionary considerations bear on the analysis of the adaptive problems surrounding navigational decisions among different possible routes. For example, suppose that you are sitting on a bench in the park and choosing between different paths that you could take to reach the bus stop that you see beyond the tennis courts. Or suppose that you are out camping and need to collect some wood for a fire and are choosing between two wooded areas that you see in different directions (where the two wooded areas seem equally likely to have firewood). What determines how you visually experience the distance of these different routes? *Evolved Navigation Theory* approaches this question by placing it in a distinctly evolutionary frame of reference.

According to Evolved Navigation Theory, we must keep in mind that there would have been significant selection pressures in ancestral environments not only to discriminate and prefer efficient routes but also to discriminate and prefer routes that reduce risk. In addition to such risks as potential exposure to predators and hostile adversaries, there would also have been the ordinary danger of falling. Even today falling is a common cause of injury, sometimes serious injury. Falls are the second most common cause of workplace injury in the United States and the single most common cause of accidental injury among the elderly (Jackson 2009). But in the environments in which our ancestors would have been navigating, in the absence of modern treatments for injury, even a small fall—leading to a twisted ankle or an open wound—could have been fatal.[1] These considerations have led proponents of Evolved Navigation Theory to predict that *risk* should be taken into account in evaluating routes. One simple way to do this would be to estimate the distance of riskier routes as longer (proportional to the level of risk) than less risky routes. By building these navigational costs into distance estimation, a simple preference for shorter routes would allow one to automatically factor in risk in selecting what is on balance the best route.

When it comes to the risk of falling, this means that routes with vertical slopes or routes alongside vertical drops should appear longer than equidistant routes that are less hazardous. Moreover, as studies of injuries show, the level of risk is also determined by the direction one takes along a route. With vertical navigation, descending is typically more dangerous than ascending (e.g., when descending a rock face or descending a ladder). Descending often means leading with your feet as opposed to your hands, it involves poorer visibility (you can't see where to place your feet unless you do something risky like push your body away from the wall to get a better view of what's just below you), and when descending it is harder to test potential handholds and foot placements at the outset of a

[1] In Chapter 4, we saw that one of the main objections rationalism's critics raise against evolutionary psychology is that we can't know anything at all about the environment of evolutionary adaptiveness (the EEA). But there is absolutely no question that our ancestors faced the risk of falling. There are many fairly basic features of the EEA like this, features which aren't in doubt and that offer interesting materials for making predictions about rationalist learning mechanisms.

venture to get a feel for the challenge ahead (unlike ascending, where the risk associated with an initial test is falling a few inches, not falling the entire vertical length). These additional facets of the risks involved in selecting a route led Jackson and Cormack (2007) to predict a specific cognitive quirk—that people should overestimate distance for descending routes more than ascending routes. They call this phenomenon the *descent illusion*.

To investigate the descent illusion, Jackson and Cormack examined people's estimate of the very same vertical distance when standing on the top of a ridge looking down and when standing on the ground looking up. (Participants indicated their estimate in each case by specifying what they took to be the equivalent horizontal distance along a flat surface.) Participants overestimated the vertical distance in both situations, but the overestimation was considerably larger when estimating from the top of the ridge than the bottom. When viewed from the top, people in this experiment overestimated distance by a massive 84% relative to the true distance. To get a sense for how powerful this illusion is, note that the more familiar Müller-Lyer illusion results in seeing the same line segment as having approximately a 10% difference in length depending on whether the line is capped by regular arrow heads or inverted arrow heads.[2] In fact, the descent illusion is the largest known real-world distance illusion (Jackson and de García 2017). Yet incredibly, the extent to which people overestimate vertical distances and the fact that they do so considerably more for downward routes than for upward routes were only discovered very recently when these researchers began to explore Evolved Navigation Theory in earnest.[3]

Here is another relatively recently discovered illusion associated with Evolved Navigation Theory. If asked whether the distance extending towards a cliff edge looks to be of a different length than the same distance extending away from a cliff edge, people often suppose that it should look shorter extending towards the cliff edge. (Possibly they think that actually seeing the cliff edge should make more vivid the need to keep one's distance from the edge). However, Evolved Navigation Theory predicts exactly the opposite, since moving towards a cliff imposes the greater risk, which, according to the theory, is factored into perception by increasing perceived distance. Once again, Evolved Navigation Theory makes the correct prediction about a quirky feature of distance estimation. Jackson and Willey (2013) tested people's distance estimates standing at the edge of a steep slope with the slope to their back (as if they were going to walk away from the slope), compared to facing the slope (as if they were going to approach it).

[2] In the Müller-Lyer illusion, the central line segment in figure (a) appears longer than the central line segment in figure (b) even though the line segments are identical in size:

(a) >——————<
(b) <——————>

[3] See also Jackson and de García (2017) for cross-cultural evidence from a small-scale society in support of the illusion.

The very same distance looked longer when participants estimated the distance towards the slope (the risker route). This quirky and unexpected phenomenon is an effect known as the *plateau illusion*.

These examples involve potentially life-threatening dangers, like plunging over a cliff. However, the overestimation of distance has been found to be graded according to the level of risk and to take place even with surprisingly modest levels of risk. This is seen in a study that tested people's estimates of horizontal routes, comparing a route over and across a ditch (high falling risk), one alongside the ditch (moderate falling risk), one on a nearby curb (low falling risk), and one on nearby solid ground (very minimal falling risk). Interestingly, participants overestimated the distance in all but the last condition, with overestimates that were proportional with the degree of risk even though the true distance was identical across these conditions. These results were again highly unexpected, particularly in the case of the curb, which involved a drop of only 6 inches but nonetheless generated an overestimation (9%) comparable in size to that involved in the Müller-Lyer illusion (Jackson and Willey 2011).

We will mention one last prediction made by Evolved Navigation Theory to round out our discussion of this example. This has to do with the way that vertical retinal images contribute to distance perception. Sometimes a vertical retinal image corresponds to a vertical surface in the environment, but sometimes it corresponds to a flat and perfectly safe horizontal surface (as when you are standing in the middle of a straight paved road with your feet on the centreline and are oriented so that the centreline traces a path ahead of you from your feet to the horizon). Evolved Navigation Theory predicts that the exaggeration of distance should only occur for environmentally vertical surfaces since only these pose a risk for falling. A contrasting hypothesis holds that the exaggeration of distance is a relatively low-level visual phenomena (in particular, that it is an artefact of the retinal image, where a vertical line on the retinal image corresponds to a greater distance than the distance represented by the same-sized horizontal line).

To test for whether the exaggeration of distance is truly sensitive to the representation of environmentally vertical surfaces, Jackson and Cormack (2008) had participants estimate distances in which their position always generated a similar vertical retinal image, half the time because of an environmentally vertical surface and half the time because of the projection of an environmentally horizontal surface (see Figure 15.1). The result, as predicted, was that participants overestimated the distance only when it corresponded to an environmentally vertical surface, with greater overestimation for greater vertical heights (which introduce greater risks). Jackson and Cormack call this effect the *environmental vertical illusion*.

Notice that the pattern of errors that infuse people's distance estimates isn't one that people are aware of (indeed, no one knew about it before this work was done). So it is extremely unlikely that the pattern is acquired in childhood from

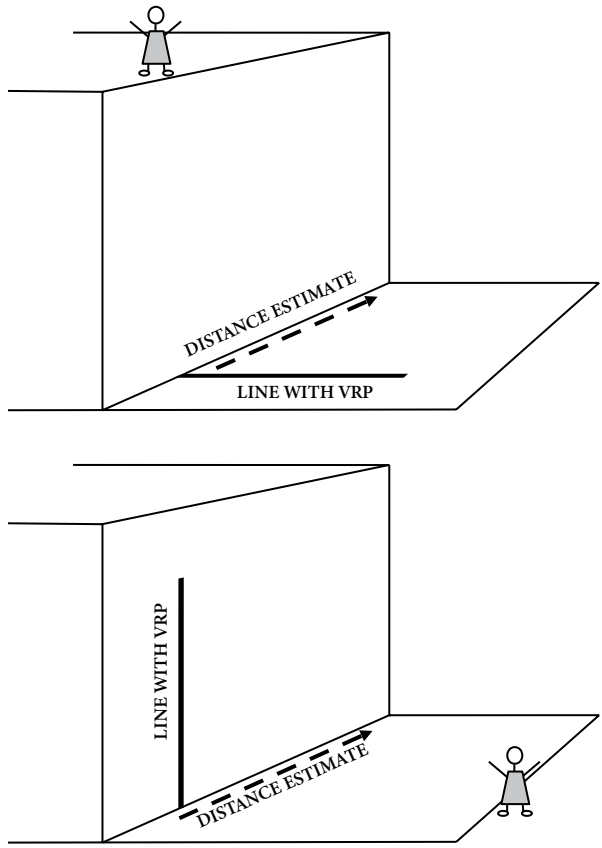

Figure 15.1 The environmental vertical illusion. The observer looks at the line, which has a vertical retinal projection (VRP), and estimates its distance by referring to a corresponding distance on the ground that runs perpendicular to the line. The viewer in the top image, who is looking down, estimates the length of the distance extending in front of her along a no-risk flat horizontal surface on the ground below. The viewer in the bottom image, who is looking up, estimates the length of a distance along a high-risk vertical surface extending upwards in front of her. (Figure based on Jackson and Cormack 2008, figure 1.)

being taught how to represent riskier routes as longer. It is also unlikely that the pattern is acquired by an individual's general experience of distance, as what would have to be learned is radically at odds with the true facts about distance. Riskier routes, after all, are *not* in fact longer; what makes these illusions is that they are being systematically *mis*represented as longer.

An empiricist could try to argue that experience of effort in navigation explains how these erroneous distance judgements come to be made. For example, the extra effort in ascending a slope could, in principle, lead to viewing ascending and descending routes rather differently. But this proposal makes the wrong

prediction about perceived distance. If the more difficult route is perceived to be longer (so route preferences continue to follow the simple rule "choose the shortest route"), then effort-based theories should maintain that a vertical slope is perceived to be shorter when descending. Yet the central finding in these investigations is that the exaggeration of distance is *greater* for descending routes. (Again, this is predicted by Evolved Navigation Theory because of the increased risk of falling when descending.) Nor could this empiricist hypothesis explain why routes along flat ground are estimated as longer when they approach or run alongside vertical drops.

In contrast, these quirky phenomena all make sense in light of the innate mechanism for distance estimation posited by Evolved Navigation Theory. This mechanism provides a unified and principled explanation for a diverse set of otherwise highly puzzling findings. In fact, the rationalist hypothesis that this mechanism subserves route selection *predicts* that we ought to find such quirky details concerning how distance (and risk) is experienced. This is because the hypothesis, which is motivated by thinking about the adaptive problems associated with navigation, is that distance estimation in navigation isn't just about distance—it incorporates an assessment of a route's level of risk. The resulting representation of distance effectively treats distance representations as summary representations of both distance and risk, allowing distance estimates to provide a simple way to select optimal routes that take into account both of these factors.

It's also important to recognize that these visual illusions are strongly tied to cues of risk that may be fairly abstract. When there are strong contravening cues—cues that there is no risk—a more accurate representation of distance is free to form. In an elegant experiment that shows how risk cues are integrated into distance estimates, Jackson and Cormack (2010) essentially repeated the setup in which they previously documented the environmental vertical illusion, but this time by using a virtual reality environment rather than real-world conditions, and by deliberately decoupling participants' view of the environment from their body posture (so that a head turn no longer had the normal consequence of changing the view concordant with moving in the environment). In this situation, people lose their sense of *presence* in the virtual world and don't experience any feeling of risk, though the strictly visual conditions for experiencing the illusion are still present. The result is that the environmental vertical illusion completely disappears. Remove the perceived risk and this removes the overestimation of distance for a vertical surface—for an illusion that otherwise distorts real distance by as much as 50%.

What's more, although the illusions we have been talking about are visual illusions, the mechanisms involved have the function of delivering their input to systems for navigational decision making and interact with the presumably

innate (defeasible) preference for shorter routes. Thus we have good reason to suppose that the mechanisms that create these quirky illusions are articulated in the sense of Chapter 2 in that they participate in a larger arrangement of rationalist learning mechanisms that turn on a way of representing the world that is geared to ancestral navigational priorities.[4]

How does this example relate to concept nativism? The concepts that it speaks to are the distance representations involved in navigational judgements and decision making. These categorize possible routes in ways that are then used to determine if a route is viable, to compare competing alternatives, and for planning which route (if any) to take on a given occasion. As with a few of our earlier examples, there is some room to debate whether such representations are truly conceptual or whether they should be considered nonconceptual. For example, theorists who require concepts to meet a strong form of the Generality Constraint may wish to claim that these distance representations aren't concepts on the grounds that they don't combine with other concepts that have nothing to do with navigation. But on other accounts of the conceptual/nonconceptual distinction, they straightforwardly count as concepts—for example, because they aren't iconic, engage with higher cognitive processes like decision making, and are clearly quite abstract in that they factor risk (and perhaps other dimensions that matter) into route selection. For our purposes, there is no need to decide between these options. For those who are happy to view these distance representations as concepts, then this is another domain that can be added to the case for concept nativism. For those who aren't, then this can be viewed primarily as arguing for there being further characteristically rationalist psychological structures in the acquisition base—structures which contribute to the origins of other representations that are uncontroversially conceptual, for example, OPTIMAL ROUTE.

Let's now turn to another example illustrating the argument from cognitive and behavioural quirks, which concerns aspects of social categorization. We will begin with a brief discussion of some relevant background regarding the representation of social categories. In social psychology, it has long been thought that an initial stage in impression formation is the categorization of a perceived person with reference to a small number of highly salient social categories: *age*, *sex*,

[4] Other research has uncovered cognitive quirks related to the distance representation associated with looming versus receding sounds and has found that looming sounds are uniquely heard as being closer than they really are, yet only when plausibly associated with moving objects (Neuhoff 1998). Both humans and non-human primates are subject to this auditory illusion (Ghazanfar et al. 2002). Moreover, the effect is more pronounced for listeners with lower fitness levels (Neuhoff et al. 2012). Once again, this is a perplexing set of circumstances until an adaptive problem is given its full due, namely, the problem of responding in a timely manner to a looming threat, particularly for those who, ancestrally, may have required a larger margin of safety and thus would benefit from more of a head start. There is now also complementary evidence showing that newborn human infants integrate auditory and visual cues for looming events (Orioli et al. 2018). So looming provides a further, related example embodying several of our seven argument forms.

and *race* (Fiske 1998; Fiske et al. 2018). These three have been thought to be privileged compared to other social categories (*occupation*, *religion*, etc.) in being automatically activated, rapidly processed, and difficult to suppress. They are also generally thought to possess this privileged status because they correlate with properties that are easy to perceive, particularly via visual features, and because they have a great deal of cultural meaning.

Standard accounts in social psychology don't offer a deeper explanation of why these three social categories stand out among others. However, from an evolutionary perspective, it makes sense that two of these are on the privileged list: age (i.e., *life stage*) and sex (i.e., the sexual morphs *male* and *female*). Throughout human evolution, the age and sex of encountered individuals would have had significant consequences for a range of actions with an impact on fitness. Since enduring conditions like these provide the materials for natural selection to fashion an adaptation, it shouldn't be all that surprising if there turned out to be innate specialized systems for categorizing in terms of age and sex, and accompanying downstream psychological systems for organizing inferences, motives, and actions that are specific to these categories. Perhaps, then, there is a deeper account to be had of why age and sex should be singled out in person perception. They may be products of innate systems that are designed to categorize people along a small number of dimensions with a history of fitness relevance.

Suppose for a moment that this sort of evolutionary account is on the right track. Still, it raises a puzzle in that, while it may account for the privileged status of the representation of age and sex, it cannot account for the privileged status of the representation of race. The problem is that ancestral hunter-gatherers wouldn't normally have travelled sufficient distances in their lifetimes to encounter individuals with the type of systematically different superficial physical characteristics associated with being considered to be of a different race (Diamond 2013).[5] Thus it is unlikely that there would have been selection pressure to develop systems for representing race. Why, then, would the representation of race be as automatic and irrepressible as the representation of age and sex?

[5] As we have noted, concepts and conceptual domains do not entail the reality of the categories they are about. So the fact that people think and categorize in terms of races does not mean that races actually exist. There are a number of competing philosophical views about the nature of race and whether races exist (Mallon 2016; Glasgow et al. 2019). For example, some philosophers, who are eliminativists about race in that they hold that races do not exist, take races to be biological categories grounded in heritable biological properties that all and only members of a given race possess. Another philosophical view about races takes races to exist, and to be socially constructed categories that factor importantly in understanding conceptions of racial identity and in social and political movements for racial justice. The work that we discuss here (which is concerned with racial concepts and categorization and *not* with the questions of what races are or whether they exist) is compatible with these and other views on the nature of race. We will use the term "race" (and its cognates) to refer to categories employed in racial categorization, with no implication that there is any deeper reality to these categories.

It was against this background that Kurzban et al. (2001) suggested that although racial categorization looks like it is on a par with the categorization of age and sex, this is misleading. Unlike cognition regarding age and sex, our minds do not contain an innate mechanism for tracking race. The reason that it appears that we are designed to track race is that race is acting as a proxy for something else that we *are* designed to track, namely, coalitions or alliances.[6]

Coalitions are small groups of people who work together for a common aim, often in competition with other coalitions. Some coalitions are relatively stable, others highly changeable. Given how important coalitions are in all human societies—and presumably in human evolution—it is reasonable to suppose that there would have been selection pressure for the development of psychological systems dedicated to navigating these social relationships. This includes systems for detecting incidents of cooperation and conflict, for assessing the strengths and weaknesses of current and potential coalitions, and for motivating involvement in profitable coalitions. And because any individual will simultaneously be a member of multiple cross-cutting coalitions, and because the coalitions that an individual belongs to can change over time, a well-designed coalition psychology should also include systems for flexibly identifying context-sensitive cues that predict the coalitions (or types of coalitions) that are relevant to the current moment. Perhaps, then, it is because we have an innate *coalitional psychology*, and because we happen to treat race as a probabilistic cue for coalition, that it appears as if we are designed to track race.

From this perspective, the noteworthy fact about the standard view in social psychology that racial categorization is a privileged category—on a par with age and sex—is that tests for this claim didn't manipulate conditions that disentangle race from coalition. As a result, race could have been activated as a default coalitional cue, one that derives from living in a society with a history of racial divisions. This would predict that if race were pitted against coalition—even for fleeting and unimportant coalitions—that we should track coalition rather than race, with racial cognition being strongly diminished. At the same time, it would predict that if age or sex was pitted against coalition, this would have little or no effect on our ability to track age or sex, since, according to this view, age and sex—unlike race—*are* fundamental categories of person perception and part of our innate psychology and so shouldn't be affected by the concurrent tracking of coalitions.

To test for this possibility, researchers have used a *memory confusion task* (Kurzban et al. 2001; Pietraszewski et al. 2014; Pietraszewski 2021). This research method involves presenting participants with a set of pictures of people and corresponding statements or contributions to a conversation made by these people,

[6] The terms *coalition* and *alliance* are used interchangeably in this literature, where the proposal that race functions as a proxy for coalition(/alliance) is sometimes referred to as the *alliance hypothesis of racial categorization* (see, e.g., Pietraszewski 2022).

and later asking them to recall who said what (for this reason the memory confusion paradigm is also sometimes known as the *who-said-what paradigm*). However, at the outset of the experiment, the participants are only instructed to form an impression of the people they will see; they aren't told that they will later be asked to remember what different people said. Not surprisingly, this is a difficult task—participants make many errors. But this is actually the point of the experiment, as the participants' errors reveal how they were representing the observed individuals. For example, suppose speakers are wearing different coloured shirts in the photographs, some wearing red shirts, the others wearing blue. If participants then systematically misattribute statements made by speakers wearing one colour of shirt to other speakers wearing the same colour shirt, this would show that they were initially categorized according to shirt colour. One of the reasons why age, sex, and race have been considered so basic to social categorization is that people's recall in memory confusion experiments has tended to cleave to these three categories. Misattributions of what someone said tend to be *within* these categories rather than *across* them. For example, people are more likely to misattribute statements by one white man to another white man or statements by one young person to another young person than they are likely to misattribute what a white man said to a black man or what a teenager said to a senior citizen.

The key innovation that was needed to apply this method to the coalitional hypothesis was to add coalitional cues that cross-classified with race (Kurzban et al. 2001). For example, in one set of experiments, participants read statements from individuals from two competing sports teams (a type of coalition), each of which was composed of both black and white players or both male and female players (Pietraszewski 2021). A baseline measure determined the extent to which an individual was confused with another of the same race or the same sex when there wasn't any coalitional information, and this was compared to the situation when coalitional cues of different strengths were given. The result was that the representation of race was greatly diminished in the presence of coalitional cues in that there was less of a tendency to confuse speakers within a racial category than across a racial categories. In contrast, the representation of sex was not reduced nearly as much and, in some cases, not at all—there continued to be just as strong a tendency to confuse male speakers with one another and female speakers with one another as in the baseline condition. What's more, further work has shown that the drop in racial categorization happens specifically when race is cross-classified with coalition. It remains robust when cross-classified with another social category, just like categorization in terms of sex (Pietraszewski 2022).

This overall pattern shows that the representation of race and the representation of sex are grounded in distinct systems for social categorization. Categorization by sex continues to be automatic and mandatory when cross-classified with coalition, consistent with the hypothesis that natural selection has left us with an innate domain-specific mechanism for the rapid identification of

an individual's sex.[7] In contrast, categorization by race is not automatic and mandatory, diminishing dramatically when cross-classified with coalition, suggesting that race is being treated as a proxy for coalition and that it is coalition, not race, that is the privileged form of categorization on a par with sex and age.[8]

The crucial fact from the point of view of the argument from cognitive and behavioural quirks is that, among the major social categories that have been thought to be privileged in person perception, racial categorization alone can be dramatically reduced—if not completely eliminated—when cross-classified with coalition. Moreover, this is true even when the coalitions are relatively insignificant ones—unknown sports teams that are implicitly introduced in the course of a brief experiment.

This quirky fact about racial categorization is highly unexpected on a domain-general account. Why would a domain-general learner acquire racial concepts that are so fragile when orthogonal coalition information is introduced, but other social category concepts (sex, age) that remain robust under comparable conditions? Standard domain-general accounts of so-called fundamental social categories hold that they are learned by observing perceptual cues that have social significance in one's community. If this were right, then, at least in the United States (where most of these experiments took place), racial categories *should* have turned out to be fundamental categories of person perception, since they are certainly associated with perceptual cues (especially skin colour) with enormous social significance. But as we have seen, racial categorization is surprisingly fragile and easily reduced with a simple experimental manipulation. Even for fleeting and insignificant coalitions, fairly minimal indications of coalitional membership that cross-classifies with race have been shown to produce dramatic drops in racial categorization.[9] This is a striking result that is deeply puzzling given a domain-general learning framework. But it begins to make sense given a rationalist framework. On the assumption that there is an innate mechanism for identifying and reasoning about coalitions and that race (in the right cultural context) is treated as a proxy for coalition, we should expect to find just the pattern of results that we have described. Racial categorization should appear to be robust—until it is pitted against coalition. Under these conditions, the categorization of

[7] Kinship, another highly important social category, appears to behave much like sex (and age) in a memory confusion experiment of this type, suggesting that it is also a privileged category of person perception and grounded in an innate domain-specific mechanism (Lieberman et al. 2008).

[8] Further support for this view comes from an intriguing variation on these experiments in which race was cross-classified with another type of coalition membership, namely, political party affiliation. In this study, race, sex, and age were all cross-classified with political party affiliation. The result was that categorization by race was greatly diminished but categorization by sex and age was not (Pietraszewski et al. 2015).

[9] We should note that the base-rate correction used in research in the memory confusion literature prior to 2018 (which was adopted from earlier work in social psychology) has been shown to be problematic (Bor 2018). However, Pietraszewski (2018) reports that a reanalysis using a new base-rate correction that avoids this problem actually *strengthens* the results in the work on race and coalition.

coalition should take precedent, and any effects in terms of racial categorization should be dramatically reduced.

In the cases of route selection and coalitional thinking, the cognitive and behavioural quirks at issue are ones that are present in adults and that are products of a fully developed cognitive capacity. For our last illustration of this argument, we will turn to an example where the quirks manifest themselves earlier in life and while a cognitive capacity is still developing. The case that we will consider concerns the development of physical reasoning in infancy.[10]

One of the core features of physical reasoning is an understanding that objects continue to exist when we can't perceive them, also known as *object permanence* (see Chapter 10). Evidence of an understanding of object permanence appears very early in development, in infants as young as 2.5 months old. For example, infants this age are surprised if a cover is placed over an object and, when the cover is removed, the object has disappeared (Wang et al. 2005).[11] It is important to keep in mind, however, that an understanding of object permanence is just one part of a much broader physical-reasoning ability. Among other things, infants must at some point come to understand when and to what extent one object will block another object, when and in what ways one object will be supported by another, which properties of objects can change over time and which cannot, and when one object will fit behind, inside, or under another object and no longer be visible. Here we will focus on the last set of these facets of physical reasoning—the conditions in which an object will no longer be visible when moved behind, into, or under another object.

By 2.5–3 months of age, infants expect an object placed behind another object to reappear when the second object (the occluder) is removed (Hespos and Baillargeon 2001a).[12] But suppose that infants are shown a tall object and a much shorter potential occluder. It turns out that infants don't initially expect that the tall object will remain visible when placed behind such an occluder. It is not until 3.5 months of age that infants are surprised if a tall object becomes completely hidden behind a much shorter occluder (i.e., by means of an experimental trick that the infants are not aware of) (Baillargeon and DeVos 1991). Before this age,

[10] We briefly touched on physical object representation previously in relation to the argument from animals (Chapter 10). We will also return to this issue in Part III when considering methodological empiricism as an objection to concept nativism (Chapter 17).

[11] This and related work suggests an argument from early development for a rationalist account of the origins of object permanence.

[12] As in Chapter 10, we are following standard psychological usage in which an occluder is an object that blocks the view of another object. However, as will become clear in a moment, we now need to distinguish between, on the one hand, an object blocking the view of another object because it is acting as an occluder with the second object *behind* it and, on the other hand, a case in which the object blocks the view of another object because the second object is *inside* of it (it is acting as a container) or *under* it (it is acting as a cover). In this chapter, "occlusion" will refer to only the sort of case in which one object is behind another, and "occluder" will refer to the object that the occluded object is behind.

they represent occlusion but fail to take into account the relative heights of the object and the occluder in determining whether the object should remain visible behind the occluder. This is somewhat peculiar, but of course, if infants' understanding of physical objects develops over time, there ought to be some notable differences between their expectations about physical events and normal adult expectations. However, work with slightly older infants shows that this is far from the end of the story.

Quite unexpectedly, once infants do begin to register surprise when a tall object becomes completely hidden behind a much shorter occluder, they are still not surprised when a similarly tall object appears to become completely hidden when placed *inside* a short container (i.e., a container that is just as short as the short occluder) (Hespos and Baillargeon 2001b). In fact, infants that are surprised in the occlusion case do not show surprise in the containment case even when they are tested with occluders and containers that are virtually indistinguishable from one another. This has been shown using cylindrical containers (with bottoms but no tops, like an open can) and occluders that were shaped and coloured exactly like the front half of these containers (without a bottom) (see Figure 15.2). Since only the fronts of the containers and occluders were visible in the test conditions (the backs and bottoms weren't visible), the container and occluder test events would have looked the same from the infants' perspective.[13] Remarkably, however, infants were surprised when the much taller object disappeared behind the short occluder, but *not* when it disappeared inside of the equally short container.[14]

Further experiments—testing infants at monthly age intervals from 3.5 months old on—have shown that infants aren't surprised when a tall object disappears inside a much shorter container until they are 7.5 months old—a full four months after they are surprised to see a comparable case of occlusion (Hespos and Baillargeon 2001b). And infants treat covers differently from both occluders and containers. It isn't until they are 12 months old that they are surprised when a tall object disappears beneath a much shorter cover—more than *four months* after

[13] Prior to the test trials, the containers and occluders were rotated so that the infants could see that the container had a back and bottom (in the container condition) or that the occluder did not have a back or bottom (in the occluder condition). But during the test trials, the container and occluder were placed upright and stationary, so that they looked nearly identical and could only be conceptualized differently—*as a container* and *as an occluder*—by remembering the demonstrations that preceded the test trials. Infants' looking times in the short container and short occluder test conditions were measured relative to their looking times in tall container and tall occluder test conditions, respectively. These tall container/occluder conditions were the same as the short container/occluder conditions, except that the container/occluder used was tall enough to completely hide the object and so infants should not have been surprised by the fact that the tall object was completely hidden in these cases.

[14] This pattern of results held even when the *very same object* was used both as the occluder and as the container. Infants *were* surprised when the tall object disappeared when placed *behind* a short container, but not when placed *inside* this same short container.

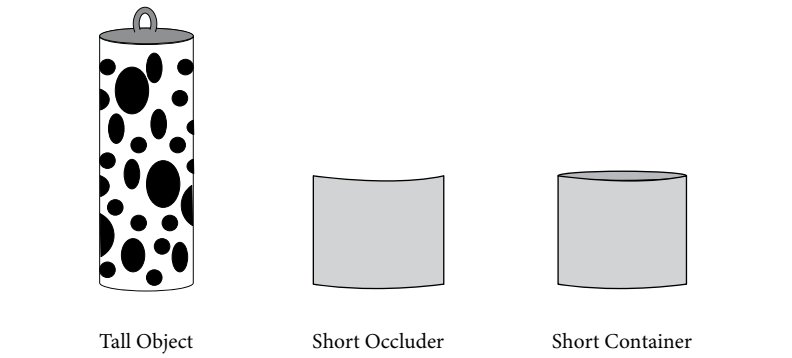

Figure 15.2 Hespos and Baillargeon's (2001b) study. 4.5-month-old infants are surprised when a tall object completely disappears behind a short occluder, but are not surprised when the same tall object completely disappears inside a short container that is perceptually indistinguishable from the short occluder in the test condition. (Figure based on Hespos and Baillargeon 2001b, figure 1.)

they show surprise in the case of containment and over *eight months* after they show surprise in the case of occlusion (Wang et al. 2005). This is true even when the cover is perceptually indistinguishable from the container used in these other experiments (i.e., when the cover is simply the previously used container turned upside down).[15]

This work reveals a startlingly quirky pattern of expectations in infants that is quite puzzling on the assumption that this developmental progression is a response to domain-general learning. Once relative height is recognized as relevant to whether an object remains visible when interacting with another object, it is hard to see why a domain-general learning system should care whether the second object is an occluder, a container, or a cover—or why it should take *so long* to apply what it has learned in one case to these other cases. Moreover, if it isn't assumed that the domain-general learning system already has the concept of an object and the idea that objects continue to exist when they aren't visible, then it is hard to see how to make any sense of these results at all. If infants saw the world solely in terms of something like shaped, coloured, and patterned surfaces, there would be no way for them to represent the difference between the containment event and the occlusion event. Both would involve exactly the same shaped surface gradually disappearing as it moves to the point at which it meets a surface with the shape of the front of the occluder/container (this second shape being identical for the occluder and the container).

Interestingly, there is also a further layer of quirkiness associated with this example. This is that although these lags are a normal part of development, it is possible to completely eliminate them if infants are shown just a few events with the right type of structure. How this works and why it is so effective illuminates how cognitive development plays out in this domain, showing it to be, in good part, a learning process in which infants seek explanations of experiences that conflict with their understanding of a given type of physical event. Moreover, as we will see, this learning process has the hallmarks of *rationalist* learning, offering strong evidence that infants have access to an innate domain-specific physical-reasoning system.

Let's look at how this learning process works. First, to review, so far we have seen that infants as young as 2.5 months old have an understanding of object permanence in that they expect an object to reappear after it is temporarily concealed by another object. And we have also seen that infants have a basic understanding of different types of interactions between objects, such as occlusion and containment. Finally, we have seen that there are significant developmental lags

[15] Rationalists are often accused of ignoring cognitive and conceptual *development* (see Chapters 4 and 17). But this type of painstaking rationalist developmental research, which tests infants as they age month by month to uncover subtle developmental patterns, shows just how wrong this criticism is.

as to when infants understand that specific features of objects are relevant to whether an object will fully disappear in these different types of events. For example, once they begin to take into account the height of objects and occluders for occlusion events, there is a substantial lag before they also take height into account for containment events, and then a further substantial lag before they take it into account for covering events.

While these sorts of developmental lags are a robust feature of typical development, Wang and Baillargeon (2008) showed that infants can be "taught" to take height into account in understanding covering events substantially earlier than the normal developmental progression—accelerating this aspect of development by a full three months (from 12 months old to 9 months old). Moreover, this acceleration only requires exposure to a few carefully orchestrated events.

In Wang and Baillargeon's procedure, infants were initially shown a tall and a short cover (which were otherwise identical) resting next to a tall object, making their heights relative to the tall object easy to register. They were then shown the inside of each cover prior to seeing it placed over the object. When the tall cover was placed over the object, it completely hid the object, since this cover was slightly taller than the object. When the short cover was placed over the object, it failed to completely hide the object, since this cover was much shorter than the object. The covers were then placed beside the object once more, so infants could again observe their heights relative to the object's height and to one another. The infants then saw all of this one more time with the same object but with what was clearly a new pair of covers—again, one that was tall and one that was short. Finally, they were tested with a new tall object and new tall and short covers. They were shown each cover placed over the object, and in each case, the object became completely hidden.

Though infants are not normally surprised when a short cover completely hides a taller object until they are a full 12 months old, with just these two teaching events, 9-month-olds *were* surprised at this outcome—they looked significantly longer when the tall object disappeared under the short cover.[16] What they seem to learn, however, is specific to this one type of event (covering events). This can be seen in further work using the same teaching method. Nine-month-olds who rapidly learned to appreciate the significance of height to covering events didn't generalize this knowledge to events in which an object enters a tube. They continued to show no surprise when a tall object became completely hidden in a tube that was substantially shorter than the object (and that was visually identical to the tested covers apart from having no top or bottom) (Wang and Kohne 2007).

[16] Moreover, the effect wasn't transient. The same result was obtained even when the teaching events were presented on one day and the test events were presented a full twenty-four hours later.

What accounts for the success of this type of intervention in allowing infants to learn about the role of relative height in covering events? Wang and Baillargeon (2008) identified three crucial features: (1) contrasting outcomes (teaching events in which the tall object is fully covered by the tall cover but not by the short cover), (2) a clear correlation between the relevant object property (relative height) and the contrasting outcomes, and (3) the availability of a causal explanation for the contrasting outcomes. Without all three of these features, infants fail to learn to take height into account in processing covering events, and so do not expect that a tall object will fail to be completely covered by a short cover.

To show that contrastive outcomes are essential, the experiment was repeated as in the original teaching experiment except that the object that was used was slightly shorter than the short cover, so that both covers were big enough to completely cover the object. As before, infants were given exposure to comparative height information—the covers were initially positioned beside the object, were then placed over the object, and subsequently repositioned beside the object. But without the contrasting outcomes, no learning about the significance of relative height for covering events was triggered.

To show that there must be a clear correlation between the relevant object property and the contrasting outcomes, another small variation on the original experiment was employed in which the covers were never placed directly beside the object (but everything else about the experiment was the same). In this variant, instead of being placed directly beside the object, the covers were held in the air above or off to the side of the object, making the differences in relative heights less salient and harder to detect. Here, too, teaching failed.

Finally, to show that the availability of a causal explanation for the contrasting outcomes is essential, the original experiment was repeated again with a different small variation. This time when the infants were shown the insides of the covers prior to the covers being placed over the object in the teaching events, they could see that the covers' insides were not as deep as they appeared to be from the outside. In particular, the tall cover's inside was shown to be quite shallow, with an inside depth even shorter than the height of the short cover. This change undermines the causal explanation for why tall covers, but not short covers, should be able to completely hide a tall object. And, when it came to the test events, the teaching again failed. The infants didn't look any longer when the tall object became completely hidden under the short cover than when it became completely hidden under the tall cover.

There are three important morals to draw from this work on how and when infants come to understand that an object will become completely hidden behind, inside, or under another object—both the work that has identified the striking developmental lags in children's ability to predict what will happen with these different types of events and the work that has showed how to accelerate infants' learning in relation to them.

First and foremost, the specificity of the processes involved in eliminating such developmental lags effectively rules out an explanation in terms of domain-general learning. The learning is tied to a particular event type (e.g., specifically to *covering events* in a way that doesn't generalize to highly similar looking *tube-containment events*). In this way, inferences about the relevance of height (and similar variables) is *category bound*, just as the developmental lags are.[17] Recall that with the lags, even once infants determine that height matters to one type of physical event (e.g., occlusion), they *don't* infer that is relevant to another (e.g., containment). Their physical knowledge develops in piecemeal fashion, with the knowledge pertaining to one type of physical event not spilling over to knowledge of other types of physical events, even though these types of events are so similar from an adult point of view.

The second moral is that while domain-general learning accounts are not in a position to explain these lags and the ways that they can be eliminated, rationalist learning accounts are. Here we will just sketch the basic outlines of a highly promising rationalist account.[18] Consider again the fact that 9-month-olds are not surprised when they see a tall object become completely hidden when a much shorter cover is placed over it, but that they are surprised when the same tall object becomes completely hidden when placed inside of a much shorter container. In order for this to be possible, infants have to be capable of representing the relative heights of the tall object and the short container. If they weren't capable of doing this, then height shouldn't make a difference to them when interpreting the outcome of a containment event. So, one part of the infant's mind—one of several systems pertaining to the representation of objects—must represent the heights of the objects (following Lin et al. (2022), we will refer to this system as the *object-file system*). At the same, however, this same information about height is systematically ignored when the infants are confronted with the covering event. As Reneé Baillargeon and her colleagues have argued, the best explanation for this fact is that, in addition to the object-file system, the mind also possesses another distinct system—a *physical-reasoning system*—which is responsible for making sense of interactions among physical objects (Baillargeon et al. 2011; Stavans et al. 2019; Lin et al. 2022). The physical-reasoning system represents physical events in a schematic manner, only including details that it takes to be relevant to the type of event in question, where the properties it deems relevant can be modified over time through learning.[19] By hypothesis, at 9 months of age,

[17] See Strickland and Scholl (2015) for evidence of category bound effects of containment and occlusion in adult object representation.

[18] For more on the learning model here and the innate domain-specific systems involved, see Baillargeon et al. (2011); Baillargeon et al. (2012); Baillargeon and DeJong (2017); Stavans et al. (2019).

[19] Why should the infants' physical-reasoning system take into account only a small subset of the properties that they are capable of detecting? It might seem more sensible for the physical-reasoning system to take into account all of the properties that infants can detect in the scene. The problem is

this system has already learned to take height into account when reasoning about containment events, but it hasn't learned to use height when reasoning about covering events. To a 9-month-old's physical-reasoning system, height has nothing to do with covering, so there is nothing surprising about a tall object completely disappearing under a much shorter cover.

What happens, then, when the developmental lag vis-à-vis covering is eliminated? The natural explanation is that the minimal teaching events help the infants to figure out that a new variable needs to be included in their reasoning about this type of event. In other words, when 9-month-olds learn that a tall object cannot be completely hidden under a much shorter cover three months ahead of schedule—and when typically developing infants learn the same thing by 12 months of age—their physical-reasoning system has become modified through learning to include information about height when drawing inferences about covering events. The important point here is that the learning model involves multiple distinct systems relevant to understanding events involving objects, each with its own structure, processes, and representational resources.[20] Since these domain-specific systems are not plausibly learned, they must be assumed to be innate. Thus, the best and most plausible account of these developmental quirks is a rationalist account.

The third moral to draw from this work is that the nature of the learning process underwriting the development in infants' physical reasoning argues that the representations involved—OBJECT, CONTAINMENT, OCCLUSION, etc.—are concepts, rather than nonconceptual representations. We saw in Chapter 6 that there are a variety of different ways of drawing the conceptual/nonconceptual distinction, with no consensus on how it should be drawn. But most theorists would

that taking into account additional variables increases processing demands. So, it would be computationally burdensome, especially for infants with highly limited processing resources, to take into account all the detectable properties of objects in physical reasoning.

[20] There is a question about how many distinct systems are needed in this sort of multiple system account. Lin et al. (2022) model this data in terms of two distinct systems—the *object-file system* (which identifies the objects in a scene and builds a representation of both their features and their spatiotemporal properties—what Lin et al. refer to as "what" and "where" information) and the *physical-reasoning system* (which categorizes an event in terms of its causal structure—for instance, taking an event to be a covering event—and then draws appropriate inferences). But much the same data might be modelled in terms of three distinct systems instead, taking the functions associated with the object-file system to be mediated by two distinct systems, one that encodes the features of objects and explains the ability to categorize and recognize individual objects as such (the *object-representation system*) and one that is devoted simply to tracking a small number of objects on the basis of their spatiotemporal properties (the *object-tracking system*) (Baillargeon et al. 2011; Baillargeon et al. 2012). The object-tracking system is also sometimes referred to as the *object indexing system* (Leslie et al. 1998). It's worth noting that infant object representation is sometimes simply equated with object tracking, and as a result, it is assumed to be a relatively simple capacity that is subserved by an attentional mechanism that employs only nonconceptual representations. As should be clear from the text, this way of thinking about infant object representation is mistaken. Object tracking is only one part of a set of interrelated capacities regarding infants' representation of objects and, on its own, cannot explain infants' ability to engage in causal-explanatory reasoning as they analyse object interactions and learn about their physically relevant properties.

agree that an inferential process that is sensitive to causal-explanatory reasons operates at the conceptual level and that the representations it employs are concepts. Now consider again the learning process involved in this development, as illuminated by the teaching experiments that showed that the developmental lags can be eliminated. This learning process has all the hallmarks of a causal-explanatory reasoning process. It is responsive to subtle forms of evidence as infants attempt to make sense of experiences with disparate outcomes that their current rules for understanding physical events treat in the same way. Typically developing 9-month-olds do not automatically take height into account for covering events; they are not surprised if a tall object completely disappears under a short cover. But if height differences are made salient, and correlated with contrasting outcomes, and if there is a causal explanation available that can make sense of the contrasting outcomes, then they *can* learn to take height into account in these types of events. However, they won't learn this if there is no causal explanation available that could explain the contrasting outcomes they observe.[21]

We have seen that the development of physical reasoning in infancy is quirky in at least two ways. First, there are robust and highly surprising lags in infants coming to see that the same variable (e.g., height) applies across different types of physical events (occlusion, containment, covering, etc.) even when these events are perceptually indistinguishable. Second, despite how robust these lags are, they can be readily eliminated (for a given category of events) with just a few teaching events of the right sort. Both of these striking findings are unexpected on the assumption that infants rely exclusively on domain-general learning processes to acquire an understanding of how physical objects behave.[22] But they make sense if this understanding is rooted in a rationalist learning system, an innate domain-specific system for reasoning about physical events. This would be an innate system with schemas for representing different types of physical events—occlusion, containment, covering, etc.—and that is updated in piecemeal fashion as infants learn about the variables that are relevant to these types of events. On this view, concepts such as OBJECT, OCCLUSION, and CONTAINMENT are innate components of an innate physical-reasoning system.

There are many facts about people's minds and behaviour that are highly surprising or unexpected and remain so on the assumption that all learning takes the form of domain-general learning but that begin to make sense given a rationalist perspective. We have worked through a variety of examples of this kind (involving

[21] See also Stahl and Feigenson (2015) for related evidence that when a physical event violates infants' expectations, infants treat this as a learning opportunity and actively engage in exploratory actions on their own to test possible explanations of the violation. This adds to the case that the representations are conceptual in that infants also appear to be involved in planning and undertaking actions that would provide them with further evidence about how to interpret a previously experienced violation of expectation.

[22] For discussion of the relation between statistical learning and explanation-based learning of the type discussed here, see Wang (2019).

such concepts as ROUTE, ALLIANCE, COOPERATION, OBJECT, and CONTAINER) to illustrate the last of our seven arguments for concept nativism—the argument from cognitive and behavioural quirks. Like the quirks that have been discovered regarding geometrical representation and navigation, the quirks associated with the estimation of distance, with racial categorization and the representation of coalitions, and with infants' representations of objects and different types of physical events are all deeply puzzling for empiricists—empiricist accounts just aren't equipped to explain them. By contrast, these quirks not only make sense when viewed in light of rationalist accounts, but attempts to explain them open new avenues of research, further enhancing the explanatorily fruitfulness of rationalist approaches. These and other quirky aspects of the human mind support another line of argument in favour of the view that the acquisition base contains a rich variety of characteristically rationalist psychological structures that are critical to understanding the origins of concepts across a broad range of conceptual domains.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. DOI: 10.1093/9780191925375.003.0015

16
Conclusion to Part II

Though it is rarely recognized in either philosophical or scientific circles, the positive case for concept nativism is overwhelming. Part of the problem is that the claims of concept nativism have been misunderstood in ways that were discussed in Part I. But another part of the problem is that the *arguments* for concept nativism have frequently been misunderstood too, and haven't been given their full due. For example, from Locke's time to the present, concept nativism's critics have been too cavalier in their handling of the argument from universality and the argument from animals. They have held rationalists to unrealistic and inappropriate standards of evidence for arguments like the argument from early development. They have largely overlooked or failed to appreciate the significance of the argument from initial representational access or the argument from prepared learning. They have not really even recognized the argument from cognitive and behavioural quirks or the argument from neural wiring. And they haven't fully appreciated the scope of the case for concept nativism, in terms of the variety of arguments for concept nativism, the interplay between these arguments, and the wealth of empirical evidence that factors into these arguments.

In Part II, our aim has been to distinguish and clarify these seven important types of argument for concept nativism. We have tried to explain exactly how these arguments are supposed to work and have highlighted both their individual contributions to concept nativism and the fact that they often provide mutually supporting considerations that converge on the same rationalist conclusions. In the course of explaining these arguments, we also sketched some of the empirical evidence for a number of case studies where these arguments apply, drawing on a range of content domains to indicate the breadth of the case for concept nativism (the large number of content domains that are rooted in rationalist local acquisition bases) and its depth (the interactions of mutually supporting arguments and findings that support these claims). As we've tried to emphasize throughout, there is an intricate interplay between theory and evidence. Interpretations of empirical evidence must be defended against numerous possible alternative interpretations and in light of the arguments and evidence that might be marshalled in favour of these alternatives.

Perhaps the most important moral of Part II is that it is a mistake to take the case for concept nativism to be a deductive argument. Instead, it takes the form of an inference to the best explanation. Each of the seven arguments we have elaborated on is itself an inference to the best explanation. But also, different packages

of these arguments work together to form even richer explanatory arguments, and ultimately all seven arguments taken together cumulatively constitute a single broad inference to the best explanation argument for our version of concept nativism—the view that many concepts across many content domains are either innate or acquired via rationalist learning mechanisms.

We have divided our discussion into seven separate arguments, but as we noted at the outset, this was largely for expository convenience. Our intention was never to imply that these are mutually exclusive or that they exhaust the case for concept nativism. In fact, there are other ways of carving up the landscape that might in principle be just as useful.

For example, we previously noted that the argument from animals could easily be divided into multiple arguments. One is a type of argument in which animal data is used to show that it's possible to acquire certain concepts in carefully controlled conditions with total or near total deprivation of relevant input prior to being tested—something that is not possible in the human case. Another is a type of argument in which animal data is used to show that a given conceptual capacity can be acquired without the sort of powerful general-purpose learning mechanisms that have only been attributed to human learners. These might be called *the argument from deprivation experiments* and *the argument from the absence of powerful general-purpose learning*. Or, to take another example, the argument from neural wiring could likewise be divided into a number of distinct arguments. For instance, one type of argument we gave in the chapter on the argument from neural wiring was that neural structures and functions relevant to conceptual representation can be preserved in the face of dramatic variation in sensory input, including cases where there is a complete absence of relevant sensory input. A different type of argument in the same chapter was that certain concepts can *fail* to be acquired despite their enormous utility and despite the presence of intact domain-general learning mechanisms (this happens, for example, in certain cases of focal brain damage). These might be called *the argument from preserved neural structure and function* and *the argument from the failure of compensatory neural plasticity*.

At the same time, although we presented the seven arguments in Part II each as a separate argument, there is enough conceptual overlap between some of them that in principle they might have been merged. For example, a number of the arguments might be viewed as variations on the more general structure of a poverty of the stimulus argument.

The core idea behind poverty of the stimulus arguments is that the environmental input to acquisition (the stimulus) is insufficient in some way for an empiricist learner to reliably acquire the psychological trait in question given the known facts about development (where an empiricist learner is one whose learning traces back solely to characteristically empiricist psychological structures in the acquisition base, especially domain-general learning mechanisms). This idea

can be developed in many different ways. For example, it may be—as in the argument from early development—that the stimulus is impoverished in that the acquisition timeframe in which a representational capacity is acquired is too short. There simply *isn't enough time* to ensure exposure to the data that would be needed for an empiricist learner to arrive at this outcome. Or it may be—as in the argument from universality—that the stimulus is impoverished in that the relevant input *is not available (or not reliably available) in certain environments* where the capacity is acquired. In these cases, an empiricist learner wouldn't be able to reliably acquire a capacity that is known to develop across a range of variable environments. Or, to mention one further example, it may be—as in the argument from initial representational access—that the stimulus is impoverished in that the relevant input *is simply inaccessible* to an empiricist learner given the types of representations such a learner can call upon to represent the stimulus.

Given this way of recasting these arguments, one could easily say that there is a general master argument for concept nativism that covers a lot of the ground we have covered with these separate arguments. But distinguishing our seven arguments for concept nativism has significant advantages. It has allowed us to highlight different strands in this master argument that argue for concept nativism in importantly different ways, which we think brings greater clarity to how a vast body of data links up with concept nativism's central claims.

Despite the many arguments and supporting findings we have given in Part II, we are keenly aware that we have had to be selective in choosing which data to discuss, which competing interpretations of the data to consider, and which conceptual domains to use to illustrate our main points. It is our claim, however, that the balance we have struck between the breadth of the conceptual domains examined and the depth of the evidence and arguments marshalled in each of these cases overall makes a compelling case for concept nativism.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0016

PART III

ALTERNATIVE EMPIRICIST PERSPECTIVES

17
Methodological Empiricism

We have been building our case for concept nativism in stages. We began in Part I by comprehensively rethinking the foundations of the rationalism-empiricism debate, both in general and as it applies to the origins of human concepts. Building on others' ideas, we presented a new account of what is at stake in these debates while showing how easy it is to misunderstand the debate in unproductive ways which can cause theorists to unwarrantedly reject the entire debate (or rationalist views in particular). One central moral of Part I was that the rationalism-empiricism debate about concepts isn't just—or even primarily—about innate concepts. Rather, it's about the contents of the acquisition base—whether it contains characteristically rationalist psychological structures and, if so, which ones—and about the extent to which rationalist learning mechanisms are involved in concept acquisition and conceptual development.[1]

In Part II, we presented the heart of our positive case for concept nativism, distinguishing seven distinct but complementary arguments for concept nativism and showing how they support a rationalist perspective on conceptual development across a broad range of conceptual domains. But our case for concept nativism can't end there. This is because our concept nativism faces opposition on two fronts. On the rationalist side, it stands opposed to implausibly extreme versions of rationalism, such as Jerry Fodor's radical concept nativism, which rejects the possibility of any type of concept learning—rationalist or empiricist—and holds that virtually all lexical concepts are innate. Despite the implausibility of Fodor's claim, his view cannot simply be dismissed, and as we will see, there is much to be learned from a detailed examination of his arguments for the view. On the empiricist side, it stands opposed to views that aim to get by with just domain-general learning mechanisms. With the heart of our positive case for concept nativism having been laid out in Part II, we now need to say something about both of these rival approaches to conceptual development. Our response to radical concept nativism (and a fuller discussion of the role of learning in cognitive development) can be found in Part IV. In this part of the book—Part III—we will critically examine empiricist alternatives and objections to concept nativism.

[1] For readers not reading the chapters in order, there are a number of technical terms that were introduced and explained earlier that we will continue to rely on in Part III, including "acquisition base", "rationalist learning mechanism", "characteristically rationalist psychological structures", "articulation", and "alignment". Brief summaries of how we are using these and other terms may be found in Boxes 1–7 in Chapter 2 and in Box 8 in Chapter 6.

Empiricism has arguably been the dominant position in the rationalism-empiricism debate historically. And it remains very well represented today—in philosophy and cognitive science and across the social sciences. So there are a great many empiricist positions that could be discussed—far too many to provide anything like an exhaustive discussion here. Our approach instead will be to work through a representative sample of views that have been thought to show that concept acquisition can be fully accounted for without having to postulate anything like our rationalist acquisition base. The empiricist views that we will examine also have the advantage that they offer a natural point of departure for taking a more detailed look at a number of further conceptual domains. Some of these views are tied to general approaches to explaining conceptual development, for example, to theories that model development using domain-general neural networks or to theories that are bound by the assumption in embodied cognition research that concepts must be realized in sensorimotor and affective systems and shouldn't be understood as amodal representations. Others are tied to more specific objections to concept nativism, for example, to scepticism about the way that rationalists sometimes draw inferences about conceptual development by examining its breakdown in individuals with a particular type of developmental disorder. Some are driven by general theoretical or methodological considerations that are taken to favour empiricism. Others are based on wide-ranging experimental research programmes. Some are open to the existence of at least some characteristically rationalist psychological structures. Others adopt a radical form of empiricism, rejecting these entirely.

We will argue that none of these empiricist alternatives undermine concept nativism. Quite the contrary: Critical evaluation of concept nativism's empiricist opposition only reinforces the explanatory appeal of concept nativism. We should note, however, that, in rejecting empiricism, we don't reject all of the tools that have been associated with empiricist theorizing. It turns out that many of these are perfectly consistent with concept nativism and should be adopted by rationalists too. In fact, one of the morals of Part III is that these empiricist innovations are capable of making a far greater contribution to our understanding of conceptual development when incorporated into a rationalist framework.

In this chapter, we will begin our discussion of empiricist alternatives by examining a type of argument that plays an enormous role in the opposition to rationalist accounts and in framing empiricist approaches to the debate about the origins of concepts. This argument is about who has the burden of proof in the rationalism-empiricism debate. Since the argument focuses on this general methodological question rather than on any particular finding, we will refer to the view that the argument aims to support as *methodological empiricism*. According to methodological empiricism, empiricism should always be considered the default position in the rationalism-empiricism debate about the origins of

concepts, and rationalist accounts of the origins of concepts should only be adopted if all possible empiricist alternatives have been ruled out.[2]

Why has empiricism been taken to hold this privileged position in the rationalism-empiricism debate? One reason, which we discussed in Chapter 4, is the supposition that empiricism is more parsimonious than rationalism. Rationalism, by comparison, is supposed to be extravagant with its commitment to innate concepts and numerous special-purpose rationalist learning mechanisms. A second reason is that empiricists are often sceptical about the viability of some of the key experimental methodologies that are associated with rationalist theorizing about conceptual development. This scepticism has led many empiricists to conclude that rationalist theories are empirically unfounded—that they stem from an inflated reading of data that can be readily explained in terms of the sorts of domain-general learning mechanisms whose existence empiricists *and* rationalists already acknowledge.

In an influential critique of rationalist work on infant cognition, Haith (1998) accuses rationalist researchers of "over-interpretations of findings as evidence for high-level cognitive operations…in the absence of adequate definitions or anchoring observations or procedures" (p. 168). He is especially critical of rationalists' reliance on experiments in which conclusions are drawn from measuring how long infants look at different visual stimuli:[3]

> One difficulty is that people developed this paradigm to address sensory and perceptual questions, not questions of high-level cognitive processing. Many factors affect looking, including variations in the perceptual dimensions of objects and people, familiarity, novelty, recency, predictability, and the time lapse between stimulus exposures.

It is in this context that Haith expresses his commitment to methodological empiricism:

> Of course, a paradigm that is created for one purpose may be adapted for another, but investigators who pursue high-level cognitive constructs must play the default game. That is, one must fend off every possible perceptual interpretation of differences to entertain default cognitive interpretations. Surely, there are alternative interpretations for any experiment, but the use of perceptual

[2] We have discussed this view briefly already in Chapters 1 and 4. But given its importance in empiricist theorizing, it deserves a fuller discussion here. This discussion will also allow us to expand on a number of important points concerning conceptual development, the experimental methods used in developmental psychology, and the representation of physical objects.

[3] Haith puts his critique in terms of doubts about the "looking paradigm". However, this term is a misnomer as a number of distinct experimental methodologies rely on measures of infants' looking time (see below and Chapters 8 and 9).

> paradigms tends to favor well established perceptual explanations. Even when an immediate perceptual explanation is not obvious, there is the danger that one will come along. (p. 170)

In much the same spirit, Prinz (2012) claims that empiricism is "more economical" and that this makes it "the default position until evidence weighs in favour of Rationalism" (p. 90). And like Haith, Prinz holds that rationalists face a daunting burden:

> To prove that core knowledge must be innate, Rationalists would need to show that the kind of knowledge they attribute to infants is of a type that would be impossible to learn by observation. But… [rationalists] rarely make any effort to do this. In fact, it is easy to imagine that the knowledge they attribute to infants is learned. (p. 110)

Although Haith's and Prinz's views aren't particularly unusual in empiricist circles, it is important to recognize how extreme they are. For example, Haith insists that for any given body of infancy data, rationalists first need to fend off *every possible* perceptual interpretation. For theorists who suppose that infants respond mostly to low-level perceptual similarities and differences, this may seem innocuous. But by insisting that all possible perceptual interpretations have to be eliminated before a non-perceptual interpretation is even contemplated, Haith is effectively postponing the consideration of a non-perceptual interpretation indefinitely. After all, there is an unlimited supply of possible perceptual interpretations for any behaviour if we include interpretations that might be considered ad hoc.

The first thing to note here is that, in line with our discussion in Chapter 4, there are at least two reasons why parsimony can't bear the weight that Haith, Prinz, and other empiricists have asked of it. First, just as there isn't a default position in the rationalism-empiricism debate about the mind in general, there isn't a default position in the rationalism-empiricism debate about the conceptual system. The problem, as before, is that empiricists, without justification, seize on a very particular, narrow sense of parsimony at the expense of an appropriately comprehensive understanding of parsimony. This narrow sense simply counts the number of types of distinct psychological structures in the acquisition base. On this way of measuring how parsimonious a theory is, it is certainly true that empiricist theories of conceptual development are more parsimonious than rationalist theories. However, as Spelke (1998) points out, in considering how parsimonious a theory of development is, we need to take into account the whole proposed process of development. We shouldn't focus exclusively on one aspect of this process, such as the quantity of psychological structures in the acquisition base. Rather, to the extent that parsimony is relevant, we need to pay attention to the entire account of how a mature state is said to be acquired—in this case, how

parsimonious a theory is in explaining the acquisition of a given conceptual capacity. Since this can't be decided in advance of having a relatively worked out theory and ascertaining how it proposes to account for the evidence, there simply is no default view to be had.[4]

Second, as philosophers of science have argued, and was emphasized in Chapter 4, parsimony only really comes into play when two theories are equally capable of explaining the same evidence. Thus, it is not applicable until we are in a position to determine whether competing empiricist and rationalist theories *are* equally capable of explaining the evidence. And to do that, we need to have relatively worked out theories to compare. It makes no sense to say that one of these approaches—the empiricist approach—can be seen in advance of inquiry to be inherently more parsimonious.

Still, this leaves the concern about rationalist experimental methodologies, particularly experiments in which we are asked to draw conclusions about how infants represent the world based on evidence about what they look at. In the rest of this chapter, we want to address this concern by revisiting and extending our examination of research on how infants represent physical objects.

Let's begin with what Haith and Prinz have to say about rationalist theorizing that relies on infancy looking-time data. Haith and Prinz both proceed by selecting a few influential studies and by proposing alternative explanations of the data—alternatives that they take rationalists to have overlooked. Since these alternatives don't require a commitment to rationalist learning mechanisms, they are supposed to illustrate the recklessness of rationalist theorizing as well as the resilience of empiricism.

One of Haith's core examples is an important early violation of expectation experiment from the 1980s—Baillargeon's (1987) drawbridge study. In this study, 3.5- and 4.5-month-old infants faced a cardboard screen that rotated smoothly through a full 180° arc like a drawbridge (see Figure 17.1). Once they were familiarized with this event, a box was placed in the screen's path at a point that would ordinarily block it from completing its rotation (at 112°). Crucially, from the infants' perspective, the box became hidden behind the screen as the screen rotated towards the box, so that any contact between the screen and the box couldn't be seen—it would have to be inferred. This was followed by one of two

[4] Relatedly, as we noted earlier in connection with the rationalism-empiricism debate about cognitive traits more generally, there are many types of parsimony to consider in addition to parsimony regarding the number of elements in the acquisition base. For example, there is parsimony regarding the number of computations that are required to learn something, parsimony regarding the amount of memory and attention needed, parsimony regarding the energetic costs of the learning process, and parsimony regarding changes posited over evolutionary time in order for the acquisition base to realize this type of process—what we referred to earlier as *computational parsimony*, *energetic parsimony*, and *phylogenetic parsimony*. These different types of parsimony pull in different directions. And it is precisely because empiricist theories are more parsimonious in terms of the quantity of psychological structures posited in the acquisition base that they tend to be significantly less parsimonious in other ways, for example, less computationally parsimonious.

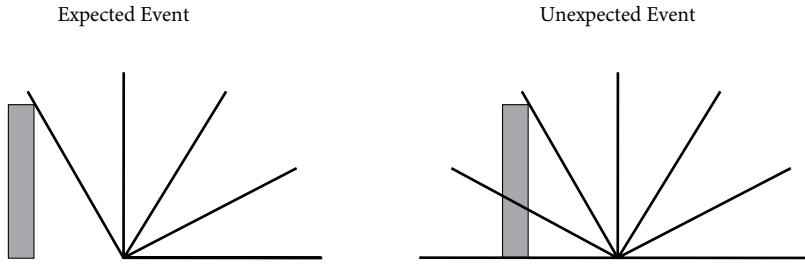

Figure 17.1 Baillargeon's (1987) drawbridge study. This figure presents a side view of the experimental setup. The infant views the screen (from the right-hand side) facing the block, so that they can no longer see the block once the screen rises like a drawbridge. Infants looked longer at the unexpected test condition on the right, where the drawbridge rotates 180° as if there is no impediment. This suggests that infants represent the existence of the occluded box and infer that it should block the rotating screen. (Figure based on Baillargeon 1987, figure 1.)

different test conditions. In the *expected event*, the screen stopped at the point of contact, as if the box blocked it from completing its rotation. In the *unexpected event*, the screen completed the full 180° arc as if there were no impediment. The result was that infants looked significantly longer at the unexpected event than the expected event.[5]

Recall that in a violation of expectation experiment, the terms "expected" and "unexpected" are generally defined relative to how adults represent the events. The logic of this experimental method is that we can ask whether infants are representing core aspects of certain types of events in similar ways to adults by determining whether they show greater interest in an event that, by adult standards, violates an expectation. In this case, the expectation in question is that the box will continue to exist even when it can no longer be seen and that it ought to block the rotation of the screen because one solid object can't pass through another. Since infants apparently have this expectation, rationalists have often interpreted Baillargeon's finding as providing evidence that 3.5- to 4.5-month-olds are already showing signs of object permanence—taking an object (the box) to continue to exist and to have its usual effects even though it is not currently perceptible.

Haith will have none of this. He argues that a rationalist interpretation is unsupported because alternative empiricist explanations are always available. Haith illustrates this by suggesting the empiricist alternative that it is possible that

[5] The results were actually more complicated in ways that critics have taken to be important. While the 4.5-month-olds looked longer at the unexpected event, the 3.5-month-olds' looking depended on how quickly they lost interest in the familiarization events that preceded the appearance of the box. The 3.5-month-olds who lost interest more quickly looked longer at the unexpected test event, while the 3.5-month-olds who lost interest more slowly looked equally at the two test events. The significance of these mixed results with 3.5-month-olds is discussed below.

infants have "lingering sensory activation" during the test condition, so that it is as if they are still seeing the box for a few seconds even when it is actually concealed by the screen:

> Is it possible that infants have some lingering sensory activation, in a way, still "seeing" the barrier box as the drawbridge swings backwards? Let's rid ourselves of the occlusion part of this experiment so that we can think about it more clearly. Say the infant looks at this episode at an angle from the side, so that she can actually see the box behind the drawbridge as it moves toward the box and magically moves through it. We are not at all surprised if infants look at that episode longer than when the drawbridge stops on contact. Why? Because infants often see one moving object contact another but never see one solid object go through another…I believe that there is nothing about the typical occlusion event that requires us to use different principles. (Haith 1998, p. 173)

Haith's thought is that there is no need to claim that infants represent the existence of an unperceived object. The unexpected nature of the event can instead be explained simply on perceptual grounds. Because of the infants' lingering perception of the box, they effectively "see" the screen pass through the box. Perceptual events of this type are unexpected since infants are unaccustomed to seeing one object pass through another.

Prinz discusses this same experiment but suggests a different alternative explanation. On Prinz's account, infants look longer at this event not because it is unexpected, but because it is familiar and because infants look longer at familiar events.

> Researchers customarily interpret long looking times as indicating surprise. But we all know that infants and children take great pleasure and interest in repetition. Thus, longer looking times might initially indicate that something is consistent with expectations, rather than contrary to expectations ('Yay! More of the same!'). (Prinz 2012, p. 92)

Prinz's point is that the test event that we have been describing as unexpected is also potentially more familiar within the confines of the experiment. This is because the familiarization events that precede the test events are of the screen completing the full 180° rotation too. So maybe infants are simply exhibiting a preference for the familiar when they look longer at the event in the condition that from the adult's perspective is unexpected. Prinz suggests that this alternative explanation is bolstered by Schilling (2000), who found that infants failed to look longer at this type of event when there were roughly twice as many familiarization trials as in Baillargeon's original study. The reason this is supposed to support Prinz's explanation is that infants may enjoy repetition but only up to a

point, after which they get bored. For Prinz and Schilling, the repetition in Baillargeon's original experiment was still enjoyably familiar, explaining why the infants in Baillargeon's study looked longer at what we're referring to as the unexpected event (the old 180° rotation). But the repetition in Schilling's experiment meets the boring threshold, explaining why the infants in Schilling's study looked longer at what we're referring to as the expected event (the novel 112° rotation). According to Prinz, this goes to show that "[t]he evidence that infants understand physical principles disappears" (2012, p. 92).

It is hard to know how seriously Haith and Prinz take these particular proposals regarding what is going on in the drawbridge experiment. Sometimes they sound as if all that matters to them is that, in principle, there exist any kind of empiricist alternative to a rationalist explanation, not that there is any real support for the specific empiricist account that they propose. Still, these are the alternatives they offer and that they presumably suppose to be representative of their methodological critique of rationalism. So we should ask whether these alternatives are better than Baillargeon's explanation, which instead attributes to infants a precocious grasp of object permanence.

In Haith's case, the proposed alternative depends on infants continuing to "see" objects for a few seconds after they are occluded. However, there is no evidence whatsoever for this supposition, and Haith doesn't offer any. There is also considerable evidence that Haith's supposition is mistaken. For one thing, as Spelke (1998) points out, the duration of visual sensory persistence in adults is known to be *far* shorter than the three to four seconds Haith estimates infants would need on the lingering perceptual contact account. For another, additional experiments by Luo et al. (2003) show that infants respond in much the same way—looking longer at the physically unexpected event—even when the gap between the disappearance of the object and the test condition is as much as three or four *minutes* long, a delay that eliminates any possibility of their having the needed "lingering sensory activation".

What about Prinz's suggestion that infants in the drawbridge study are expressing a preference for the familiar? Prinz is certainly right that in some circumstances infants do prefer the more familiar of two stimuli. However, it is a dubious suggestion to make in this case since Baillargeon explicitly tested this alternative and refuted it in this same study. She did this with a simple and elegant control condition in which infants saw the same familiarization events as before (with the drawbridge rotating through a full 180°) but with no box placed in the path of the screen in the test trials. So although the test event in which the screen rotated 180° was still more familiar, it was no longer unexpected, as there was no longer the appearance that one solid object passed through another. On Prinz's proposal, the infants should have looked longer at the 180° test event in this case as well, and for precisely the same reason, namely, because they enjoy repetition. (It was to address and rule out exactly this "repetition" hypothesis that Baillargeon

ran this control condition.) What Baillargeon found, however, was that the infants in this condition looked equally at the 180º and 112º test events.[6]

Prinz's methodological empiricism takes the same form elsewhere in his critique of concept nativism. His overall approach is to offer alternative empiricist explanations of a small number of experiments that are associated with rationalist theorizing. As we have just seen, Prinz's alternatives may not be terribly plausible ones—indeed they are sometimes even ones that his rationalist opponents have anticipated and experimentally ruled out. Nonetheless, Prinz seems to think that just mentioning these alternatives is all that an empiricist needs to do.[7] Because, in his view, empiricism is the default position in the rationalism-empiricism debate, it is enough to undermine a rationalist theory of a representational capacity to show that the data is "open to interpretation" (Prinz 2012, p. 92) and consequently that the situation is "inconclusive" (p. 109).

In his critique of rationalist views of physical object representation, Prinz (2005) discusses another example—one that we encountered earlier (in Chapter 10). He offers an alternative empiricist interpretation of Kellman and Spelke's (1983) study of the development of object representation, from which they draw the

[6] Schilling's (2000) study, which Prinz cites in support of his interpretation, also faces a number of problems (see Baillargeon 2000). Perhaps the most important of these is that it confounds the habituation-dishabituation method with the violation of expectation method. In habituation-dishabituation studies, it is crucial that the participants become bored with the habituation stimuli (i.e., lose interest in the stimuli to a significant degree) so that they can dishabituate to the test stimuli if they represent the test stimuli as different in kind than the habituation stimuli. However, in a violation of expectation experiment, there is no need for infants to become bored by viewing the familiarization trials prior to witnessing the test events (in fact, doing so may well be counterproductive). The function of the familiarization trials is merely to prepare infants for the complex events in the test condition so they are able to fully process them. The difference between these two methods wasn't perfectly clear when Baillargeon published her original study; at that time, she described her familiarization trials as "habituation trials". But in hindsight it is quite clear that what matters in experiments like those in the drawbridge study is not whether the infants view the test trials as different in kind than the initial trials. Rather, what matters is that the infants have sufficient familiarization with the experimental setup so that they are able to fully process the test trials. This is particularly clear in light of the contrasting looking times in the experimental and control conditions given exactly the same type of initial trials.

So what are we to make of the fact that Schilling's infants didn't look longer at the unexpected event when given twice as many familiarization trials? As Baillargeon points out, probably what happened is that the infants became so bored with the experimental setup that they weren't able to fully engage with the test events. This would have led them to respond to a relatively superficial change—the drawbridge merely rotating a different amount. Further experimental work that Baillargeon (2000) reports shows that slightly older infants (6-month-olds) lose interest with the number of familiarization trials that Schilling used but do fine with as few as *one* familiarization trial. Baillargeon also notes that the looking times in the test conditions in Schilling's experiments are much lower than in Baillargeon's original experiments—a further reason for supposing that Schilling's infants weren't sufficiently engaged by the test events.

[7] Regardless of what one thinks of Prinz's form of argument, it is terribly misleading for him to suggest that his rationalist opponents have *overlooked* these alternatives. Readers who are relying on Prinz's summary of the infancy research could easily come away with the false impression that Baillargeon was unaware of the possibility that infants could be looking longer at the unexpected event in the drawbridge study because it is more familiar. Not only was she aware of the kind of empiricist alternative proposed by Prinz, but she explicitly tested it experimentally—and was able to reject it—in the very study Prinz targets for criticism.

(a) Habituation Stimuli (b) Test Stimuli

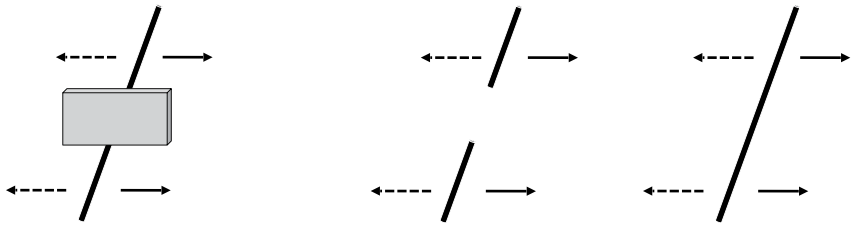

Figure 17.2 Stimuli from experiment 1 in Kellman and Spelke (1983). (a) Infants saw rod ends moving in sync, one above and one below a block. (b) Following habituation, infants saw two stimuli that were both compatible with the habituation event: (1) two smaller rods and (2) a single continuous rod. (Figure based on Kellman and Spelke 1983, figure 3.)

conclusion that infants as young as 4 months old can represent partially occluded objects as such. For Kellman and Spelke, their data suggest the rationalist theory that infants possess "an unlearned conception of the physical world" (p. 521). Recall that infants were shown an event that could be seen in two ways: (1) as a single intact rod's top and bottom protruding out from behind a block, or else (2) as two aligned smaller rods moving in sync, one at the top of the block and the other at the bottom of the block (see Figure 17.2(a)). After becoming habituated to this event, the infants were subsequently shown a single rod and two smaller rods (in both cases, without an occluder, as in Figure 17.2(b)), and were found to look significantly longer at the two smaller rods. This suggests that they were perceiving the habituation event like adults—as involving a single connected rod whose middle was hidden behind the block—and were responding to the novelty of the two smaller rods in the test trials.

Prinz deems this conclusion "too hasty" (2005, p. 689) and proposes instead that "perhaps they were staring longer simply because two objects are more interesting than one. Despite a variety of clever control conditions, Kellman and Spelke do not adequately rule out this hypothesis" (2005, p. 689). According to Prinz's hypothesis, the reason that infants look longer at the two shorter rods in the test trials has nothing to do with how they interpret the stimuli in the habituation trials. They look longer at the two rods simply because they find two objects more interesting than one—that is, they simply have a baseline preference for the more complex stimulus. Prinz claims that this hypothesis isn't refuted by Kellman and Spelke's many experimental conditions. In arguing for this claim, Prinz considers a single condition—one that he takes to provide prima facie evidence against his hypothesis—and builds a case that the outcome of this condition turns out to be consistent with his hypothesis after all.

In evaluating Prinz's response to Kellman and Spelke, there are a number of factors that need to be taken into account, including Prinz's characterization of the experimental condition he focuses on and the plausibility of his alternative

hypothesis. But just as important—perhaps even more important—is his argumentative strategy or the form that his anti-rationalist argument takes in his response, which exemplifies some of the central elements of methodological empiricism. We will start with this strategy.

To assess the argumentative strategy, it will help if we adopt the simplifying assumption that Prinz is right that his proposed explanation can account for the specific results he mentions—that these results are consistent with the hypothesis that infants merely have a baseline preference for two rods over one. Does this assumption undermine Kellman and Spelke's argument that infants can represent partially occluded objects as such? Not at all. The problem is that even if Prinz's hypothesis is consistent with the outcome of this one condition, he has done nothing to show that it is also consistent with the dozen or so further conditions that Kellman and Spelke's paper examined. It should go without saying that Prinz's hypothesis could be consistent with one condition (the one he happens to mention) and inconsistent with others. We will shortly argue that this is exactly the situation Prinz is in—there are a number of experimental conditions in Kellman and Spelke's paper that Prinz fails to mention where the results strongly argue against his hypothesis.

So why does Prinz only consider whether his hypothesis is consistent with a single experimental condition? Although he doesn't say much about these methodological issues, one possibility is that he takes this condition to be especially problematic for his proposed hypothesis. In that case, perhaps his thinking is that, if he can show that his hypothesis is compatible with the results in this condition, then he has undermined the main or most difficult challenge it faces. Another possibility is that Prinz has chosen one of Kellman and Spelke's conditions more or less randomly, on the assumption that this condition is representative of the others or at least not unusual. In that case, the thought would be that if an empiricist explanation can be given of the data in this condition, a comparable empiricist explanation should be possible for any of Kellman and Spelke's experimental conditions.

Either way, in order to see why this type of strategy is flawed, it is worth reflecting on why Kellman and Spelke employed so many experimental conditions to begin with. This was because they were trying to simultaneously test a range of competing hypotheses that are associated with different theories about the origins of object representation. They took different experimental conditions to be relevant to different hypotheses and were interested particularly in these three:

> *The gestalt hypothesis*—according to which infants have an innate set of gestalt principles and represent objects only in terms of general perceptual factors related to good form or good continuity. As Kellman and Spelke note, "on this view, infants should perceive the complete shapes of partly hidden objects as soon as they can detect certain configural relationships in visual scenes, such as the alignment of visible surfaces and the similarity of their colors and textures" (p. 485).

> *The Piagetian hypothesis*—according to which infants only gradually develop an understanding of objects over the first two years of life. On this view, the 4-month-olds in this study should not be capable of representing partial occlusion as such; they should only see arrangements of visible surfaces.
>
> *The object occlusion hypothesis*—according to which infants initially take the world to be populated by physical objects that move in front of and behind one another, leading to incidents of occlusion. On this view, infants "should perceive the surfaces as parts of a single object if the surfaces move rigidly together" (p. 486), but may remain agnostic about whether such surfaces constitute a single object in the absence of such cues. This is the hypothesis that Kellman and Spelke endorse.

Importantly, all three of these hypotheses—along with Prinz's hypothesis—are compatible with individual conditions in Kellman and Spelke's study when considered in isolation. So when Prinz claims that his preferred hypothesis is consistent with the data from one of these conditions, that can't tell us very much. The important question is which if any of these hypotheses can accommodate the full range of findings. What Kellman and Spelke show is that the theory that does best by this measure is the object occlusion hypothesis and consequently that we have good empirical grounds for supposing that 4-month-olds represent object occlusion.

We won't work through all of the considerations that favour this hypothesis relative to the competition. But just to give a sense of the overall structure of Kellman and Spelke's argument, we will briefly mention the sorts of considerations that Kellman and Spelke used to discriminate between the gestalt hypothesis and the object occlusion hypothesis. On the object occlusion hypothesis, synchronized motion is an important cue that two surfaces are part of a single object. If two surfaces rigidly move together, this is a good indication that they are surfaces of the same object even if they aren't seen to be connected (due to the presence of an occluder). On the other hand, the gestalt principles regarding good form and good continuity aren't grounded in this particular type of motion cue. So unlike the object occlusion hypothesis, the gestalt hypothesis predicts that infants will behave the same regardless of whether an occluded rod is in motion or not. To adjudicate between these two hypotheses, then, we can compare the result from the original occlusion condition (Figure 17.2) with a comparable condition that is completely static. Kellman and Spelke did just this and found that the motion cues are critical. When they are removed in the static condition, infants behaved differently than in the original occlusion condition: They now looked equally at the single connected rod and the two rod ends in the test trials.

What this goes to show is that we shouldn't focus on whether a given hypothesis is compatible with the result from any single experimental condition. Both the object occlusion hypothesis and the gestalt hypothesis are compatible with

some of these results. It is only when we turn to how they fare with the broader range of experimental conditions that we can tease them apart and see that the gestalt hypothesis doesn't do nearly as well as the object occlusion hypothesis.

Let's return now to Prinz's hypothesis that the infants simply have a baseline preference for two rods rather than one. As we noted earlier, there are a number of conditions in Kellman and Spelke's paper that Prinz does not mention that strongly argue against his hypothesis. Consider, for example, the very first experiment in Kellman and Spelke's paper. This experiment compared two situations: (i) the occlusion condition discussed above, where there are two visible rod segments moving in sync at the top and bottom of a stationary occluder (see Figure 17.2(a)) and (ii) the same arrangement but with just the top segment moving (see Figure 17.3(a)). (For adults, the latter doesn't give rise to the impression that there is a single, centre-occluded rod.) Note that Prinz's hypothesis predicts that infants should look longer at the two rod segments in *both* cases, since on this hypothesis they are simply responding in accordance with their baseline preference for two objects over one. By contrast, Kellman and Spelke found that infants looked equally at the two test events in the second condition, showing no preference for the two rod segments. Likewise, in another experiment, Kellman and Spelke asked how infants would behave if the visible rod segments and the block all moved in sync (see Figure 17.3(b)). Again, Prinz's hypothesis predicts that infants should look longer at the two rod segments in the test trials, since on this hypothesis they are simply responding in accordance with their baseline preference for two objects over one. But Kellman and Spelke found that here, too, infants showed no preference for the two smaller rod segments. Other experiments in Kellman and Spelke's study also argue against Prinz's baseline preference hypothesis, including experiments that employed stationary stimuli (none of the rods were in motion) and experiments in which the block was behind the rods instead of in front of them (so that the long rod was fully exposed). Infants in none of the relevant conditions in these other experiments exhibited a preference for the more complex test stimulus.

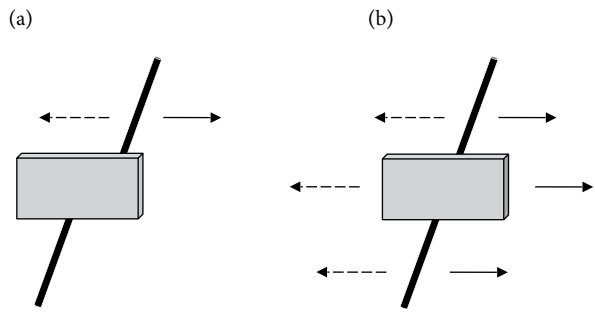

Figure 17.3 Habituation stimuli from additional experimental conditions in Kellman and Spelke (1983) that argue against Prinz's alternative hypothesis. (Figure based on Kellman and Spelke 1983, figures 3 and 11.)

Just as compatibility with the results of one or two experimental conditions is not enough to vindicate the gestalt hypothesis or the Piagetian hypothesis, it is not enough to vindicate Prinz's hypothesis. Hypotheses need to be evaluated against the full set of experimental findings in a given study. More generally, as Spelke rightly highlights elsewhere (as one of four guidelines she proposes for research on early cognitive development): "All theories of early cognitive development must encompass all the relevant data . . . no account merits attention if it is based on a small subset of findings and ignores contrary results" (1998, pp. 191–192).

What about the one experimental condition that Prinz considers when making his case for the baseline preference hypothesis? The experiment Prinz has in mind is one in which there is no occluder in the habituation trials. Infants are simply habituated either to a moving and fully visible long rod or else to two smaller aligned rods moving in sync (as in Figure 17.2(b)). As Prinz put it:

> In the control condition, infants see two bars moving with a gap in between them. This looks just like the two bars in the test condition . . . In other words, the appearance of the bars in the control condition is just like the appearance of the bars in the test condition, because there is an unfilled gap. (2005, p. 689)

Prinz reports that infants subsequently don't look longer at the two small rods in the test trials and hence that this might appear to argue against their having the baseline preference that he is proposing. But he goes on to suggest a reason why appearances may be misleading. He claims that their baseline preference is obscured by the fact that the habituation trials in this condition repeatedly show two rods, which looks just like the two-rod test stimulus. So by the time they see the test stimuli, they have lost interest in the stimuli, despite their baseline preference for the two rods. In contrast, although infants see two rod segments (above and below the block) multiple times in Kellman and Spelke's initial experiment too (see Figure 17.2), this is in a different context—the habituation trials include the stationary block as well. According to Prinz, removing the block in the test trials is just enough of a change to renew their interest and bring them back to their baseline preference.

There are a number of problems with Prinz's discussion of this experimental condition. One problem is that Kellman and Spelke intended this experimental condition (which is slightly different than Prinz's description) to test a completely different hypothesis—the gestalt hypothesis.[8] So Kellman and Spelke didn't see

[8] The closest experiment in Spelke and Kellman (1983) to the one described by Prinz is the experiment we described above for testing the gestalt hypothesis. When we described the experiment, we only presented the experimental condition, but there were two control conditions, one of which was similar to the condition Prinz describes. In the control conditions, participants were presented in the habituation trials with either a stationary complete rod in front of a block (rather than behind it and partly occluded by it) or two stationary smaller rods similarly presented in front of the block. This last condition is the closest to the condition Prinz describes. The purpose of the control conditions was to

this condition as a test to rule out Prinz's hypothesis. This makes sense, in any case, since this condition would be a particularly poor way to test Prinz's hypothesis. Once infants are habituated to a stimulus, they are less likely to respond to the same stimulus precisely because they start to tune it out—that is the very principle behind research using the habituation method. So the condition that Prinz focuses on is neither particularly challenging for his hypothesis nor representative of the broad set of conditions in Kellman and Spelke's paper. The more general problem with Prinz's discussion, however, is the *methodological* one that we have outlined above—that Prinz is wrong to focus on the question of whether his hypothesis is consistent with a single finding, especially when there are numerous conditions he doesn't mention that strongly argue against his hypothesis when their results are taken collectively.

Before we close our discussion of methodological empiricism, we will briefly address three further criticisms that Haith and Prinz have of rationalist research on object representation: (1) further evidence argues against rationalist accounts of object representation, (2) mere exposure to objects makes it easy to acquire object representations, and (3) rationalists deny that physical knowledge develops gradually.

(1) Does further evidence argue against rationalist accounts of object representation? Haith's and Prinz's appraisals of rationalist views of object representation focus on Baillargeon's drawbridge study and on Kellman and Spelke's partial occlusion study—experiments with 3.5- to 4.5-month-olds. While Haith and Prinz claim that this work is unable to show that infants at this age represent hidden or partly occluded objects, Prinz (2005, 2012) also employs another line of criticism, one that accepts for the sake of argument that infants at this age do have the abilities that Baillargeon (1987) and Kellman and Spelke (1983) say they have. Even so, Prinz maintains, infants at this age have had more than enough time to have acquired knowledge of object occlusion through general-purpose learning, and so there is no reason to suppose any of it is innate.

In developing this further line of criticism, Prinz recognizes that there is research that suggests that this ability may appear at younger ages—perhaps as young as 2 months old—but citing work by Slater et al. (1990), he claims that newborns haven't been shown to represent the unity of partly occluded objects. Hence there is a gap of at least two months in which this learning could take place. Prinz also points to the fact that while the older infants in Baillargeon's study (4.5-month-olds) looked longer at the unexpected event, the results for infants just one month younger were mixed: whether the 3.5-month-olds looked

serve as a comparison with the experimental condition. As we noted earlier, in the experimental condition, the infants looked equally at the single complete rod and the two-rod displays. In the controls, unsurprisingly, the infants looked longer at whichever stimuli they hadn't seen in the habituation trials.

longer at the unexpected event turned on whether they readily became bored during the familiarization trials. For Prinz, this finding constitutes evidence "against innateness" and suggests "that the principle of persistence is learned between the third and fourth month of life" (2012, p. 93). Finally, Prinz adds that infants fail to reach for hidden objects until they are considerably older than the "top performers in Baillargeon's study" (2012, p. 93). In particular, there is a period in development when infants can reach for an object but fail to do so if the same object is placed under a cover or behind a screen. It isn't until they are 9 months old that infants reliably reach for objects in this type of situation. Prinz's claim is that this mismatch between younger infants' looking behaviour and their reaching behaviour is grounds for supposing that they aren't truly representing the existence of hidden objects.

In our view, these criticisms carry little weight. Take the claim that newborns can't represent the unity of a partly occluded object. As we saw in Chapter 10, performance factors need to be addressed, especially newborn infants' difficulties in perceiving motion. Newborns have to be in a position to detect the motion of the two exposed rod pieces so that their synchronous movement can be registered. (If you can't see that the two pieces are moving together, then there is no reason to suppose that they form the ends of a single connected rod.) When this essential information is highlighted using stroboscopic motion, newborns behave in the same way as Kellman and Spelke's 4.5-month-olds (Valenza and Bulf 2011).[9]

What about the mixed results with the 3.5-month-olds (in contrast with the 4.5-month-olds) in Baillargeon's drawbridge study? Though the fact that 3.5-month-olds failed to look longer at the unexpected event when they were slow to become bored during the familiarization trials might seem problematic, it doesn't pose a significant difficulty for rationalism. This is because, as Baillargeon argues, there is a perfectly plausible account in terms of performance factors that can explain why younger infants failed to look longer at the unexpected event despite their ability to represent occluded objects. The crux of Baillargeon's suggestion is that the point of familiarization trials in a violation of expectation experiment isn't to induce boredom; it is to prepare infants so that they are in a position to appreciate the test events.[10] Thus the likely explanation of the fact that these infants didn't quickly become bored in the familiarization trials is that they hadn't fully processed what was happening in these complex events. Notice that if this is right they wouldn't have been prepared to fully take in the relevant

[9] Recall as well that the argument from animals offers additional support for an innate system of object representation. For example, newborn chicks have been found to represent the unity of a centre-occluded object when it is the very first object that they see (Regolin and Vallortigara 1995; Lea et al. 1996).

[10] Again, the violation of expectation method isn't the same as the habituation method (see footnote 6).

aspects of the test events. But then there would be no reason they *should* be surprised by (and hence look longer at) the unexpected outcome.

Finally, there is Prinz's claim that younger infants are unlikely to be able to represent the existence of hidden objects given that it isn't until much later in development that they reach for hidden objects. In fact, Prinz understates the mismatch between the way that infants look at physically unexpected events and their reaching behaviour. Since Piaget's (1954) seminal work on the development of the representation of physical objects, it has been widely thought that infants progress through a series of critical stages in being able to represent hidden objects, and that much of this is only manifest in infants' increased success in reaching for and retrieving objects that can no longer be seen. One of these developmental milestones is the one mentioned by Prinz—in which infants only start to reach for hidden objects when they are 9 months old. But for Piaget, even this accomplishment doesn't show that infants understand that hidden objects have an objective existence. One sign that they don't is that they have to overcome the so-called A-not-B error. To induce an A-not-B error, an experimenter visibly hides an interesting object in one location (location A) (e.g., placing a toy under a cover) and lets the infant reach for it, repeating this sequence a number of times. Then the experimenter visibly places the object in a new location (location B). What Piaget found is that infants less than 1 year old continue to reach for location A even though they can see right in front of them that the object has been placed in location B. So there is evidence of a serious discrepancy between rationalists' claims that 4-month-old's looking-time behaviour shows that they represent hidden objects and the failure of even 12-month-olds to reach for hidden objects.

What should we make of the mismatch between very young infants looking longer at physically unexpected events and the surprising failures of significantly older infants to correctly reach for objects of interest? Does this undermine the idea that younger infants are able to represent hidden objects? No it doesn't. Once again, performance factors have to be taken into account.

Consider first the fact, highlighted by Prinz, that infants only start to reach for hidden objects when they are 9 months old. Notice that when a cover is placed over an object, a more complicated action plan is needed to retrieve the object than when there isn't a cover. To determine if infants who fail to reach for objects can nonetheless represent the correct location of the hidden object, we need to determine whether the failure is due to competence or performance issues. One way to determine this is to see whether they reach for hidden objects in conditions with reduced task demands. For example, we could look to see whether young infants reach for the object in a condition where the reason that the object is hidden is simply that it is no longer visible because the lights have been turned out. Under these circumstances, the object is hidden, but there is no additional object (the cover) that is introduced in order to hide the original object. In this

simpler hidden-object condition, it turns out that even 5-month-olds will reach for it (Hood and Willatts 1986).[11]

What about the A-not-B error? Why do infants search for an object in an old hiding location (location A), when they have just seen it placed in a new hiding location (location B)? Psychologists have offered a number of possible explanations for this behaviour. Performance factors again seem to play an important role. One important factor has to do with younger infants' memory limitations and their inability to inhibit a response (Diamond et al. 1994). However, given the huge impact of the A-not-B error on empiricist theories of object representation, we think it is particularly significant that further rationalist research suggests another factor may be at the centre of the phenomenon. The search failures in this case may not be owing to infants having an impoverished understanding of the situation that they find themselves in but rather to their having an unexpectedly sophisticated understanding of it.

Notice that the standard A-not-B experiment isn't just comprised of an object appearing in one location for a number of trials and then subsequently appearing in a different location. It occurs in a rich social setting, with the experimenter making eye contact with the infant, purposefully engaging the infant's attention, and purposefully placing the object repeatedly in location A. Topál et al. (2008) reasoned that these features of the situation may serve as signals to the infant that the experimenter is communicating generalizable information about the object's hiding place—something along the lines that *objects like this belong here (in location A) and that this where they can be found*. If infants understand the situation in these terms, then even when they see the object placed in a new location B (rather than the usual location A), they would still have reason to look for it in location A because that is where it ought to be based on the information conveyed by the experimenter.

To test this hypothesis, Topál et al. conducted several variations on the A-not-B experiment. In one variation (the communicative version), as the experimenter repeatedly placed the object in location A and then in location B, she engaged in the usual communicative interactions with the infant (eye contact, calling the infant by name, etc.). In another variation (the non-communicative version), as the experimenter repeatedly placed the object in location A and then in location B, she did not engage in any communicative interactions with the infant (so no eye contact, no calling the infant by name, etc.). The difference was striking.

[11] See also Baillargeon et al. (1990) for evidence that infants of the same age fail to reach for occluded objects not because they are unable to represent their existence but because of difficulties with actions that depend on certain types of problem solving. Interestingly, both infants' and adults' predictive reaching for moving objects is worse when the object is temporarily hidden by an occluder than when it is temporarily hidden by darkness—occlusion seems to be cognitively more difficult even for adults, where there is no question that the occluded object is represented as continuing to exist (Hespos et al. 2009).

Infants in the communicative condition conformed to the usual pattern, with a strong tendency to reach for location B, while infants in the non-communicative condition reached towards location A and location B equally. This suggests that, along with performance factors involving memory and processing load, expectations driven by pedagogical communication play a large role in explaining infants' failure to reach for desirable objects in A-not-B tasks in spite of their representing the objects as continuing to exist.[12]

(2) Does mere exposure to objects makes it easy to acquire object representations? At times, Prinz seems to suggest that learning about the existence and properties of ordinary objects can't be particularly difficult given that infants are situated in the physical world and consequently have abundant experience with physical objects:

> How does an infant know that a potato continues to exist when it is briefly taken out of view? Must that knowledge be innate? Not necessarily. From the very start of life, infants who can see experience numerous occasions when an object they are watching disappears from view. In fact, this happens every single time they blink… Hypothetically, infants could form the belief that the world disappears with each blink, but that's a pretty sophisticated inference that requires concepts of inexistence…each time an object passes behind another or gets engulfed in a shadow or occluded by an infant's own hands, it reappears a moment later, so the expectation of persistence is reinforced… The point is that infants have ample opportunity to discover that objects persist. Their expectations are informed by experience of a world in which objects rarely flicker out of existence. (2012, pp. 110–111)

Unfortunately, these remarks greatly underestimate the difficulty of the task that infants face on the empiricist assumption that everything they come to know about objects has to be learned from scratch. Seeing why this is the case can be illuminating. The problem, in brief, is that having experience with objects isn't the same thing as experiencing them *as objects*. Prinz claims that infants can learn about occlusion simply by seeing that an object that disappears behind another

[12] This research is part of a larger research programme that argues for the existence of a rationalist learning mechanism for acquiring certain types of socially transmitted information, which is paired with a reciprocal mechanism that disposes "teachers" to provide cues in the form of ostensive signals, which activate this form of learning. The learning mechanism involved in this system of reciprocal mechanisms, which are already active in infants and young children, disposes learners to be receptive to cues that a teacher is providing generalizable information about a kind—for example, not just information about how *this* bottle can be opened (that it can be opened in such and such way) but information about how *anything of this type* can be opened (that things of this type are to be opened this way). This system—which is referred to as a system for *natural pedagogy*—ensures that children respond to ostensive signals (such as pointing and eye gaze), that they have referential expectations associated with these signals, and that they interpret such signals as efforts to convey generalizable information about a kind (Csibra and Gergeley 2009, 2011).

object often reappears shortly thereafter (and that the alternative requires explicitly representing inexistence). But if infants don't already have the means to represent objects as such, why would they take the entity that disappears and the entity that appears to be one and the same? And why would they represent this one entity as moving *behind* the other as opposed to (e.g.) simply no longer representing it as being there? (Notice also that ceasing to represent it doesn't require representing inexistence.)

The problem here can be difficult to appreciate. Essentially the empiricist is committed to infants initially representing the world in *alien terms*—not as adults do (i.e., as a world populated by physical objects), but in some other way. What other way? Prinz himself doesn't have much to say about this. But other empiricists who have taken this question more seriously can help us see just how bizarre an infants' perceptual world might be following empiricist strictures. For example, Piaget held that infants start out with no idea whatsoever that there is anything in their experience beyond fleeting, disjointed images that come into being as a result of their own actions:

> During the first two stages (those of reflexes and the earliest habits), the infantile universe is formed of pictures that can be recognized but that have no substantial permanence or spatial organization. (Piaget 1954, p. 4)
>
> failing to locate himself at the outset in space, and to conceive an absolute relativity between the movements of the external world and his own, the child at first does not know how to construct either groups or objects and may well consider the changes in his image of the world as being simultaneously real and constantly created by his own actions. (Piaget 1954, p. 7)

And Quine maintained that young children represent the world without distinguishing between individuals, stuffs (e.g., water), and properties:

> We in our maturity have come to look upon the child's mother as an integral body who, in an irregular closed orbit, revisits the child from time to time; and to look upon red in a radically different way, viz., as scattered about…Water, for us, is rather like red, but not quite; things are red, stuff alone is water. But the mother, red, and water are for the infant all of a type; each is just a history of sporadic encounter, a scattered portion of what goes on. His first learning of the three words is uniformly a matter of learning how much of what goes on about him counts as the mother, or, as red, or as water. It is not for the child to say in the first case 'Hello! mama again', in the second case 'Hello! another red thing', and in the third case 'Hello! more water'. They are all on a par: Hello! more mama, more red, more water.
>
> The child can learn 'mama', 'red', and 'water' quite well before he ever has mastered the ins and outs of our adult conceptual scheme of mobile enduring physical objects, identical from time to time and place to place. (Quine 1960, p. 92)

It is not at all easy to imagine what it would be like to conceptualize the world in any of these ways. Nor is it easy to see how anyone would come to conceptualize the world as it is ordinarily understood, as populated by physical objects, given such a starting point.[13] In his criticism of rationalist theories of the origins of physical knowledge, Prinz underestimates the difficulty of addressing these challenges. By supposing that infants can already see objects disappearing behind one another and reappearing, Prinz is effectively attributing to them the very sort of representations that his anti-rationalist stance says they have to learn. In contrast, if empiricist restrictions are maintained, then the explanatory project is far more difficult—and far more interesting (assuming it can be pulled off). This is because the transition from infancy to later childhood would be tantamount to the adoption of a new ontological framework, a radically different way of looking at the world.

(3) Do rationalists deny that physical knowledge develops gradually? Finally, let's consider the charge that rationalists mistakenly claim that there is nothing for infants to learn about physical objects since a rationalist model would simply postulate that the full adult competence is innate. The accusation of "dichotomous" thinking about cognitive capacities is a major theme in Haith's methodological discussion of rationalist research on infant cognition. He rebukes rationalists for taking "indications of the earliest fragments of a concept as evidence for virtual mastery of the concept" (Haith 1998, p. 168) and charges rationalists with ignoring and obscuring conceptual development:

> the use of the looking paradigm for cognitive inquiry has encouraged dichotomous answers to cognitive questions also. Can infants do arithmetic? Do infants perceive causality? Do infants appreciate continuity, cohesion, and inertia in object motion? (p. 171)
>
> the need for a full-scale developmental model of representation that incorporates the notion of partial accomplishments is obvious. (p. 175)

This is a theme that is echoed by many other critics of rationalism, both in philosophy and in cognitive science. For example, Cottingham (1988) highlights the view that rationalism is widely seen in philosophy as being anti-learning and anti-developmental:

> Perhaps the greatest source of unease which most people nowadays feel on being confronted with the theory of innate ideas is that it does not seem to do justice to the way in which human beings appear to acquire knowledge via a gradual process of learning. (p. 71)

[13] The difficulty here is effectively a variant of the problem of initial representational access discussed in Chapter 12.

And Thelen and Smith (1994) accuse rationalist psychologists of being anti-development:

> Despite the seemingly hidden competencies revealed in special tasks and amid the continuities across development, children do develop; babies and adults are not the same. The fact of development is not explained by a list of innate ideas. (p. 30)

Similar sentiments can be found in Munakata et al. (1997); Karmiloff-Smith (1998); and Cohen et al. (2002); among many others.

Earlier we pointed out how unfounded this charge is when directed at rationalism as a general framework for studying the mind (see Chapter 4). In fact, one of the key morals that we have been at pains to emphasize throughout this book is that rationalism is in no way incompatible—or even in any tension at all—with learning and development. Rationalism just denies that *domain-general* learning mechanisms are more or less the whole story regarding the acquisition of our cognitive and conceptual capacities; it holds that rationalist learning mechanisms are needed too.

The representation of physical objects is no different in this regard. In Chapter 15 we saw that rationalists have systematically investigated how infants' understanding of physical objects changes on a month-by-month basis and how these changes argue for an approach involving rationalist learning mechanisms. This painstaking work has revealed a wealth of highly surprising data, including the quirky finding that infants draw appropriate inferences about one type of physical interaction months before they draw comparable inferences for events of another type that are virtually perceptually indistinguishable—for example, learning that a tall object should be seen when it is *behind* a small occluder before they learn that a tall object should be seen when it is *inside* a small container of the same size as the small occluder. We also saw that further rationalist research, which aims to uncover the structure of the domain-specific rationalist learning mechanisms involved in this learning, has successfully identified the conditions under which a developmental lag of this type can be effectively eliminated—where this can happen on the basis of as few as two teaching events with the right form. Far from ignoring development, rationalists have been at the forefront of documenting and explaining conceptual development in this domain.

In this part of the book (Part III), our aim is to examine some of the most important and influential contemporary empiricist views that stand in opposition to concept nativism. This chapter has focused on what we have been calling *methodological empiricism*—the widely held view that the rationalism-empiricism debate can largely be settled in favour of empiricism on methodological grounds. Empiricists who subscribe to methodological empiricism see empiricism as having a special status of being the default view regarding the origins of concepts.

Although it is often said that empiricism has this special status because it is the more parsimonious theoretical framework, we have seen that empiricist and rationalist theories may both be parsimonious (in different ways), and that a general evaluation of how parsimoniously they explain the development of a cognitive capacity cannot be made in advance of enquiry but requires the examination of worked out theories. Proponents of methodological empiricism have also built their case against concept nativism by proposing alternative empiricist explanations of some famous experimental results that have been taken to support concept nativism, claiming that their opponents have hastily drawn rationalist conclusions overlooking these alternatives. The problem with this claim, however, is that the proposed empiricist explanations tend to flout Spelke's principle that proposed explanations should not be "based on a small subset of findings" and ignore contrary results (1998, pp. 191–192)—a principle that would be taken as too obvious to even require stating in the context of another science, like chemistry or biology. Moreover, as we have seen, the proposed empiricist explanations are often *not* ones that rationalists have overlooked, but ones they have anticipated and refuted, sometimes in the very same studies that these empiricists have taken as the focus of their critique. Finally, while empiricists have taken the example of object representation to illustrate how methodological considerations favour empiricism, the situation is exactly the opposite. The example of object representation demonstrates the need to take more seriously the way in which domain-specific rationalist learning mechanisms guide infants' learning about the physical world and illustrates the importance of providing a rich explanatory account of a wide range of data. Questions about conceptual development need to be settled on the basis of the explanatory merits of competing theoretical explanations, not on methodological grounds.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0017

18
Neo-Associationism

In this part of the book (Part III), we are taking a critical look at some of the most influential empiricist alternatives and objections to concept nativism. Accounts based on associative processes have been the cornerstone of empiricism throughout its long history. Indeed, it would hardly be an overstatement to say that associative processes are the quintessential empiricist tool for explaining cognitive and conceptual development and for countering rationalist accounts. The focus of this chapter is on what we will call *neo-associationist* accounts—contemporary empiricist views where processes of association are central to empiricist opposition to concept nativism.[1]

Contemporary empiricists' accounts make use of associative processes both in objecting to rationalist accounts and in developing empiricist alternatives. It will be useful to distinguish two forms of neo-associationism to reflect these two distinct roles for associative processes and consider the two forms of associationism separately. The first of these—which we will refer to as *deflationary neo-associationism*—is critical of the evidence that rationalists have cited for various proposed innate representations and rationalist learning mechanisms. Associative processes in this case are taken to show that rationalists' interpretations of their own data are misguided and that the data are better explained in terms of simple associations involving low-level perceptual properties being acquired via domain-general learning mechanisms. The second form of neo-associationism—which we will refer to as *constructive neo-associationism*—is more ambitious. Its goal is to provide an empiricist account of the origins of certain abstract concepts or conceptual abilities in terms of domain-general associative processes.

We maintain that concept nativism is *not* undermined by neo-associationism of either kind. Of course, a comprehensive assessment of neo-associationism would require examining numerous instances where associative processes have been claimed to support empiricism over rationalism. We cannot do that here. But what we can do is work through some representative concepts from a domain that is especially favourable for empiricist accounts. In particular, we will look at

[1] Processes of association have been central to empiricist theories of the mind going back at least as far as the British empiricists in the seventeenth and eighteenth century. We use the term *neo-associationism* to emphasize that our focus is on contemporary theorizing and the contemporary rationalism-empiricism debate.

concepts that are central to what is known as *sociomoral cognition*.[2] These concepts provide a particularly good test case for considering neo-associationist empiricist accounts, since the concepts involved are highly abstract and ones that are not ordinarily thought to be present in infants or very young children, but rather to be acquired later in childhood, and even then, only on the basis of extensive teaching or training. If neo-associationism (in either form) provides a robust challenge to rationalism, these are precisely the kinds of concepts one should expect it to apply to (see also the discussion of Heyes' deflationary neo-associationist account in Chapter 9). Examining these concepts involved in sociomoral cognition also has the advantage of allowing us to consider another conceptual domain at issue in the rationalism-empiricism debate about the origins of concepts.

We will begin with the deflationary form of neo-associationism, which claims to undermine concept nativism by showing that well-known developmental data have been overinterpreted, or mistakenly interpreted, by rationalists. But first we need to provide a bit of context. Not long ago, rationalist accounts of sociomoral cognition had to make do with a narrowly circumscribed body of developmental data for preschoolers and older children (see, e.g., Nichols 2004; Dwyer 2007; Mikhail 2007). However, recent years have seen an outpouring of important publications reporting that considerably younger children are sensitive to sociomoral matters, including many studies showing such sensitivity in preverbal infants. What's more, these newer studies (some of which were mentioned earlier in the book) have explored a much broader range of sociomoral categories. For example, preverbal infants have been found to engage in complex patterns of inference regarding who others will affiliate with (Hamlin and Wynn 2012; Hamlin et al. 2013; Powell and Spelke 2013, 2018); they take language and accent to be indicative of group membership and use this information to draw inferences about third-party social relations (Liberman et al. 2017); they represent and reason about stable social dominance relations (Mascaro and Csibra 2012; Gazes et al. 2017); they have expectations about the outcomes of interactions between novel types of agents based on the agents' relative sizes, their number of groupmates, and whether their groupmates are in a position to see the conflict (Thomsen et al. 2011; Pun et al. 2016, 2022); they expect that social allies will step in to aid one another in an intergroup conflict (Pun et al. 2021); they have a rudimentary understanding of resource transfers between social agents, one that distinguishes *giving* from instances of *taking* that involve the very same motions (Tatone et al. 2015); they expect distributions among agents to be fair (Buyukozer Dawkins et al. 2019); they expect a form of indirect reciprocity in which an agent's fair

[2] The term *sociomoral cognition* covers a range of represented categories and distinctions that are relevant to group living, whether or not the person thinking in these terms is directly impacted by the other people's actions. Sociomoral cognition includes preference formation and various forms of reasoning pertaining to in-group/out-group status, social rank, cooperation, affiliation, fairness, loyalty, punishment, and other related social phenomena.

distribution is later rewarded by others who are in a position to know about what happened (Meristo and Surian 2013); and when witnessing a bystander reward or punish an agent who distributes resources to others, they expect fair distributors to be rewarded and unfair ones to be punished (Geraci and Surian 2023). Collectively, this work—all with preverbal infants—makes an impressive case that an assortment of sociomoral concepts are present at a remarkably young age.

One of the earliest and most important papers in this literature is Hamlin et al. (2007), which showed that 6-month-old infants socially evaluate and form preferences about individuals based on how these individuals have interacted with other agents that the infants don't know—that is, the infants are bystanders observing third-party interactions.[3] Infants in this study first saw an agent (a wooden block with large googly eyes) move in such a way that it appeared to be repeatedly attempting to climb a steep hill. Then two other agents—also blocks with large googly eyes—alternately interacted with this "Climber". The "Hinderer" appeared to oppose the Climber's efforts, knocking into it in such a way that the Climber ended up rolling down to the bottom of the hill; the "Helper" appeared to aid the Climber by nudging it up the hill, after which the Climber did a little celebratory dance by bouncing up and down. Adults readily interpret the Helper to be a prosocial agent who intentionally assists the Climber in achieving its goal, and the Hinderer to be an anti-social agent who intentionally thwarts the Climber. Six-month-old infants appear to share this interpretation. The infants preferred the Helper over the Hinderer when given a choice between them.

But why think that the infants were manifesting a distinctively *social* preference? Maybe the infants were responding to the patterns of motion associated with the Helper and Hinderer and not to the social significance of their actions. Hamlin et al. took this possibility into account by running another version of the experiment that eliminated the cues of agency—duplicating the motions of the Helper and Hinderer in a context in which the Climber had no eyes and no signs of self-propulsion. In this case, the infants showed no preference between the Helper and the Hinderer. Taken together, these results strongly suggest that infants possess genuinely social preferences at the age of 6 months. What's more, in related work that uses a preferential looking measure (see Chapter 9 for explanation), Hamlin and her colleagues have shown that infants as young as 3 months old also have a preference for the Helper over the Hinderer and, moreover, this preference is manifest across diverse types of scenarios, where the agents have different goals and intentions, and where what counts as helping or hindering involves highly varied physical movements (Hamlin and Wynn 2011).

Other research has shown that infants' social preferences are surprisingly subtle. For example, 8- to 10-month-olds are sensitive to intentions versus outcomes,

[3] Hamlin et al.'s study builds on earlier, groundbreaking work by Kuhlmeier et al. (2003), which investigated 12-month-olds' responses to third-party interactions using a similar hill-climbing scenario.

and to intentional harms versus accidental harms (Hamlin 2013; Woo et al. 2017). And while infants as young as 5 months old prefer Helpers who assist prosocial characters, they don't *always* prefer Helpers. In fact, they prefer a Hinderer who hinders a character that was previously seen to be anti-social to some other third party. In other words, to a first approximation, they seem to be thinking in terms of the somewhat Machiavellian principle *an enemy of my enemy is my friend* (Hamlin et al. 2011; Hamlin 2014).

As with the work on objects in the previous chapter, empiricists have been sceptical that infants are representing the world in such complex, abstract ways, drawing inferences about goals, intentions, helping, hindering, and so forth. One common empiricist response has been to argue that infants are only responding to simpler low-level properties of the stimuli. For example, Scarf et al. (2012) point out that in the original Hamlin et al. (2007) study, the Climber bounces up and down after the Helper nudges it to the top of the hill (the celebratory dance), and that it doesn't bounce after the Hinderer knocks it down to the bottom of the hill. Scarf et al. propose that infants have a positive attitude towards the bouncing motion and that their preference for the Helper over the Hinderer has nothing to do with helping or hindering or any other social factor. Instead, all we need here is "a simple association hypothesis" (p. 1). According to this hypothesis, infants like the bouncing motion, they come to associate this motion with the Helper because this bouncing motion only occurs in the Helper scenario (not in the Hinderer scenario), and this causes infants to prefer the Helper over the Hinderer. Infants needn't represent that the Helper is doing anything prosocial. In fact, they needn't represent the Helper as *doing* anything at all—they might represent it simply as a moving block and not even as an agent, much less a prosocial agent. On this view, all that is happening is that infants are learning to prefer the Helper by coming to associate it with a salient positive event—the bouncing. They simply like bouncing, which leads them to like things that are broadly associated with bouncing events (in this case, other blocks).

What we have, then, is a nice example of the deflationary neo-associationist strategy. If Scarf et al.'s hypothesis is right, Hamlin and colleagues' data don't support a rationalist interpretation after all. Moreover, Scarf et al.'s proposal isn't merely a possibility to consider. They have tested it with an experiment that was designed to dissociate the bouncing from the helping. This experiment used materials based on Hamlin et al.'s work (blocks with eyes) but with three distinct conditions—bouncing occurring with helping, bouncing occurring with hindering, and bouncing occurring with both helping and hindering. They found that infants preferred the Helper in the first of these, the Hinder in the second, and showed no preference in the third, suggesting that, as Scarf et al. predicted, infants only care about the association with bouncing.

Does this mean that infants aren't attuned to the social structure of these interactions? Are they really only attending to something as simple as a bouncing motion and the entities it is associated with? No. As Hamlin et al. (2012) have

pointed out, Scarf et al.'s experimental setup altered crucial features of the stimuli, making this an inadequate test for the simple association hypothesis. For example, in Hamlin et al.'s original experiments, the eyes were suggestive of goal-directed behaviour—the position of the pupils was fixed to give the impression that the Climber was looking towards the top of the hill. By contrast, in Scarf et al.'s experiments, the eyes were not fixed in an upward gaze. As a result, as the Climber went up the hill, the pull of gravity caused the "pupils" in its googly eyes to drop, giving it the appearance of looking in the wrong direction.

To assess the relevance of these differences, Hamlin (2015a) ran two further experiments. The first experiment compared two variants of the helping scenario (where upon reaching the top of the hill, the Climber bounces up and down): one variant used the original stimuli from Hamlin et al.'s study and the other used Scarf et al.'s stimuli. The result was that Hamlin replicated the finding that infants prefer the Helper but *only* for Hamlin et al.'s stimuli—when shown Scarf et al.'s stimuli, infants showed no preference. Hamlin's second further experiment examined how infants' preferences are affected by bouncing stimuli. In one condition, infants saw helping/hindering events without bouncing but with good cues that the Climber had the goal of reaching the top of the hill. In another condition, the Climber bounced after being helped to the top of the hill but lacked the goal-directed cues. Hamlin found that infants preferred the Helper in the first condition but not in the second. This strongly suggests that infants aren't merely responding to an association with a positive low-level physical feature (bouncing). Rather, they are interpreting these events in terms of their social significance. When it is clear that agents with goals are involved, infants have a clear preference for prosocial (helping) agents over anti-social (hindering) agents irrespective of any associated bouncing.[4]

[4] Hamlin's work on infant and toddler prosociality (and related rationalist research) has attracted a great deal of attention, both positive and critical. Some of these studies have replicated or extended Hamlin's rationalist findings (e.g., Scola et al. 2015; Woo and Spelke 2023), while others have failed to replicate Hamlin's original findings (e.g., Cowell and Decety 2015; Schlingloff et al. 2020). What should we make of these mixed results? Several points are worth bearing in mind in evaluating them. First, while it is important to take failures of replication seriously, as we noted in Chapter 9 a failed replication attempt does not simply invalidate the original finding. Second, individual findings should be seen in the context of a broader range of findings. In this regard, a meta-analysis examining both published and unpublished work in this area concluded that, taken as a whole, it supports the view that infants have a preference for prosocial agents. This was true even after correcting for possible publication bias, though it was concluded that the effect size is likely to be lower than that in published work (Margoni and Surian 2018). Third, the details of replication attempts matter. Even a small change to the stimuli or testing procedure can result in infants failing the task despite possessing the concepts at issue (as when Scarf et al.'s stimuli inadvertently eliminated cues that indicated the agent's goals in Hamlin et al.'s 2007 experiments). This is one of the key morals illustrated by the exchange between Hamlin and colleagues and Scarf and colleagues. Teasing out infants' underlying competence is especially challenging in work of this sort, which involves presenting infants and young children with complex scenarios. Even a highly faithful replication attempt may fail if it employs stimuli or procedures that subtly affect the infants' attention, memory, or processing of the events, leading infants to fail the task due to such performance variables masking their underlying competence. This may be what is involved in other failed attempts to replicate rationalist findings of prosociality in

In short, the deflationary neo-associationist strategy fails to account for how infants respond to helping/hindering scenarios. Recall also that we noted earlier that the helping/hindering studies are part of a much larger trend in recent developmental psychology. Many aspects of sociomoral cognition are now under investigation, with different research teams identifying an increasingly broad set of sociomoral representational capacities in very young children—many in children less than 1 year old. Suppose we take this research at face value. What does this tell us about the way that infants represent the social world?

The studies by Hamlin and colleagues argue that infants as young as 3 months of age recognize the existence of agents with goals, represent interactions between agents in terms of whether they engage in prosocial or anti-social actions, and have specific preferences regarding who to interact with based on these agents' social interactions with third parties. This suggests that by this early age infants already possess a wide range of sociomoral concepts. Given the age of the children and circumstances that evince these types of concepts, this favours the rationalist view that the foundation of sociomoral cognition involves innate concepts or rationalist learning mechanisms. However, just as with the examples we touched on in our treatment of the argument from early development in Chapters 8 and 9, there are a number of questions this leaves open. One that is worth highlighting concerns the precise content of the infants' concepts, which may differ from that of related adult concepts. For example, it is certainly possible that the infants in the helping/hindering experiments are engaging in moral evaluations of the prosocial and anti-social agents, determining whether the agents act in accordance with a normative moral principle to be helpful (or to not be anti-social). Another possibility, however, is that the infants are engaging in an *affiliative social evaluation*, asking instead who it would be desirable for them to associate with—an evaluation grounded in self-interest.

The interpretation in terms of affiliative preferences is the one that Hamlin herself seems to favour. And it is supported by a further line of argument (Baillargeon et al. 2015). Consider the hill-climbing paradigm once more, in which infants are initially familiarized with the prosocial Helper and the anti-social Hinderer before they are asked to choose between the two. It turns out that although infants prefer the Helper when given a choice of which to associate with, they don't look longer at the anti-social actions than the prosocial actions when they see both types of scenario in the familiarization trials—suggesting that they are not at all surprised by either type of action. But they ought to be surprised by

infants, such as Schlingloff et al. (2020) (Hamlin, personal communication). Supporting this interpretation, Tan and Hamlin (2022) have found in a further study that the number of familiarization trials is a crucial factor, and that infants prefer the Helper if given more familiarization trials (allowing them to more fully process the scenarios in the test events). Relatedly, by using eye tracking, they also found that infants' preference for the Helper depends on their successfully extracting the Climber's goal (as indicated by whether they look to the top of the hill as the Climber is ascending the hill).

(and so look longer at) anti-social actions if they viewed such actions as counternormative.

There is evidence, then, that infants' tendency to choose prosocial over anti-social agents may initially be about their affiliative preferences and not a reflection of normative moral reasoning. At the same time, the current trend in developmental psychology, we think, suggests that there are numerous elements of sociomoral cognition involving innate concepts or rationalist learning mechanisms, including ones focused specifically on moral domains. For example, as we mentioned earlier, one line of investigation has examined expectations about fairness and has revealed that infants as young as 4 months old expect resource distributions to be fair (Buyukozer Dawkins et al. 2019).[5] Importantly, these are cases where the infants themselves aren't recipients of these resources; the infants merely observe distributions to other unknown individuals. In addition, slightly older children—but still under 2 years of age—seem to understand that a fair distribution isn't always an equal distribution. They take into account merit (in the form of the relative amount of work completed by the different potential recipients) (Sloane et al. 2012).

Interestingly, with only minimal cues to group membership, infants expect members of the same group, but not those of different groups, to provide support to one another (Jin and Baillargeon 2017). This suggests that there may be a deeper story to tell about why the infants in Hamlin and colleagues' work appeared not to expect prosocial behaviour, one having to do with the social context of the scenes they observed. Jin and Baillaregeon (2017) note that the Helper/Hinderer stimuli in this research did not include cues about the group membership of the characters. So it is entirely possible that young infants do expect prosocial behaviour and represent some form of harm norm after all—so long as this is tied to *in-group* interactions (for related findings, see Bian et al. 2018; Ting et al. 2019).

In considering deflationary neo-associationist accounts in the sociomoral domain, we have found no real support for such accounts. Instead, what the data suggest is that infants possess a rich set of representations underpinning a remarkably early initial understanding of prosociality, fairness, and other sociomoral categories. Of course, this is just one conceptual domain, and so it remains possible that in principle such accounts may fare better for other conceptual

[5] Earlier work had produced conflicting findings as to whether infants expect windfalls to be distributed equally. Some authors (Meristo et al. 2016) found that 10-month-old infants expected equal distributions, while others (Ziv and Summerville 2017) had found that slightly younger infants (9- and 6-month-olds) did not. But, apart from the age difference in the infants tested in these studies, there was another difference. In the study with older infants, the unequal distribution involved giving all the items to one individual and none to the other (two vs. zero); in the other study, both individuals got something, although one got more than the other (three vs. one). Buyukozer Dawkins et al. (2019) showed that age was not the crucial factor. Nine-month-olds—and even 4-month-olds—are sensitive to inequalities when the unequal distribution involves giving all the items to one individual and none to the other.

domains. But as we noted at the start of the chapter, the sociomoral domain provides a particularly strong test case for neo-associationist accounts. After all, this domain ought to be one of the most favourable domains for developing an empiricist neo-associationist account, since the conceptual capacities in question in this domain are of a kind that empiricists take to fall well outside the scope of those that infants' possess. If deflationary neo-associationist accounts are unsuccessful here, even for young infants, they are even less likely to succeed for less sophisticated conceptual domains. As a result, it is highly unlikely that deflationary neo-associationism can provide a robust general alternative to rationalism.

Let's turn now to the second form of neo-associationism—*constructive neo-associationism*—which aims to provide a domain-general associationist account of the origins of sociomoral concepts. In particular, constructive neo-associationism holds that children learn sociomoral norms through socialization, especially instruction that includes positive and negative reinforcement in response to behaviours that conform to, or violate, society's moral norms. Children are taken to represent the association of these positive and negative outcomes with various actions in a way that simultaneously shapes their behaviour and causes them to internalize these norms.

No doubt, many take positive and negative reinforcement to be the crucial factor in successful moral education. Among philosophers and cognitive scientists, Jesse Prinz provides an especially clear and intuitively compelling version of a neo-associationist view based on this idea. Prinz holds that there are *no* innate moral psychological structures: "Morality is a byproduct—accidental or invented—of faculties that evolved for other purposes" (2007, p. 368). He develops an account of moral norms in which moral norms are grounded in patterns of emotional reactions to norm transgressions.[6] The principal mechanism driving the acquisition of moral norms is supposed to be *emotional conditioning*:

> Emotional conditioning (the main method used in moral education) may allow us to construct behavioral norms from our innate stock of emotions. If caregivers punish their children for misdeeds, by physical threat or withdrawal of love, children will feel badly about doing those things in the future. Herein lie the seeds of remorse and guilt. (2007, p. 404)

Prinz adds that emotional conditioning affects not only behavioural rules but also "rules about what people should feel" (2007, p. 404), and notes that perspective taking may be important too.

[6] Prinz holds that "moral rules involve both self-directed emotions and other-directed emotions," that "our emotions must be directed at third parties" and not simply at those who transgress against ourselves personally, and that "mature moral judgments are enforced by meta-emotions" such as anger towards a transgressor who doesn't feel guilt (2007, p. 369).

He argues that cross-cultural facts support his approach, that there are functional explanations of what are purported to be universal moral norms, and that poverty of the stimulus arguments for rationalist accounts of cognitive development get little traction for moral norms. Regarding the cross-cultural facts, Prinz argues that cross-cultural variation undermines rationalist accounts here. For example, in discussing harm norms, he writes:

> Is there a universal prohibition against harm? The evidence is depressingly weak. Torture, war, spousal abuse, corporal punishment, belligerent games, painful initiations, and fighting are all extremely widespread. Tolerated harm is as common as its prohibition. There is also massive cultural variation in who can be harmed and when. Within our own geographic boundaries, subcultures disagree about whether capital punishment, spanking, and violent sports are permissible. Globally, every extreme can be found. In the Amazon, Yanamomo warriors engage in an endless cycle of raiding and revenge (Chagnon 1968). Among the Ilongot of Luzon, a boy was not considered a man until he took the head of an innocent person in the next village; when he returned home, women would greet him with a chorus of cheers (Rosaldo 1980). (Prinz 2007, p. 373)

Prinz's point is that this variability speaks against the existence of an innate harm norm since an innate harm norm would ensure universal agreement about what sorts of harms are prohibited. But even if some such a norm were universal—or if it were universal that every society adopts some harm norm or other—Prinz holds that these universals wouldn't require a rationalist explanation anyway. Rather, they could be explained in terms of their benefits to society as a whole. For example, Prinz remarks:

> Of course most cultures prohibit *some* harms, but there are non-nativist explanations for that. Such prohibitions are a precondition for social stability. (2007, p. 373)

Alternatively, universal harm norms might just be straightforward and inevitable effects of social living:

> We dislike it when our loved-ones are harmed. Human friendship promotes caring, which, in turn promotes the formation of rules that prohibit harm. Prohibitions against harm may be byproducts of the general positive regard we have for each other. (2007, p. 375)

In response to poverty of the stimulus arguments for rationalist accounts of cognitive development in this domain, Prinz argues that the information children receive about norms may not be so impoverished after all. Children may get

sufficient feedback from peers and adults, for example, feedback regarding the difference between violations of conventional norms and violations of distinctively moral norms:

> children get socialized into moral competence by observation of adults outside of the household and from social interactions with peers. A child who violates a conventional rule may be ridiculed by peers, but she is unlikely to incur worse than that (imagine a child who wears pajamas to school one day). A child who violates a moral rule, however, is likely to incur her peers' wrath (imagine a child who starts fights). In short, different kinds of misdeeds have different ramifications, and a child is surely cognizant of this. (2007, p. 393)

Prinz also cites anecdotal evidence in support of this claim:

> I was recently at a party with four one-and-a-half-year olds, and I made three casual observations: these children did not show remorse when they harmed each other; at such moments parents intervened with angry chastisement, social ostracism ("Sit in the corner"), and reparative demands ("Say you're sorry"); and parents never exhibited anger or punitive responses when children violated conventional norms, such as rules of etiquette. (2007, p. 393)

What should we make of these arguments for Prinz's neo-associationist view and against a rationalist treatment of moral norms? We will start with the argument that turns on the existence of cultural variability regarding harm. The first thing that needs to be said about this argument is that we agree that there is absolutely no question that there is a great deal of variability here—both across cultures and within cultures. Substantial variation of this kind has been confirmed by interviews, by cross-cultural experiments, by historical analyses, and other methods. Many of the findings can be surprising to those unfamiliar with the groups in question. For example, Shweder reports the following moral norm for a subculture in India:

> One male [Oriya Brahmin] informant put it this way: "If the wife touches her husband on the first day of her period, it is an offense equal to that of killing a guru. If she touches him on the second day, it is an offense equal to killing a Brahman. On the third day to touch him is like cutting off his penis. If she touches him on the fourth day it is like killing a child". (Shweder 1991, p. 262)

Direct comparisons of different societies, including numerous small-scale societies, have also identified considerable variation regarding how people respond to the ultimatum game, with mean offers varying from around 25% to over 50% across cultures, suggesting widespread variation in what is considered fair

(Henrich et al. 2005).[7] Or, to take another example, societies vary substantially in the degree to which they take intentionality into account in moral judgements, with some small-scale societies treating such things as taking someone's belongings accidentally vs. intentionally (or hitting someone in the face accidentally vs. intentionally) as effectively morally equivalent to one another (Barrett et al. 2016).

Additionally, there is much variation in development. Children in every society change in systematic ways as they learn the particular moral norms in their community. And, of course, individuals often come to change their views in personal or idiosyncratic ways as they think more deeply about such things as slavery, capital punishment, vivisection, or whatever happen to be the controversial moral issues in their community. There simply is no question that there is a great deal of variation in the moral norms that people endorse across cultures and at different times. No one should deny this.

But does all of this variability show that we need something along the lines of Prinz's constructivist neo-associationist account and that rationalist accounts of moral cognition can be rejected? Not at all. As we were at pains to point out in our discussion of the argument from universality, variation in and of itself is not incompatible with rationalist accounts of cognitive development. Since we addressed this issue at length in Chapter 11, we can be brief here.

First of all, it is worth remembering that concept nativism isn't about the existence of universal cognitive traits per se. Universality is relevant to the rationalism-empiricism debate only to the extent that patterns across cultures are best explained in rationalist or empiricist terms. In the case of moral cognition, a norm might be universally accepted without appearing universally in behaviour for a number of reasons that are perfectly consistent with an overall rationalist account of where the norm comes from. At the very least, we have to take into account such things as people's understanding of their own self-interest. Moral norms are not deterministic laws of nature. Sometimes, when a norm one holds conflicts with self-interest, self-interest can win out. Also, there may be competing moral principles pulling people in different directions. For example, while many people might consciously subscribe to moral norms with broad coverage, they may also have processes that are geared towards morally parochial ways of thinking—not only favouring their in-group but possibly adopting norms that purport only to apply to their in-group (Bernhard et al. 2006; Schmidt et al. 2012; Balliet et al. 2014; Jin and Baillargeon 2017; Bian et al. 2018; Ting et al. 2019).[8] How someone in this state of inner conflict behaves would turn on which system, in the moment, is in control. For these and other reasons, it should go without

[7] Recall that in the ultimatum game, one player proposes how to split a pot of money, and if the other player rejects the offer, neither gets anything.

[8] In instances where a group has been dehumanized, it may also be difficult to see that a moral norm that one subscribes to applies to individuals in that group (Haslam 2006).

saying that violating a moral norm in one's behaviour in no way entails failing to endorse the moral norm.

More generally, Prinz seems to be arguing against the existence of a harm norm that is as inclusive as possible—something along the lines *never harm anyone for any reason*. But any reasonable proposal for a universal harm norm would allow for a variety of behaviours where some element of pain or distress is to be expected, including many on Prinz's list (e.g., initiation rites and contact sports). Some forms of pain and distress might also be taken to be morally justified in light of a greater good (e.g., war and corporal punishment).

The biggest problem with Prinz's argument based on variability, however, is that his examples are all drawn from *adult* behaviour. But as we have emphasized throughout the book, concept nativism is a thesis about the acquisition base and does *not* entail that innate traits be rigidly fixed.[9] The version of rationalism appropriate to moral norms may well be one where an innate starting state is eventually replaced or overridden. Or it may be one where the acquisition base provides an innate domain-specific starting point that provides just the beginnings of what will eventually become a full-blown moral faculty. With either of these rationalist frameworks, relevantly different environmental circumstances would lead to differing outcomes.

Let's turn now to Prinz's argument that we should reject rationalist explanations of universal moral norms because the universality of such norms can be explained without the need to postulate any rationalist learning mechanisms. Prinz suggests that (if they exist) at least some universal moral norms could be accounted for in terms of their beneficial effects, for example, because they support social stability. There are two problems with this suggestion. The first problem is that, as we argued in the previous chapter, the mere existence of a possible empiricist alternative explanation does not give us grounds to reject rationalist accounts. The bigger problem, however, is that the fact that a norm (or anything else) is beneficial gives us no reason at all to suppose that its existence can or should be explained in domain-general empiricist terms. It merely suggests that there is some selection pressure to maintain such a norm once it arises. However, norms might be both beneficial and inexplicable in domain-general empiricist terms. So, pointing to beneficial effects of a norm doesn't actually provide a concrete alternative to a rationalist account of the norm, or even argue that some alternative or another of this type exists.

What about Prinz's suggestion that harm norms are a product of the high regard we have for our loved ones? The main problem with this suggestion is that it fails to explain why anything specifically moral should be acquired. Notice, for example, that many other species apart from humans have a positive regard for

[9] Prinz is not alone in suggesting that concept nativism, to its detriment, is committed to invariant moral norms. See, e.g., Sterelny (2010).

other members of their species, yet this does not lead to their having moral norms. At the very least, then, more would need to be said before this suggestion could be taken seriously as an explanation of the origin of harm norms.[10]

Finally, we have Prinz's claim that there is no hope for a poverty of the stimulus argument regarding moral norms. Recall his anecdote about the way the children and their parents behaved at that party—the children's lack of remorse and the parents' differing responses to transgressions of conventional and moral norms. Working from a very different theoretical perspective, Dwyer (2007) has drawn attention to contrary anecdotal evidence, remarking that "some parents get just as hot under the collar about conventional transgressions as they do about moral transgressions. (In some middle-class households, etiquette is taken very seriously.)" (p. 416). But we take it that anecdotal evidence of either type is fairly limited regarding what it can tell us. And beyond the anecdotes, Prinz's assertion that children receive relevant evidence to differentiate moral from conventional norms is just that—an assertion. For what it is worth, there is at least some experimental evidence showing that children themselves are deeply bothered by the violation of conventional rules (Rakoczy and Schmidt 2013).[11] Moreover, with the recent discovery regarding the very young age at which children possess sociomoral knowledge, it becomes less plausible that children have had experiences sufficient for them to learn this using only domain-general learning systems.

In short, Prinz's arguments for his empiricist emotional conditioning account carry little weight. But his emotional conditioning account also faces a number of serious challenges. Recall that Prinz's appeal to emotional conditioning holds that moral norms are acquired through social reinforcement as children experience or observe the consequences of moral transgressions—parents "punish their children for misdeeds, by physical threat or withdrawal of love" leading children to "feel badly about doing those things in the future" (2007, p. 404). We will discuss three considerations that challenge this approach and that speak to some of the difficulties facing neo-associationist models of moral norms more generally.

(i) Norms without reinforcement. The emotional conditioning account maintains that moral norms must be reinforced and that this is how they are acquired.

[10] It is also likely that regard for friends and loved ones is grounded in characteristically rationalist psychological structures. So even if this account is supposed to be an anti-rationalist account of moral norms, the deeper set of facts that it depends on may support a rationalist account of other aspects of social cognition.

[11] It is sometimes suggested that there is no clear psychological distinction between moral norms and conventional behavioural norms, and that this is a problem for rationalist views of moral cognition since it means that there is nothing specifically moral about the innate systems underlying norm acquisition. However, if there is no clear psychological distinction between moral norms and conventional behavioural norms—if there is only a more general normative domain—then this would not decide the issue in favour of empiricism. Rather, it would just shift the rationalism-empiricism debate regarding the origins of moral cognition to the question of what sorts of systems—rationalist or empiricist—underlie the acquisition of norms in this more inclusive normative domain. (For a sample rationalist account of the origins of norms in general, see Sripada and Stich 2006.)

But some moral norms seem to be in place prior to their being reinforced. We have already seen that infants expect third parties to conform to certain norms. Given how young the infants in question are—for example, 4-month-olds who expect fair distributions—it is highly unlikely that they themselves have been subject to reinforcement learning of the type that the emotional conditioning account requires. Children this young have a severely limited range of action, and parents and caretakers don't typically show anger at infants who are less than a year old for their failure to conform to moral norms. Nor do they attempt to instil guilt, disgust, or anger at moral transgressions (as Prinz's account requires) when it comes to children this young. Nor is it plausible that the infants could glean the norm through some form of verbal instruction, since infants at this age lack the needed linguistic skills.[12] And the acquisition of principles of fairness, or other moral norms, is unlikely to depend on learning based on the observation of siblings' behaviour and subsequent patterns of reinforcement. Many of the infants in these kinds of studies don't even have siblings to observe. For example, Hamlin (2015b) reports in relation to her work on infant understanding of helpers and hinderers that as many as 50% of the infants in her studies are firstborn children with no siblings at the time of testing.

(ii) Norms despite contrary reinforcement. The emotional conditioning account holds that moral norms are instilled through patterns of reinforcement. This might make sense in cases in which children's norms line up with their parents' and caretakers' norms—and hence with the patterns of reinforcement adults would provide. But while young children endorse some of the same moral norms as adults, this isn't always the case.

Take young children's views about ownership. Among young children, disputes frequently arise over who should be allowed to play with a given toy. If a toy belongs to one child but another child wants to play with it, parents typically say that the child who owns the toy should let the other child play with it. As it turns out, 3- to 7-year-old children disagree, taking the rights conferred by ownership more seriously than adults do. In one experiment, Neary and Friedman (2014) presented children with scenarios in which there is a dispute between third parties over the use of an item. The owner wants to use the item, but another child is using it and requires it to complete a task. When asked who should get to use the item, children overwhelmingly sided with the owner. By contrast, adults given the same scenarios overwhelmingly side with the child currently using the item. Given the enormous disparity between children's and adults' judgements in such

[12] Sterelny (2010) suggests that children could learn about fairness by "overhearing and taking part in discussions over how spoils are to be divided" (p. 291). But again, children's expectations about fair distributions are in place well before they can comprehend or engage in these sorts of conversations. Such conversations may well be relevant to the further development of how children think about fairness, but the evidence suggests that they can't explain how children acquire ideas about fairness in the first place.

scenarios—where children's judgements are robust over many years and directly contrary to those of adults—it seems very unlikely that the norm the children are endorsing is the product of reinforcement learning.

A similar dynamic is presumably at play regarding 17-month-old infants' view that only in-group individuals are required to help one another (Jin and Baillargeon 2017). In the experimental work showing this, infants were given only minimal information regarding group membership—some characters exclaimed *I'm a bem!*, while others exclaimed *I'm a tig!*. The individuals in question all appear to come from the same local culture in that they speak and dress in much the same way. It is implausible to suppose that parents and caretakers emotionally condition such young children to adopt moral norms where agents only help members of their own group. On the contrary, to the extent that infants of this age are encouraged to adopt helping norms at all, they are usually encouraged to adopt norms to help anyone who is in need of assistance.[13]

(iii) Limited explanatory potential of reinforcement. On Prinz's neo-associationist view, reinforcement is supposed to explain how children come to adopt particular moral norms. But even if reinforcement learning explained why children endorse the norms that they do, this would leave much about moral cognition still to be explained. For example, it leaves as an open question how children are able to formulate the norms they come to hold in the first place and how they are able to represent the situations that are relevant to evaluating and applying them.

Many moral norms—even in the simplified form that children grasp—inherently involve highly abstract ways of thinking which are not themselves explained by reinforcement learning. A moment ago we noted that young children have their own norms about ownership. But clearly, ownership isn't an *observable* property. There isn't a physical feature, like proximal contact, that connects instances of owners with the items they own, or that makes it possible to literally see an owner's entitlements. Likewise, fairness is more complex than just equal distribution. Among other things, fairness takes merit into account and only applies to distributions that are intentionally made by agents (not accidental or chance distributions), complexities that are understood from a very early age. Moreover, sociomoral norms routinely make reference to properties (e.g., *helping* or *hindering*) that are realized by highly diverse types of physical situations, which infants and young children spontaneously treat as instances of the same behavioural type. All of these further complications highlight the fact that it is very easy to underestimate the difficulty of explaining the origins of moral

[13] Infants of this age spontaneously engage in helping behaviour without parental encouragement (or even parental presence) (Warneken and Tomasello 2013) and even help others anonymously (Hepach et al. 2017). Surprisingly, rewarding children for helping others may be counterproductive. In one study, extrinsic rewards were actually found to undermine, rather than to promote, altruistic tendencies in slightly older children (Warneken and Tomasello 2014).

cognition—much as it is easy to underestimate the difficulty of explaining cognitive development more generally (see Chapter 5).[14]

Of course, none of this is to say that cross-cultural variation and developmental change don't occur. On the contrary, we are entirely in agreement with Prinz and other proponents of the emotional conditioning approach that cross-cultural variation and developmental change are extremely important features of sociomoral cognition that need to be taken into account. The norms involved in moral cognition certainly vary across cultures, and they can change and develop over time. As noted earlier, rationalism about the sociomoral domain isn't the view that moral norms are unalterable or that there is a fixed universal morality. What it claims instead is that sociomoral cognition is not entirely the product of domain-general learning processes. It develops partly on the basis of characteristically rationalist psychological structures in the acquisition base—such as innate concepts or innate domain-specific mechanisms pertaining to the sociomoral domain.

There is also no reason to suppose that the development of all aspects of sociomoral cognition is based on exactly the same kinds of characteristically rationalist psychological structures or that they are all learned in the same way or all follow the same developmental trajectory. This is true even if we focus relatively narrowly on the developmental origins of the representations of moral norms.

It could be that the innate basis for such norms is something along the lines we mentioned earlier regarding fairness, in which there is an initial innate principle that can be overridden (producing adult variation of the sort outlined in Henrich et al. 2005 that we mentioned in Chapter 11, despite there being universality earlier in development). However, there are many other possibilities that fall within the broad framework of rationalist accounts, including possibilities where the innate basis is more minimal. For example, one possibility is that there are no innate moral principles but just a very small number of innate moral concepts. Another is that there are rationalist learning mechanisms that are specific to learning moral norms and that have innate domain-specific biases that direct them towards certain types of moral norms. Another is that there is a single rationalist learning mechanism that is specific to learning moral norms of all

[14] Prinz is not alone in underestimating the difficulty of explaining the origins of moral cognition. Sterelny (2010), for example, claims that moral development is grounded in domain-general pattern recognition and that this explains why moral judgements are fast, automatic, and generalize to previously unobserved cases. But while moral development and moral judgement may have these features, such features don't distinguish learning based on domain-general pattern recognition from learning based on domain-specific pattern recognition. And Sterelny doesn't give any details about how domain-general pattern recognition would enable an individual who is incapable of moral judgement to become one who is. In discussing this transition, he remarks that "we convert that visceral reaction into a normative judgement" (p. 292), that "moral norms are grafted on top of our dispositions to respond emotionally" (p. 292), and that "moral cognition is a natural development of our existing emotional, intellectual and social repertoire" (p. 293). However, this leaves the crucial question of how the conversion, grafting, or natural process actually works or why it should be thought to be a domain-general process and not one that incorporates innate characteristically rationalist psychological structures (Joyce 2013).

types (or perhaps instead, norms in general) with no such biases. And so on. For each of these and many other related possibilities, it's also likely that moral norms undergo an extended process of refinement and elaboration in which domain-general processes play a part—perhaps even a very large part—in the full developmental story. The main point, though, is that they aren't the whole story.

Prinz's empiricist account of the moral domain is an especially clear and forthright example of the approach we have dubbed constructive neo-associationism. Nonetheless, the difficulties for his view are indicative of the challenges that any version of constructive neo-associationism faces. In order to address these challenges, such views will have to become more rationalist, incorporating more characteristically rationalist psychological structures into the acquisition base. More generally, although individual neo-associationist accounts pertaining to the sociomoral domain need to be considered on a case-by-case basis, there is little hope that neo-associationism in either form—deflationary or constructive—can provide a thoroughly non-rationalist alternative in this domain.

To sum up, there is no real support for neo-associationist accounts in the sociomoral domain. Neither the deflationary form of neo-associationism nor the constructive form of neo-associationism is particularly promising. Instead, there is strong support for characteristically rationalist psychological structures playing a role in the acquisition of a rich and diverse set of concepts in this broad domain. Of course, this is just one conceptual domain, and so in principle it remains possible that such accounts may fare better for other conceptual domains. But as we mentioned at the start of the chapter, the sociomoral domain provides a particularly strong test case for neo-associationist accounts. It ought to be one of the most favourable domains for developing an empiricist neo-associationist account, since the conceptual capacities in question in this domain are of a kind that empiricists take to be well outside the scope of those that infants' possess. Thus, if deflationary neo-associationist accounts are unsuccessful here, even for young infants, they are even less likely to succeed for many other conceptual domains. Likewise, if constructive neo-associationist accounts (such as emotional conditioning) are at best only part of the story about how moral norms develop and must build on a foundation grounded in domain-specific characteristically rationalist psychological structures in the acquisition base, this doesn't bode well for their prospects elsewhere. We conclude that there is no reason at all to suppose that neo-associationism can provide a plausible general alternative to the rationalist view of the origins of concepts we laid out in Part II.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0018

19

Artificial Neural Networks

From Connectionism to Deep Learning

Artificial neural network models of human cognition are enormously popular both in philosophy and in cognitive science. They are also widely assumed to vindicate an empiricist approach to the origins of concepts by showing that domain-general learning can account for the acquisition of concepts that rationalists have argued are innate or are acquired via rationalist learning mechanisms. In this chapter, we assess the bearing of research on artificial neural networks on the rationalism-empiricism debate by critically examining two important and representative types of empiricist proposals for how artificial neural networks might provide domain-general learning accounts of such concepts.

Our discussion will span approaches to neural network modelling from the *connectionist* or *parallel distributed processing* tradition and approaches that are associated with more recent developments in *deep learning* research in artificial intelligence (AI).[1] What both of these broad approaches have in common is the idea that psychological processes should be modelled as the spread of activation across a network of interconnected processing units. In these computational models, each processing unit must achieve a certain level of activation before it sends a signal to the units it is connected to further downstream in the network, and the amount of activation that is propagated between any two units depends on the weight that has been assigned to the connection between them. Learning consists in the adjustment of these connection weights across a large number of training cycles. A standard form of learning involves the network being given feedback regarding the output in each training cycle, which functions as an error signal about how well the network's output matches some target output. For example, the input could be patterns of activation corresponding to present tense irregular verbs ("bring", "ring", etc.) and the output, after numerous training cycles and adjustments, would be the network having settled into new patterns of activation corresponding to the past tense of these verbs ("brought", "rang", etc.).

[1] The term *connectionism* is sometimes used as a generic term for all approaches that model cognition using artificial neural networks. Other times it is used as a more restrictive term for the sorts of neural networks that predate deep learning research. We will follow the latter usage, using this term for the earlier forms of neural network modelling.

In that case, the network would be said to have learned a fragment of the English past tense system.

Let's turn now to how an artificial neural network might be thought to argue for an empiricist treatment of human concepts. We will begin by considering one of the most impressive and influential connectionist theories of human conceptual development—Timothy Rogers and James McClelland's account of the ontogeny of semantic memory (Rogers and McClelland 2004, 2008).

Semantic memory consists of a store of general information about all of the sundry categories that a person knows about, where the represented information plays a role in understanding their ability to recognize, interact with, and reason about items from these various categories.[2] The knowledge that babies cry, that pens are for writing, and that dogs wag their tails are all based on semantic memory. Likewise, inferences about categories and category instances are also based on semantic memory, for example, the inference that Fido will eat and grow given the knowledge that puppies eat and grow and the information that Fido is a puppy. Rogers and McClelland emphasize this feature of semantic memory in their account of how it works. They describe semantic memory as a system that allows people to draw inferences and make generalizations about properties of objects that can't be perceived in the moment. Maybe you have never seen a given coffee mug before. But owing to experience of other coffee mugs, you are in a position to infer that it may contain a hot beverage. Or maybe you have never cut into a mango before. But owing to what you've seen with other types of fruits, you might expect to find a pit inside.

Rogers and McClelland's proposal is that much of the acquisition and workings of human semantic memory can be explained in terms of a relatively simple neural network adapted from Rumelhart (1990) and Rumelhart and Todd (1993) (see Figure 19.1). In this model, the input layer is comprised of two banks of units, one that represents categories of perceived items (*pine*, *daisy*, *canary*, etc.) and one (labelled *Relation*) that represents information about general relations between categories (*X can do Y*, *X has property Y*, etc.). The output layer is a bank of units that represents various ways of completing propositions about items in represented contexts (e.g., if the input is *canary can…*, then the appropriate output would be the activation of GROW, MOVE, FLY, and SING). There are also two further layers of units—a hidden layer that stands between the input and the output, and the so-called *Representation layer*, which stands between item input and the hidden layer.

Given this setup, information flows forward through the network, modulated by the levels of activation in the units in the preceding layer and the weights on their connections. Because Rogers and McClelland's simulations of learning are

[2] In this way, semantic memory contrasts with *episodic memory*, which stores information about particular events in one's life (e.g., your dinner last night or your most recent birthday).

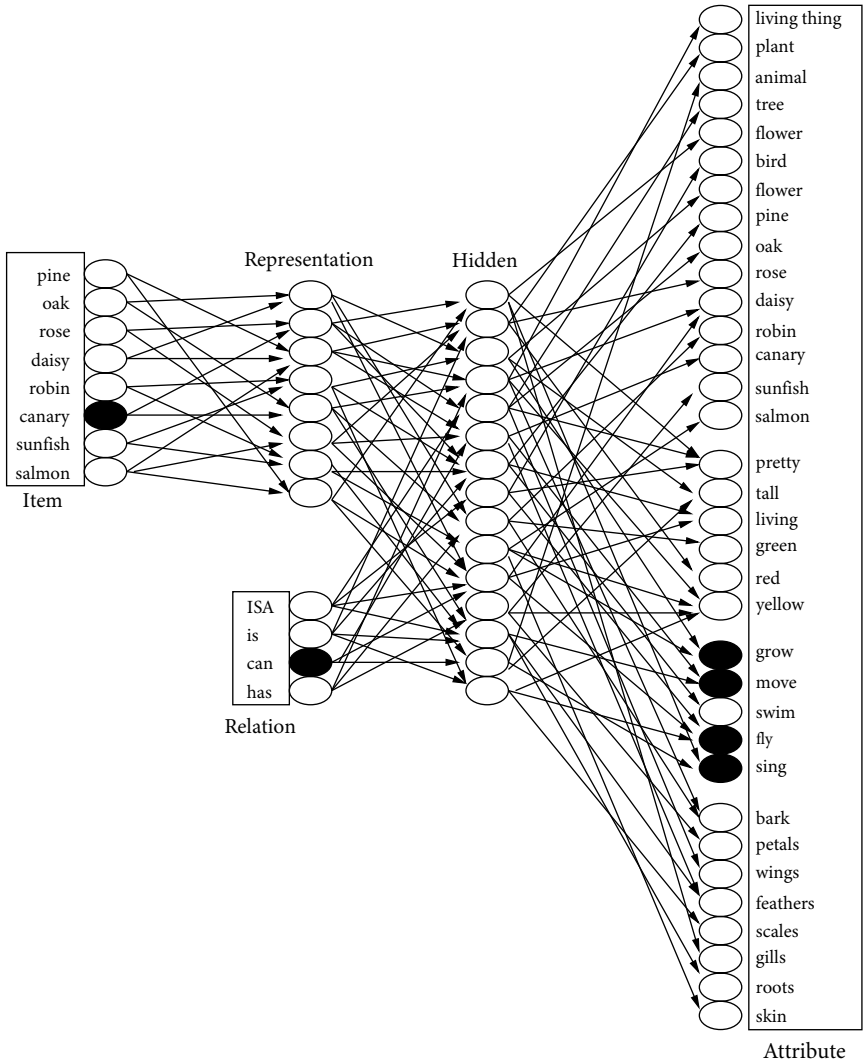

Figure 19.1 Rogers and McClelland's model of semantic memory. The example depicted is the model's representation of the query: What can canaries do? Previously the network was trained to produce the right output for each possible item/relation pair. The training consisted in comparing the output for each of these pairs to feedback given to the network as to whether the output was correct, each time leading to small adjustments to the connection weights throughout the network to better approximate the target output given in feedback. (From Rogers and McClelland 2004. Reproduced with permission.)

intended to demonstrate the power of a domain-general learning mechanism, the network's initial state is one in which, in essence, all of the connections are weighted equally (they are randomly assigned different small values). The process of learning is modelled by the network's training, in which each item/relation pair is activated, the resulting output activation is compared to feedback provided by the experimenters regarding the correct categories and relations, and small adjustments to the network's connection weights are made to diminish the discrepancy between its output and the feedback. According to Rogers and McClelland, such training over time causes the network to generally acquire a distribution of connection weights that reliably produces the right input-output pairings.[3]

In proposing this process as a model of learning, Rogers and McClelland suggest that the model's training is an abstract simulation of what happens as preverbal children form generalizations from encountered objects in particular situations. Take the situation in which a young learner sees a robin fly away when a cat approaches.[4] In this case, the child may make a prediction about what robins do in the presence of a cat and then update her future inferences based on how well her prediction matches the outcome. In the network simulation of this process, the prediction corresponds to the flow of information given input consisting of the activation of ROBIN and CAT APPROACHES, and learned changes to semantic memory correspond to the adjustments following a comparison of the resulting output with the feedback provided in this instance (the activation of FLY). As Rogers and McClelland remark in describing this process, "the environment provides both the input that characterizes a situation as well as the information about the outcome that then drives the process of learning" (2008, p. 694).

Rogers and McClelland don't deny the existence of domain-specific processing mechanisms. However, their model is intended to show that the principal features of semantic memory that are relevant to developmental psychology, including the appearance of domain-specific processing mechanisms, can all be accounted for by what are fundamentally domain-general learning mechanisms. The upshot of this work, they claim, is that "there is no clear reason at present to rely so heavily upon the invocation of initial domain-specific principles" (2008, p. 711). In other words, the logic of their argument is that we should reject rationalist views of the acquisition base because connectionist models show that innate domain-specific learning mechanisms are unnecessary.

Rogers and McClelland's central argument for their model turns on the claim that it can explain six phenomena that have been thought to characterize

[3] As a result of this training, the Representation layer also comes to represent the input items, with each type of input item producing its own unique profile in the Representation layer.

[4] This example goes beyond the representational resources of the partial model of semantic memory as it is depicted in Figure 19.1, so further units would need to be added to the model for the representation of cats, the relation of approaching, and so on.

semantic memory and its development.[5] The six phenomena they single out are: (1) the special role that causal knowledge has in semantic memory, (2) the privileged role of certain categories, which are deemed more natural and useful for purposes of generalization, (3) children's proneness to so-called illusory correlations, where they take exemplars of a given category to have properties they clearly lack, (4) the fact that the properties that children take to be more important vary depending on the type of category in question, (5) the fact that conceptual reorganization exhibits patterns akin to theory change in science, and (6) the fact that children acquire concepts for very broad categories before acquiring concepts for narrower categories and subcategories. For each of these phenomena, Rogers and McClelland present concrete simulations to show it can arise on their model through environmental stimulation and feedback. On this basis, they claim to offer a fully specified computational mechanism that explains the origin of the most basic facts about semantic memory. They argue that this is a major advantage over the rationalist accounts that they are opposed to in that, although these rationalist accounts are grounded in experimental findings, they generally have little to say about the computational operations that take place in semantic memory tasks; instead, they tend to stick to high-level descriptions of children's conceptual competence, or offer only the general outlines for a computational procedure.[6]

Although we won't be able to work through the way that Rogers and McClelland's model accounts for each of the six phenomena they cite, we will briefly consider their treatment of one of these phenomena in order to convey the strengths of their account and to see why their connectionist model might be thought to argue against concept nativism. We will focus on their account of the apparent developmental progression in which concepts for very broad categories are acquired before concepts for narrower categories and subcategories.

Mandler and McDonough (1993) argue for the existence of this developmental progression by examining young children's ability to distinguish different categories using an *object examination* task, a methodology that uses infants' spontaneous behaviour in manipulating sets of objects to draw inferences about which categories the infants are representing the objects as belonging to. In Mandler and McDonough's study, infants were given four objects from the same category to explore (e.g., four toy animals) and then subsequently given a fifth object from the same category or one from a contrasting category (e.g., a toy vehicle). Mandler

[5] We should point out that while many developmental psychologists would agree with this list of basic developmental phenomena to explain, there is room for debate about this list and whether it should constrain theories of conceptual development. For purposes of this discussion, however, we will follow Rogers and McClelland in supposing that this list may be used to evaluate the merits of competing computational models of the development of semantic memory.

[6] We broadly agree with Rogers and McClelland on this point: Rationalists *should* do more to address the issue of computational mechanisms underlying rationalist accounts. We would be delighted if our work inspired others to address this relative gap in rationalist theorizing.

and McDonough argued that if infants spend more time exploring the object from the contrasting category, this indicates that they see it as new and hence are distinguishing between the two categories. Many developmental psychologists have supposed that perceptual similarity initially organizes children's concepts. But Mandler and McDonough's results seem to suggest otherwise. For instance, they found that 9-month-olds distinguish animals from vehicles but not dogs from fish, even though the *animal* and *vehicle* categories were internally perceptually less similar than the *dog* and *fish* categories. Both the animals and vehicles included exemplars with very different shapes from other exemplars in the same category (e.g., an elephant and a bird among the animals, and a motorcycle and a train among the vehicles). In contrast, the exemplars of *dog* and *fish* were highly similar within their respective categories and dissimilar across categories (e.g., all of the dogs had four legs, heads, and ears, and all of the fish had fins). This and related work (Mandler et al. 1991; Pauen 2002) has been widely taken to show that children acquire a general concept for animals prior to acquiring concepts for particular types of animals.[7]

To show that their connectionist network can account for this type of pattern of progressive conceptual differentiation, Rogers and McClelland trained the network on the items in Figure 19.2 and examined the Representation layer at three

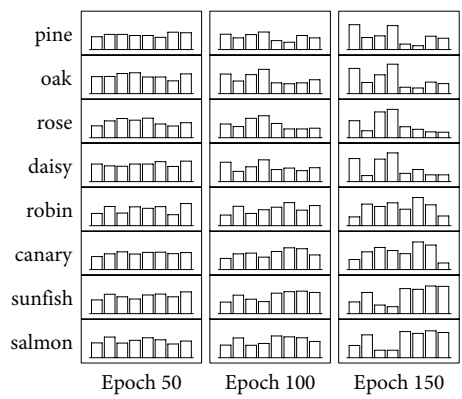

Figure 19.2 Rogers and McClelland's simulation of the pattern of conceptual differentiation in childhood in which broad categories appear before narrower categories. Each row depicts the activation levels in the Representation layer of the connectionist network for a given input item. Epochs consist in the presentation of every item-relation pair in the training set. After 50 epochs, there is little differentiation among categories. However, after 100 epochs, animals and plants are differentiated, and after 150 epochs, so are subcategories for animals and plants. (From Rogers and McClelland 2004. Reproduced with permission.)

[7] Mandler et al. (1991) caution against describing these more inclusive categories as superordinate categories and adopt the term *global categories* instead in order to make clear that children's earliest concepts needn't comprise a well-defined hierarchical system of classification.

points in the training—after 50, 100, and 150 epochs, where an *epoch* is a training sequence in which every item-relation pair in the training set is presented one time to the network. This means that in 50 epochs the network will have had fifty rounds of inputs of *pine*, *oak*, *rose*, *daisy*, *robin*, *canary*, *sunfish*, and *salmon* for each of the four represented relations (*isa*, *is*, *can*, *has*), with feedback on each round. The results can be seen in Figure 19.2. Notice that at epoch 50, the activation levels in the Representation layer of the network show little differentiation for the different inputs. But things change by the time the network has processed these item-relation pairs fifty more times (epoch 100) in that activation levels in the Representation layer for plants begin to differentiate from those for animals. And by the time the network has processed these item-relation pairs another fifty times (epoch 150), not only does the network differentiate plants and animals, but it distinguishes subcategories within these broader categories too (e.g., trees vs. flowers).

Some of Rogers and McClelland's critics have dismissed such findings on the grounds that the model isn't sufficiently realistic. And it's true that in many ways it isn't. Among other things, their training regimen depends on cycles in which every possible input item is presented to the network. That would be like a child never experiencing a pine tree without also experiencing a sunfish, and for this to happen over and over. Clearly, real children would have different levels of exposure to these categories, and the timing of these encounters would be far more variable. Notice also that the inputs already embody the distinctions between different categories (Snedeker 2008). This is to suppose that the child already has the input representations, for example, that she can differentiate the category *pine* from *oak* and can always tell when she is encountering a pine rather than an oak. The training regime also provides an unrealistic level of feedback to the child (Marcus and Keil 2008). Every time the network receives input during the training regimen (i.e., for every epoch), it is given comprehensive and accurate feedback regarding the network's predictions for each item-relation pair. That would be like a child being in a position to note in every encounter with a robin not only whether it moves and flies, but also whether it has properties that are less readily perceptible in a given viewing, such as whether it grows. Even for readily perceptible information, having accurate feedback about the correct output for every input is not realistic. Consider learning that birds sing. Since the training would indicate an error if SING isn't activated when the input is a bird, this would be tantamount to a child never seeing a bird that isn't singing.[8]

[8] Of course, it is possible that testimonial evidence could mitigate some of these difficulties, but testimonial evidence can only go so far since children aren't given comprehensive and accurate verbal feedback about every possible item-relation pairing either. Moreover, some of the developmental patterns that Rogers and McClelland aim to explain are present in prelinguistic infants and hence before they could take advantage of testimony.

However, despite these criticisms, Rogers and McClelland's simulation does demonstrate that at least in principle a domain-general connectionist network can gradually evolve in a way that mimics one of the six basic features of conceptual development that they set out to explain. And this suggests that artificial neural networks may play an important role when it comes to constructing an explicit computational model of conceptual development. But does it show that such a model must be—or ought to be—an *empiricist* model?

No, it doesn't. The key point to recognize is that there is nothing inherently empiricist about connectionist modelling. It is perfectly possible to advance *rationalist* connectionist models. This point has been understood since the early days of connectionist modelling even though it is rarely emphasized. For example, Rumelhart and McClelland's seminal *Parallel Distributed Processing*, which sparked an explosion of interest in connectionist modelling, clearly championed empiricist forms of connectionist modelling (McClelland et al. 1986; Rumelhart et al. 1986). Nonetheless, Rumelhart and MClelland were explicit that "PDP models, are, in and of themselves, quite agnostic about issues of nativism versus empiricism" (Rumelhart and McClelland 1986, p. 139).

What would a rationalist connectionist model of semantic memory look like? One possibility would be to adapt Rogers and McClelland's own model so that the network's initial state isn't restricted to random connection values. Instead, the network could start off with connection weights that are already set to values that predispose it towards certain categorical distinctions.[9] For example, the network's initial weightings for at least some categories could be set to something closer to those that appear at Rogers and McClelland's 100th epoch. In that case, the network would start out with patterns of activation that don't treat things like pine trees and sunfish indiscriminately. It would already be disposed to classify animals together and to classify plants together and to form different types of inferences and generalizations regarding these classifications.

Suppose that it is granted that we ought to be pursuing computational models of development broadly along the lines of Rogers and McClelland's model.[10] The question then becomes how to choose between empiricist versions and rationalist ones. As we read Rogers and McClelland, they claim that empiricist models are preferable on the grounds that, if a network can be trained to settle on activation

[9] Other potential rationalist modifications include a more selective arrangement of connections among the units and an arrangement in which the network appears within a larger rationalist architecture (e.g., receiving input from and providing input to other mechanisms that involve characteristically rationalist psychological structures).

[10] There are further complex issues about the viability of neural network modelling that we can't go into here, especially the need for models that implement or interface with systems of structured representations in which abstract operations can be defined over stored variables (Marcus 2001). These limitations have led Rogers and McClelland to postulate a second system that contributes to semantic memory, one that is capable of rapid learning based on minimal input (Rogers and McClelland 2008, p. 713).

patterns corresponding to the six developmental phenomena they have singled out for its evaluation, there is no need to postulate more structure. Its success in settling on the right patterns "calls into question the necessity of invoking initial domain-specific knowledge" (2008, p. 690).

Notice that this is a variant of methodological empiricism, which we discussed in Chapter 17. Rogers and McClelland are assuming that empiricism should be taken as the default view and that rationalism should be rejected if an empiricist model (in this case, an empiricist connectionist model) is possible. As before, however, the issue shouldn't be whether empiricist accounts are possible but which sort of view—empiricist or rationalist—provides the best overall explanation of conceptual development. Since both rationalism and empiricism are compatible not only with connectionism in general but also with the sorts of connectionist models that Rogers and McClelland offer, the issue should be whether empiricist or rationalist versions of these models are better at explaining the existing data. This means that we need to do more than ask if a connectionist model can in principle show how the classification of plants and animals (or other categories) emerges from a general-purpose system of representation. We also need to give full consideration to the sorts of evidence and arguments for concept nativism that were covered in Part II. To our mind, this broader set of considerations attests to the fact that, if a connectionist network is to be used to model semantic memory, then it ought to be a *rationalist* connectionist network, one that builds in certain innate categories and enough innate structure to accommodate a significant number of category-specific inferences.

Regarding the representation of animals, in particular, we have already seen that a case can be made for a number of mechanisms pertaining to animals that incorporate characteristically rationalist psychological structures. Earlier we highlighted the case for a rationalist learning mechanism for biological motion detection supporting the identification of animals (Chapter 10), for a rationalist learning mechanism that can interrupt current action in response to the presence of nearby animals (the animate monitoring hypothesis) (Chapter 14), and for category-specific rationalist learning mechanisms for acquiring socially mediated knowledge about dangerous animals (Chapter 14).

For another example of a phenomenon related to the representation of animals that is best explained in rationalist terms, consider the simple fact that animals are understood to have insides and that these insides are taken to matter to what an animal is and what it can do. How might an empiricist connectionist model explain where this understanding comes from? Presumably, if it were to emerge in preverbal children, this would require experiences of animals and their insides to guide the adjustment of its connection weights—for example, seeing a fish with a severed head or a fallen bird with a gash along its chest. Such experiences are not out of the question if we are talking about a learning process that continues through preschool and beyond; older children may well have built up the

relevant experience. But we can be fairly confident that the needed experience isn't reliably available to preverbal infants, at least in the dominant population surveyed in contemporary developmental psychology, namely, infants from urban communities in European and North American countries. They simply don't reliably have experiences of "open" animals and aren't reliably in a position to benefit from what might be said about animals' insides (that is, assuming adults have any proclivity to discuss this matter around them in the first place).[11]

For this reason, one way to test whether an empiricist connectionist network offers the right framework for modelling this important feature of children's understanding of animals is to focus on preverbal infants' expectations about which types of objects have insides. Setoh et al. (2013) approached this matter by examining 8-month-old's expectations regarding simple objects that are either animal-like by virtue of being both *self-propelled* and *agentive*, or not animal-like by virtue of lacking these key properties.[12] They found that infants look longer when a self-propelled agentive object is rotated and revealed to be hollow (i.e., to have no insides, like an empty open tin can) compared to when a similar self-propelled agentive object is rotated and revealed to be "closed" (like an unopened tin can). But infants don't look longer if the hollow object isn't both self-propelled and agentive. In other words, preverbal infants expect an object to have insides when and only when it is classified as animal-like in being a self-propelled agent.

What about the representation of plants? Notice that just as evolutionary considerations suggest that characteristically rationalist psychological structures are involved in representing and acquiring knowledge about animals, similar considerations extend to plants too. Plants were prevalent in human evolution and essential to the human diet and to other human activities (e.g., providing materials for construction). Moreover, many types of plants were also dangerous. The plants our ancestors lived among contained harmful, even deadly, chemical toxins, and often harboured mechanical defences, such as stinging hairs and thorns. Researchers who have highlighted the adaptive problems associated with these facts have begun to look for early plant-specific cognitive and behavioural dispositions in young children. What has been discovered so far is a rich set of domain-specific capacities that are unlikely to have been acquired on the basis of more fundamental domain-general processes.[13] For example, infants have been

[11] As was noted in Chapter 11, infants and very young children in these societies have little or no exposure to death. When they do encounter animal insides in the form of meat, this is typically presented in a highly antiseptic manner—to the point where children's later recognition that they are eating the insides of an animal often amounts to a shocking discovery.

[12] Setoh et al. explain that "self-propelled objects are those that start moving by themselves, without contact with other objects, whereas agentive objects are those that interact contingently with other objects, again without contact" (2013, p. 1). The agentive cue they used was whether the object appeared to initiate and engage in a brief conversation with the experimenter.

[13] Wertz (2019) refers to this collection of capacities as the "PLANT System", short for Plant Learning and Avoidance of Natural Toxins System.

found to be more reluctant to touch plants than artefacts and other natural entities (e.g., shells) (Wertz and Wynn 2014a), they touch plants less even after having made some contact with them (Włodarczyk et al. 2018), they look more at nearby adults when about to touch a plant than when they are about to touch other things (Elsner and Wertz 2019), and they are more disposed to overcome a reluctance to touch a novel plant compared to a novel artefact after witnessing an adult touching these things (Włodarczyk et al. 2020).

There are also early emerging plant-specific inferential dispositions that are connected to the representation of food. One that is particularly interesting is that infants seem to be predisposed to understand that plants are a food source, a predisposition that constrains social learning. In one study (Wertz and Wynn 2014b), 18-month-olds were shown a plant and an artefact, each of which had dried fruits attached to them, and witnessed an adult performing the very same action towards both the plant and the artefact. In one condition, the adult performed a *food-relevant action* (removing a fruit and placing it in her mouth), first from the plant and then from the artefact (or vice versa). In another condition, the adult performed a *food-irrelevant action* (removing a fruit and placing it behind her ear), again first from the plant and then from the artefact (or vice versa). In both conditions, this was followed by a second adult removing the rest of the fruits from both the plant and the artefact and placing them on two plates, with the plant's fruits on one plate and the artefact's fruits on another. Finally, the infants were permitted to reach for a fruit from either of the two plates while being asked *Which one can you eat?* (when they had witnessed the food-relevant actions towards the plant and the artefact), and *Which one can you use?* (when they had witnessed the food-irrelevant actions towards the plant and the artefact). Notice that if children are predisposed to think of plants as a food source, then after seeing the food-relevant action and being asked *Which one can you eat?*, they should show a preference for the plant's fruit rather than the artefact's. But after seeing the food-irrelevant action and being asked *Which one can you use?*, they should have no preference between them. This is exactly what happened. What's more, a second experiment revealed the same predisposition in infants as young as 6 months old.[14] Apparently, very young children expect that plants provide food.

Now, of course, with the right type of experience and a sufficient amount of time, a domain-general learning mechanism could figure out that plants are potential food sources. But what these results highlight is that, as a matter of fact, the human mind is innately disposed to think in these terms. After all, the tested infants were from New Haven, Connecticut, and presumably not routinely

[14] At this age, a violation of expectation procedure had to be used. Wertz and Wynn reasoned that if 6-month-olds are predisposed to classify plants as potential sources of food, then they should look longer when the food-relevant action is performed with the fruit taken from the artefact as opposed to fruit taken from the plant but should look equally at the food-irrelevant action performed with the fruit taken from the plant and the artefact. This is just what was found.

exposed to people removing fruits from plants prior to ingesting them—certainly not in the first six months of life. Quite the contrary, most fruit in American homes comes from non-plant sources—indeed, it comes from artefact sources, such as refrigerators, cabinets, bowls, colourful plastic containers, and so on. Moreover, typical plants encountered in American homes show no evidence of producing fruit or any other type of food, for that matter. And, yet, these 6-month-olds nonetheless expected plants rather than artefacts to be potential food sources.

To put this point more directly in terms of the needs of an empiricist connectionist learning system, if provided input for numerous observations of people acquiring food from plastic containers and other non-plant sources, it ought to adjust its connection weights to increase activation from ARTEFACT to FOOD SOURCE, while decreasing activation from PLANT to FOOD SOURCE. But then the simulation would get the wrong result; it wouldn't register real children's precocious appreciation of plants as potential food sources. In contrast, a rationalist connectionist network could implement and hence explain this early disposition by postulating initial connection weights that favour the spread of activation from PLANT to FOOD SOURCE. Likewise for the finding regarding infants' hesitancy to touch plants compared to artefacts and non-plant natural objects (Wertz and Wynn 2014a). Here too a rationalist connectionist model could postulate initial connection weights that favour the spread of activation from PLANT to DANGER, for example.

We have identified a number of problems with the assumption that connectionist research undermines concept nativism. Most importantly, we have stressed that connectionist models must take into account the sorts of arguments for rationalism put forward in Part II, which tell us that semantic memory doesn't start out as a largely undifferentiated system of representation. As a result, it would seem that the best way to develop artificial neural network models of the system of concepts involved in human semantic memory is to incorporate innate content through a range of characteristically rationalist psychological structures in the acquisition base—in other words, for researchers to pursue *rationalist* connectionist models.[15]

[15] Much the same could be said about Bayesian models of development, which are often seen as being opposed to rationalist accounts of conceptual development (Perfors et al. 2011). A major advantage of Bayesian models is that, in contrast to many neural net accounts, Bayesian models readily incorporate highly structured representations. There is also evidence that Bayesian models capture aspects of children's word learning (Xu and Tenenbaum 2007) and that infants employ Bayesian statistical inferences when forming generalizations about categories (Dewar and Xu 2010). But it's important to recognize that Bayesian models aren't inherently empiricist models even though Bayesian statistical inferences can be applied across different content domains. This is because any given Bayesian model will presuppose a certain way of representing both the data and the hypothesis space that learners are working with, as well as the relevant prior probabilities, and these could stem from characteristically rationalist psychological structures in the acquisition base. Moreover, domain-general Bayesian learning might be one component of an overall rationalist approach to a given

There is another possibility, however. This is that the problem with the connectionist models that we have been discussing so far isn't so much with their empiricism as it is with the particular kinds of artificial neural networks that they have employed. Perhaps newer types of artificial neural networks—ones that fall under the deep learning research paradigm—could do better. In fact, many proponents of deep learning see the situation in precisely these terms, advocating for recent machine learning techniques in AI while simultaneously adopting much the same empiricist assumptions that motivated earlier connectionist research.

Deep learning models differ in a number of ways from earlier types of neural network models (Buckner 2019; Serre 2019). The most obvious difference—and the main reason they are called *deep* learning models or *deep* neural networks—is that, compared to earlier models, they contain considerably more layers of units standing between a network's input units and its output units. Depth is a measure of the number of intervening layers—the more intervening layers, the greater the "depth" in the system. In addition, deep neural networks often employ combinations of hidden layers with tightly restricted connections among their units. This design can realize a set of hierarchically organized features detectors, which locate certain patterns regardless of where they appear in the network's input (or in a preceding hidden layer) and amplify these signals for further processing. Deep neural networks are also adept at dealing with very large quantities of data and are typically trained on massive data sets.

As a result of these and related characteristics, deep learning models have proven to be remarkably good at pattern classification. This has led to many advances in machine learning with a range of applications (LeCun et al. 2015). These include applications in medical diagnoses (Chilamkurthy et al. 2018), prediction of folded protein structure (Jumper et al. 2021), computer chip design (Mirhoseini et al. 2021), speech recognition (Hinton et al. 2012), translation between languages (Bahdanau et al. 2014), and obstacle detection for driverless

domain. For example, a rationalist account of conceptual development might hold that in a given conceptual domain there are innate principles, concepts, or other representations that provide a starting point for further conceptual development, and that what domain-general Bayesian learning does is fill out, or possibly override, some or all of this initial structure. Take, for example, an important recent proposal by Nichols (2021) which focuses on the origins of moral judgements. Nichols aims to explain such patterns as the fact that children acquire rules that prohibit *doing* something (e.g., stealing) rather than rules that prohibit both *doing something* and *allowing that thing to be done*, even though children don't seem to be given explicit information about this distinction. Nichols and his colleagues make a good case that this feature of children's learning may result from domain-general statistical inference—that it is a special case of the *size principle*, which favours a narrower rule over a more general rule when the available evidence is consistent with both (Nichols et al. 2016). However, this may all be true and yet there may still be a rich innate domain-specific sociomoral system that serves as part of the background to this learning. As we noted in the previous chapter, preverbal infants represent and reason about complex social relations and have expectations about such things as social dominance, aiding allies, and fairness, which are likely grounded in innate concepts or principles or rationalist learning mechanisms. Like the accounts based on neural networks we have been discussing in the text, Bayesian accounts will often best be developed within a broader rationalist framework.

cars (Huval et al. 2015), among many others. Artificial neural networks using deep learning pattern recognition techniques have also excelled in playing challenging strategic games, defeating the very best human players at Go (Silver et al. 2016; Silver et al. 2017).

Despite these and other significant achievements, and despite claims to the contrary by deep learning theorists, we will argue that deep neural networks don't significantly alter the landscape of the rationalism-empiricism debate. Like earlier connectionist models, deep learning models do not undermine rationalism, and also like earlier connectionist models, the best way to develop deep learning models of human conceptual capacities is in the context of a broader rationalist framework. We will illustrate these points by focusing on deep learning research on the categorization of visual images.

Deep learning models have had impressive results in this area (e.g., Krizhevsky et al. 2012/2017; He et al. 2016; Phillips et al. 2018; Hu et al. 2018). The training given to deep neural networks for image classification typically involves large sets of images in which the network is presented with many examples of different types of objects along with information about how they should be classified (whether it is being shown a dog, guitar, chair, lizard, elephant, and so on). For example, the network might be presented with an image of a *dog* (say, an image of a dog sitting on a rug) and be told that it is a dog, and then presented with an image of a *lizard* (an image of lizard lying in the grass) and told that it is lizard, and then presented with a different image of a *dog* (a dog running on a beach) and told it is a dog, and so forth for millions of images. Following this type of training, deep neural networks are then able to generalize beyond the exemplars that they were trained on, successfully classifying new exemplars from the categories they were trained on (e.g., correctly classifying new images of dogs that were not part of the training set). What's so exciting about this research is the phenomenal accuracy such models have achieved, which by some measures, equals or even surpasses human accuracy rates (see, e.g., He et al. 2016).[16]

In considering the bearing of these results on the rationalism-empiricism debate, it should be kept in mind that these sorts of models have been designed with a different aim than Rogers and McClelland had for their model of semantic memory. While Rogers and McClelland's model was intended to learn a broad range of facts about different types of objects—facts that correspond to people's everyday knowledge about what they are like and what they can do (e.g., that birds fly)—deep learning image classification models are directed at the problem of visual categorization. Their focus is primarily on the capacity to classify visual

[16] We should note that this favourable comparison with human categorization performance depends on precisely how categorization accuracy is measured. One common measure of accuracy deems a network to make the correct choice for a given item if the correct choice is among the network's top five choices for that image. Overall accuracy, then, is a matter of the percentage of items where the correct answer is among the network's top five choices.

images in terms of object categories (e.g., the capacity to classify images of birds *as birds*, or, in any case, to classify birds as a unified visual category).

The fact that deep neural networks have achieved such high levels of accuracy in classifying new instances of categories they have been trained on naturally suggests they are forming representations that correspond to ordinary visual object categories (e.g., *dog*, *bird*, *guitar*, and so on). However, there is another possibility, namely, that their success depends more on their generalizing across high-level patterns in the images *taken as a whole* and does not involve their ability to isolate and identify the focal objects in the images at all.

In fact, evidence from research on so-called *adversarial images* strongly suggests that something like this is what is going on. Adversarial images are ones that have been altered specifically to trick deep neural networks (see Figure 19.3). One type of adversarial image involves global changes across the entire image, changes that are subtle enough that human observers may not even detect them but that nonetheless lead state-of-the-art deep neural networks to dramatically change how they categorize the image, for example, switching from confidently saying that an image is of a Christmas stocking to confidently saying that it is of an elephant (see Figure 19.3(a) and (b)) (Moosavi-Dezfooli et al. 2017). Another type of adversarial image involves relatively small local changes to an image's background without affecting the focal object. This too can lead to dramatic changes in image categorization, for example, leading deep neural networks to switch from confidently saying that an image is of an elephant to confidently saying it is of a stingray (see Figure 19.3(c) and (d)) (Karmon et al. 2018). Highly successful deep neural networks for image classification can also be derailed by patterns that are completely unrecognizable to human observers yet lead these models to confidently categorize an image as being of a particular type of animal or artefact, for example, saying it is an image of peacock or a penguin (see Figure 19.3(e) and (f)) (Nguyen et al. 2015).

Needless to say, none of this calls into question the value of deep neural networks in a wide range of important applications. Such networks still represent major advances in medical diagnosis or speech recognition technology, for example, and accordingly are of enormous practical value. But what these specific types of failures do strongly suggest is that, when deep learning models are successful at correctly categorizing images of ordinary objects—elephants, stockings, and so on—they aren't doing so in the same way as humans.

For many researchers who work on deep learning, this disparity between artificial and human image processing is of little consequence, since their aim isn't to capture the way that human visual categorization works. Rather, like other researchers in AI, their aim is to show that certain things that humans do can be done equally well (or even better) by a machine, even if the processes the machine relies upon are radically different from the psychological processes underlying human cognitive capacities. However, if image categorization in deep neural

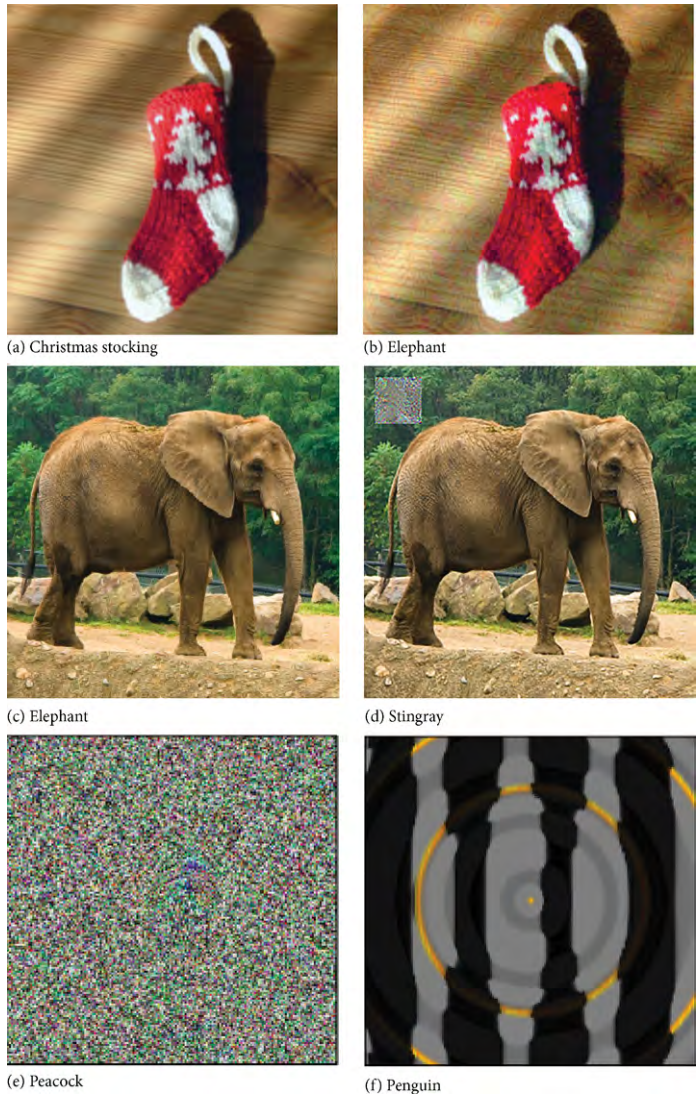

Figure 19.3 Deep learning image categorization errors. Adversarial images expose significant limitations among deep learning models that have otherwise been found to successfully categorize novel images. The labels under these images reflect categorization decisions that state-of-the-art deep neural networks made with high confidence. While image (a) is correctly categorized as a Christmas stocking, subtle changes to the entire image lead it to be incorrectly categorized as an elephant, as in (b) (Moosavi-Dezfooli et al. 2017). Likewise, while image (c) is correctly categorized as an elephant, a change made just to a small part of the image background causes the network to categorize it as a stingray, as in (d) (Karmon et al. 2018). The images (e) and (f) aren't seen as animals of any type by human observers but are categorized by highly successful deep neural networks as a peacock and a penguin with over 99% confidence (Nguyen et al. 2015). (Images reproduced with permission. Images (a) and (b) from Moosavi-Dezfooli et al. 2017, images (c) and (d) from Karmon et al. 2018, and images (e) and (f) from Nguyen et al. 2015; (c) Photo: Jason Pratt / Wikimedia Commons / CC-BY-2.0.)

networks is achieved using computational processes that are very different from the ones that are at work in ordinary human visual categorization, then it tells us very little about human psychology and the representations and processes involved in human visual categorization. This means that it will also tell us very little about the rationalism-empiricism debate and the origins of human concepts.

Moreover, even if there were a greater correspondence between the way that deep neural networks and human beings categorize visual images—even if these representational capacities were identical—deep neural networks need to be able to acquire these representational capacities under the same kinds of conditions that actual human learners are able to if deep learning research is to have a direct bearing on the rationalism-empiricism debate. As we noted earlier, deep neural networks are typically trained on massive data sets. Image classification networks are often given literally millions of examples of distinct images of the objects—with complete and accurate information in each case as to what the image is an image of. This is in stark contrast with some of the case studies that were discussed in Part II, where representational and conceptual capacities were found to be in place with little or no experience of instances of their categories (see Chapters 8, 10, 13, and 14). Representations of biological motion, partly occluded objects, number, and faces, for example, are all manifest in human infants at birth.

Even for representations of specific object categories—the sorts of animals, artefacts, and other objects that these networks have been trained on—it is not at all plausible that human learners normally make use of the massive amount of input that deep neural networks are typically given.[17] Nor are human learners provided with comprehensive and accurate feedback about how to generalize from one experience to another in the course of a learning period, making sure that they treat all of the dogs as one type of object, all the elephants as another, and so on.

None of this is to say that deep learning systems can't be useful in modelling human visual categorization or the acquisition of human concepts of particular types of objects. For all that we have said, deep neural networks may provide a key part of the story of how these concepts are acquired. There are two ways this

[17] In fact, humans can readily learn new visual object categories on presentation of a single new instance, via what has been called *one-shot learning*. Bayesian models of one-shot learning explain it by means of prior knowledge that structures this learning (Fei-Fei et al. 2006; Lake et al. 2015). Bayesian researchers generally suppose that this prior knowledge is acquired via domain-general learning mechanisms, but another possibility is that some of it is innate or acquired via rationalist learning mechanisms. A striking example of one-shot learning involving rationalist learning mechanisms is the ability of newborn chicks to acquire viewpoint invariant representations of an object based on their very first image of an object (Wood and Wood 2020). In this research, newborn chicks were reared in highly controlled conditions so that their initial visual experience after birth was limited to a single object seen only from a single viewpoint. Subsequently they were shown this same object from a different viewpoint and a new object. The chicks were found to prefer the familiar object (that is, the one they had been trained on but from a different viewpoint) over the new object.

might go. The first, which is closest to the aspirations of many of the theorists working on deep neural networks, is that these models could be modified to build in more innate components to constrain learning in a way that makes the account only a modestly rationalist one overall. This approach might include a network organization that is predisposed to segregate objects from their background and from one another and, moving beyond images to dynamic input, might incorporate the sorts of innate domain-specific principles for representing and reasoning about objects that we discussed in Chapters 10, 15, and 17.[18] It might also include rationalist learning mechanisms for biological motion detection and other special-purpose mechanisms that act as filters on the categorization process. In this way, deep neural networks for image classification could be developed to include at least a modest amount of innate content and innate domain-specific structure—structure that initially constrains how they respond to visual images and that channels a network's learning in the same direction that it would be channelled by the human acquisition base.

The second way in which deep learning might contribute to the acquisition of concepts of particular types of objects is as part of a thoroughly rationalist approach that is more in line with the form of concept nativism we argued for at length in Part II. Such an account would see deep learning models as fully integrated with a rich acquisition base with characteristically rationalist psychological structures bearing on many different conceptual domains. It wouldn't take the kinds of concepts at issue here—concepts of specific types of objects (STOCKING, ELEPHANT, etc.)—to be innate.[19] But there is good reason to suppose that their acquisition builds on a diverse range of characteristically rationalist psychological structures in the acquisition base that work in conjunction with domain-general processes.

This second way of incorporating more innate structure is motivated by the fact that many of the obstacles for empiricist connectionist models that we mentioned earlier would equally be obstacles for empiricist deep learning models. For example, as we noted in the discussion of Rogers and McClelland, the many rationalist learning mechanisms that we argued for in Part II are relevant to any account of semantic memory. The further examples we discussed in this chapter pertaining to animals and plants also illustrate the need to move towards a more

[18] Adding this structure wouldn't just make deep learning systems better able to model human learning. By building in more innate domain-specific structure, this might also make them better at fulfilling the goal in AI research of creating machines that excel at tasks that normally require human intelligence (cf. Lake et al. 2017; Marcus 2018; Versace et al. 2018).

[19] It's worth noting that such an approach would not even require these sorts of concepts to be learnable *only* via rationalist learning mechanisms. Perhaps under some circumstances (e.g., circumstances in which learners have access to unusually large amounts of data or to data of just the right type), they could be acquired by domain-general learning mechanisms. The point is that these conditions do not hold for the sorts of cases that are the focal point of competing rationalist and empiricist models—these are typically cases in which there isn't access to this type of data or the time or other resources that would be needed to make good use of it.

rationalist account. Prelinguistic infants seem to intuitively expect animals to have insides. But this isn't because they have seen a great many "open" animals. Likewise, prelinguistic infants seem to intuitively understand that plants are a food source. But this isn't because they regularly see food being extracted from plants. So, regardless of how deep a network is, if an artificial network is going to be able to explain how human children come to have such knowledge, it needs to be able to do this without having a training period that includes many examples of these types of experiences.

In sum, there is no incompatibility between artificial neural network modelling and concept nativism. While many researchers have assumed that artificial neural networks vindicate empiricism, such networks can be developed in either empiricist or rationalist ways. Given the powerful evidence for a broad range of concepts being innate or acquired via rationalist learning mechanisms (see Part II), the most plausible path forward for artificial networks to model human concept acquisition is one in which they incorporate a suite of characteristically rationalist psychological structures.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0019

20
Neuroconstructivism

Domain-specific learning mechanisms are a prominent and central feature of rationalist theories of the mind. They are also widely rejected by empiricists. This rejection represents a core element of empiricism when the domain-specific mechanisms in question are ones that are claimed to be innate or when their domain specificity is claimed to stem from characteristically rationalist psychological structures in the acquisition base. In these circumstances, the rejection of domain-specific learning mechanisms is tantamount to a rejection of rationalism. Many empiricists would go further, rejecting not just these types of domain-specific learning mechanisms but also the very idea that domain-specific mechanisms play an important role in human cognition, even in adults. However, this is not the only option for empiricists. Another possibility, briefly noted in Chapter 19, is to accept the existence of domain-specific mechanisms in mature cognition, but to hold that such mechanisms are not rationalist learning mechanisms, and instead take them to be acquired on the basis of more fundamental domain-general learning mechanisms.

In our view, this alternative approach is one of the most interesting ideas to come out of contemporary empiricist theorizing. Such a view doesn't deny that domain-specific mechanisms play an important role in human cognition; it just contends that these arise as a product of domain-general empiricist learning mechanisms. The primary architect of this type of view, which we will refer to as *neuroconstructivism*, is Annette Karmiloff-Smith, whose seminal book *Beyond Modularity: A Developmental Perspective on Cognitive Science* laid the foundations for this theoretical approach. In this chapter, we take a careful look at neuroconstructivism, focusing in particular on Karmiloff-Smith's own developments of neuroconstructivism and her highly influential critique of rationalist accounts of cognitive and conceptual development (Karmiloff-Smith 1998, 2009a; Karmiloff-Smith et al. 2003; Filippi and Karmiloff-Smith 2013). As we will see, her view is of interest not only as an alternative to concept nativism but also because it claims to undercut rationalist arguments that focus on evidence from developmental disorders.

Let's begin with a brief overview of Karmiloff-Smith's neuroconstructivism. As Karmiloff-Smith explains, neuroconstructivism maintains that the cortex has a largely equipotential initial structure that includes a modest number of *domain-relevant* mechanisms. Unlike domain-specific rationalist learning mechanisms, what Karmiloff-Smith calls "domain-relevant mechanisms" are taken to individually have

the wherewithal to process heterogeneous types of content; a given mechanism of this type, however, can turn out to be "somewhat more relevant to one kind of input over others" (Karmiloff-Smith 1998, p. 390). As Karmiloff-Smith sees it, the match between these mechanisms and certain types of input resides in small seemingly inconsequential features of different cortical areas, such as neuronal density, the amount of white matter, or neuronal firing threshold levels. Over time these small disparities can translate into big differences in cognitive development. For example, they may result in a domain-relevant area repeatedly receiving and processing the same type of input, leading it to acquire a stable specialized representational function. And this in turn may prompt a cascade of events elsewhere in the brain, potentially leading to further functional specialization. For Karmiloff-Smith, neuroconstructivism requires a major shift in thinking about the origins of domain specificity. "[R]ather than evolution providing pre-specified representations, this change in perspective places the mechanism of progressive ontogenetic change on centre stage" (Karmiloff-Smith 1998, p. 390).

It is important not to confuse neuroconstructivism's domain-relevant mechanisms with concept nativism's domain-specific rationalist learning mechanisms. Even though both of these proposed types of mechanisms are, in a sense, fit for particular kinds of information processing, Karmiloff-Smith's domain-relevant mechanisms are not domain-specific rationalist learning mechanisms with specific representational functions. They are simply supposed to be better candidates than other cortical areas for processing certain types of information because of a very rough match between some of their structural features and this type of input.

For example, Karmiloff-Smith rejects the rationalist proposal that there is an innate language acquisition system that incorporates the principles of Universal Grammar or any other rules or constraints specific to language. Instead, she proposes that there is a general-purpose learning mechanism that includes a domain-relevant mechanism—in this case, a feedback loop. According to Karmiloff-Smith, the incorporation of the feedback loop makes the general-purpose learning mechanism more suited to processing sequential input than neural tissue where there is no feedback loop and hence leads it to favour speech input. In time, she thinks, this fortuitous privileging of linguistic input leads to the development of entrenched language-specific processes even though there is nothing about the system that is initially specific to language (Karmiloff-Smith 1998, p. 390). Likewise, rather than accept the existence of an innate domain-specific mechanism involved in face recognition, Karmiloff-Smith proposes that domain-specific circuits for face recognition start out as a domain-general mechanism for coding visual patterns in general, not faces in particular. But since they happen to end up processing far more face-based input than other types of information, this causes the initially non-specialised mechanism to become a stable face-specific mechanism (Karmiloff-Smith 2009a, p. 59).

In short, we have two competing explanatory frameworks to account for any domain-specific mechanisms that are found in typical adult cognition. These are the neuroconstructivist framework (with its initial domain-relevant mechanisms, which are domain general) and the rationalist framework (in which the domain specificity often traces back to characteristically rationalist psychological structures in the acquisition base). So the question at issue between neuroconstructivism and concept nativism isn't about the existence of domain specificity, but rather about its origins—in particular, whether it should be explained solely in terms of domain-general learning mechanisms or at least partly in terms of characteristically rationalist psychological structures. Karmiloff-Smith's claim is that the rationalist framework has significant shortcomings, and that neuroconstructivism, which she takes to be immune to these difficulties, is the superior theoretical framework.

Some of Karmiloff-Smith's arguments in favour of neuroconstructivism will be familiar from earlier chapters. She objects to rationalism on the grounds that it "has a static flavour and implies genetic determination, as if states in the brain were hardwired, unchanging, and unaffected by developmental or environmental factors" (Karmiloff-Smith 2009a, p. 58). In contrast, neuroconstructivism has the virtue of capturing the complex process of development in the way that it treats the brain as a "self-structuring, dynamically changing organ over developmental time as a function of multiple interactions at multiple levels, including gene expression" (Karmiloff-Smith 2009a, p. 58). Likewise, Karmiloff-Smith claims that neuroconstructivism fits better with what has been discovered regarding the brain's plasticity. Unlike rationalist accounts of plasticity, neuroconstructivism's account of plasticity is more encompassing. It has the advantage, she claims, that it treats plasticity as a "basic feature of normal and atypical cortical development" (Karmiloff-Smith 1998, p. 397). In addition, Karmiloff-Smith (1998, 2009a) argues that neuroconstructivism is supported by work in computer modelling that shows how the different functions of the dorsal and ventral visual pathways could develop from small differences corresponding to neuronal firing thresholds (O'Reilly and McClelland 1992). In this model, the very same input passed through two channels that were initially identical apart from the speed at which their activation levels changed. This resulted in the faster channel coming to represent object location (as in the dorsal stream) and the slower channel coming to represent object identity (as in the ventral stream)—a demonstration that neural specialization can emerge from development rather than having to be "prespecified in the infant neocortex" (Karmiloff-Smith 1998, p. 392).

Since we have already discussed similar arguments against rationalist accounts of cognitive and conceptual development (see Chapters 3, 4, and 13), we will be brief in responding to these arguments here. In our view, most of the advantages that Karmiloff-Smith claims on behalf of neuroconstructivism turn on mistaken

characterizations of the rationalist position or on considerations that, contrary to initial appearances, don't actually favour her neuroconstructivism.

Karmiloff-Smith's principal argument for neuroconstructivism is that it treats development as a dynamic process. Rationalism, to its detriment, is supposed to entail a static view of the brain in which its domain-specific rationalist learning mechanisms are not affected by the environment or by prior states of development and are essentially genetically determined. Unfortunately, this purported contrast greatly distorts the rationalist position.

The charge of genetic determinism or that rationalism takes the environment to be irrelevant to development is just plain wrong. As we emphasized earlier (see Chapter 3), rationalists don't hold the view that genes alone bring about the mind's innate cognitive traits. It is recognized by all parties to the rationalism-empiricism debate—rationalists and empiricists alike—that all cognitive traits are the product of interactions between genes and the environment. Genetic determinism is a straw man position. Rationalists can and do recognize that the causal processes that build the brain involve enormously complex interactions between genes and environmental factors of all sorts (factors in the intercellular and extra-cellular environment, including which other genes are expressed or silenced).

Also, as explained earlier, rationalists assume that the environment plays a powerful role in cognitive development, not just in a few isolated cases but as a matter of course, and that much of the time it plays a distinctively psychological role providing crucial input to a rationalist learning mechanism. We have given numerous examples throughout the book of rationalist learning mechanisms that embrace this sort of substantive environmental input. To mention just two, in Chapter 15 we saw that even for cognitive domains as basic as the representation of physical objects, rationalist models of acquisition postulate learning processes that require environmental input, such as experiences of contrastive outcomes of similar physical events and processes. And in Chapter 18 we noted that rationalist accounts of moral development can embrace substantial variation in adult moral norms that is sensitive to environmental factors. For example, we suggested that early expectations regarding fair distributions (and similar sociomoral matters) can give way to diverse adult norms, with early norms potentially being overridden or modified through environmental input. Far from taking the environment to be irrelevant to development, rationalist models assume that the environment plays an essential role and has far-reaching effects on the conceptual system, just as empiricists do.

Like many critics of rationalism, Karmiloff-Smith claims that cortical plasticity argues against a rationalist view of the acquisition base. She takes neuroconstructivism to be superior in light of the ubiquity of neural plasticity. That is, plasticity is seen "as basic feature of normal and atypical cortical development". But as we noted in our explanation of the argument from neural wiring in Chapter 13,

the issue that divides rationalist and non-rationalist accounts of the brain's plasticity isn't whether cortical plasticity exists or even how common it is. Rationalists and empiricists agree that neural changes underlie all mental activity, meaning that plasticity will be ubiquitous. The pertinent issue is whether at least a significant amount of this potential for change takes the form of *constrained plasticity* in which there are substantive limits on the plasticity of neural structure and function.

We have seen that there are indeed rather striking examples of constrained plasticity and that these argue for a rationalist framework. For example, recall the finding that the same functional differentiation that normally occurs in the parietal visual cortex shows up in people who are congenitally blind, with the representation of artefacts in medial regions and the representation of living animate things in lateral regions (Mahon et al. 2009). It is hard to see what the neuroconstructivist account of this differentiation would be. If the brain is as plastic as Karmiloff-Smith maintains, then why should the same pattern of functional differentiation occur despite systematic and massive differences in input? On the other hand, this differentiation is readily explicable on a rationalist account in which the functional structure of the ventral visual cortex is significantly constrained regarding how a relatively small number of evolutionarily important categories are represented. Likewise if the brain is largely unconstrained in terms of plasticity, it is hard to see why individuals like Adam (discussed in Chapter 13), who have suffered focal brain damage very early in life are unable to compensate for this early damage despite normal facial input and the great importance of face processing to daily life (Farah and Rabinowitz 2003). It is just as true for Adam as for neurotypical individuals that a disproportionate percentage of visual inputs will be from faces, so neurocontructivism should predict that Adam should still develop normal (or near normal) face processing mechanisms even if this takes a bit longer than in neurotypical individuals.

Finally, it is certainly noteworthy that something as apparently inconsequential as a difference in the speed of activation changes in two networks can, under certain assumptions, lead a computer model to settle on differences of function similar to the differentiation in the dorsal and ventral visual pathways (O'Reilly and McClelland 1992). But this only tells us about a theoretical possibility regarding cortical development. It doesn't tell us how this development actually takes place and whether the two cortical streams realize distinct innate domain-specific systems. As with the empiricist artificial neural networks we looked at in the previous chapter, the question isn't whether it is *possible* for a computer model to produce an outcome that approximates a feature of mature cognition. It's whether the processes that mediate this outcome correspond to the processes that actually take place in real human development. Karmiloff-Smith has not provided any

substantive grounds to suppose that they do (nor to suppose that such accounts would generalize to other areas of conceptual development).[1]

For these reasons, we take Karmiloff-Smith's general case for neuroconstructivism and against rationalism to be unconvincing. But we still need to consider her critique of the way that some theorists have taken patterns of selective deficits in cases involving so-called genetic disorders to be evidence for rationalism.[2] If her critique is successful, then nothing can be learned about the origins of normally developing cognitive capacities by looking at the performance of children with Williams syndrome (WS), autism spectrum disorder (ASD), or other similar developmental conditions. And if this is true, it might seem to considerably weaken the argument from neural wiring.

The sort of rationalist argument Karmiloff-Smith takes as her target is one in which the existence of domain-specific rationalist learning mechanisms in neurotypical development is claimed to be supported by patterns in atypical development that are associated with genetic disorders. In a standard rationalist argument of this kind, specific cognitive impairments that appear alongside other relatively spared psychological capacities are regarded as providing evidence that potentially argues for domain-specific rationalist learning mechanisms. In discussing the argument from neural wiring in Chapter 13, we presented a few examples of arguments along these lines. One involved the pattern of spared and impaired abilities in children with WS. For example, although WS children are known to have numerous difficulties with spatial representation, they have relatively strong face representation capacities, suggesting that face recognition is supported by an innate domain-specific rationalist learning mechanism (among other related characteristically rationalist psychological structures in the acquisition base). Another is the finding that children with ASD can have severe difficulties with false-belief tasks but succeed on similar tasks with photographs and maps. This suggests that representing and reasoning about beliefs is likewise supported by a domain-specific rationalist learning mechanism.

Karmiloff-Smith claims that these and similar rationalist arguments are all misguided. The problem in her view is that no conclusion about typical development can be based on evidence regarding atypical development. As Karmiloff-Smith sees it, the "abnormal brain is not a normal brain with parts intact and parts impaired. It is a brain that develops differently throughout embryogenesis and postnatal brain growth" (Karmiloff-Smith 2000, p. 148).

[1] It is no easy matter to study the functional properties of the infant brain, but an investigation of newborn monkeys (using fMRI) suggests that the large-scale organization of the visual system develops in advance of visual experience and that this includes the differentiation of the dorsal and ventral pathways (Arcaro and Livingstone 2017).

[2] As we noted in Chapter 13, the term "genetic disorder" is problematic in a number of ways. Because of this we will sometimes use terms like *atypical development*.

Karmiloff-Smith's point is that each step in development crucially depends on what has happened before. Since typical and atypical development involve important early differences that contribute to differing neurodevelopmental trajectories, subsequent developmental processes will produce pervasive differences in the psychologies of neurotypical and atypical individuals. Even when children or adults with a given condition appear to perform a task in the typical range, this needn't be because their atypical development has had no effect on the capacity. The behaviour is likely to be implemented by different mechanisms and processes that manage to achieve much the same overall effect. For example, although people with WS excel at identifying faces, Karmiloff-Smith takes them to recognize faces in an atypical way. Where typical face recognition is holistic, Karmiloff-Smith holds that WS face recognition succeeds by focusing on individual facial features (Karmiloff-Smith 1998). As a result, their spared face recognition abilities are completely uninformative about whether neurotypical face recognition is grounded in a domain-specific rationalist learning mechanism.

Note that Karmiloff-Smith's argument is intended to raise a *principled* difficulty for any rationalist argument of this form.[3] Her main reason for supposing that atypical development doesn't tell us anything about typical development is her concern that small initial differences can be amplified through cascading processes in development and lead to widespread differences in the brain even if some of these differences aren't immediately apparent given current testing methods. Because of these cascading effects, we should never assume that there is anything typical about the brains of people with a genetic developmental disorder and should never conclude that a purportedly spared cognitive capacity is the same as what's found in typical development. As a result, the finding of what looks to be a spared cognitive capacity can have no bearing on whether this ability is typically grounded in a domain-specific rationalist learning mechanism.

An analogy might help to bring out the structure of this argument. Imagine trying to figure out the inner workings of a fully functioning clock radio by comparing it to other *malfunctioning* clock radios. This might be useful if the comparison were restricted to others of the same model but not when comparing the item to different (malfunctioning) models. The problem is that these other devices may realize the same general functions (clock, alarm, music, nightlight, etc.) in different ways. Suppose that the clock in one of these devices works fine despite the fact that its snooze function is faulty. This tells you nothing about the design of the fully functional device since the two are, after all, different models that may not be built from the same components or have the same overall internal organization.

[3] As we will see shortly, she supplements this general argument with a number of detailed case studies, which will need to be addressed separately.

Do these considerations undermine the argument from neural wiring? Is Karmiloff-Smith right that absolutely *nothing* can be learned about the neurotypical acquisition base by examining the pattern of abilities and deficits in people with developmental conditions like WS and ASD that are associated with particular genetic anomolies? No. While it is true that the evidence from these sources may not provide an infallible guide to the domain-specific mechanisms that support neurotypical development, it can nevertheless play an important role in the identification of such mechanisms.

The cornerstone of Karmiloff-Smith's argument is the claim that small initial differences can initiate developmental cascades that produce pervasive developmental effects. As a result, there is supposed to be no reliable point of comparison between individuals following atypical and typical developmental trajectories. They may look similar in certain respects, but given the import of small initial differences in their development, we should expect widespread differences regarding the nature of their cognitive capacities and the neural mechanism that realize them.

But notice, if the problem about using atypical development to learn about typical development is these small initial differences, then we shouldn't be in a position to draw conclusions about typically developing individuals even by studying *other typically developing individuals*. This is for the simple reason that, even among typically developing individuals, there are endlessly many small initial differences that are unique to each individual. No two people *ever* share exactly the same initial conditions (not even so-called identical twins). Consider also the enormous variability among the environmental factors that potentially affect development at every stage in the lifespan. Typically developing foetuses, infants, and children undoubtedly develop under a wide range of differing conditions. These include things like their mothers' stress levels during pregnancy, their early diet (e.g., breast milk vs. formula), their exposure to different types of bacteria and viruses, the number of siblings, time spent in structured play, access to formal education, and numerous other variables, not to mention all the different combinations of these variables. If we were to take Karmiloff-Smith's argument to its natural conclusion, no science of development would be possible. There would be no room for generalizations about any cognitive capacities and hence no room to investigate such things as the typical development of object representation, face representation, numerical representation, and so on.

But clearly there is no reason to be so pessimistic about the possibility of a science of cognitive development. Consider that brain development is grounded in biological development more generally, and that biological development is robust in the sense that the development of a trait is often insulated from variability regarding the development of other traits (Machery 2011). So, on the reasonable assumption that brain development is like other aspects of biological development, brain development ought to be robust too.

Moreover, work in neuroscience has also shown that much brain development *is* robust, resulting in the same anatomical structures and the same essential wiring across a spectrum of conditions. Consider something as basic as the parts of the brain involved in visual processing. Standard accounts identify at least a dozen subsystems (V1, V2, V3, MT, MST, VIP, etc.) that are present in all neurotypical humans—and perhaps three times that number—despite many differences in initial conditions and developmental environments across individuals (Gazzaniga et al. 2019). These highly robust neurological subsystems reliably support the same cognitive functions across individuals. No one in developmental neuroscience is much troubled about claiming that the standard wiring diagram mapping out these structures and their complex interrelations captures the shared structure of (neurotypical) human vision. But if we were to take seriously the worry about small differences affecting development, it would be a completely open question whether this diagram is informative about any particular individual even if he or she seems to see things in the same way as the very individuals whose neural structures provided the basis for scientific theories of the visual cortex.

Much the same point holds for different *groups* of typically developing individuals. For example, given the initial differences between male and female foetuses and infants—chromosomal differences, hormonal differences, differences in parental rearing, etc.—we wouldn't be able to use evidence regarding the development of the visual system in one sex to learn about the visual system in the other sex. After all, these differences could cascade in such a way that importantly different processes end up producing capacities that simply appear similar given current testing methods.

Likewise, if the concern about cascading effects were to prevent us from using evidence from preserved functions in atypical development, then we shouldn't be able to use evidence from animals either. For example, we shouldn't be able to use evidence from mammals or even non-human primates when it comes to studying the structure of the human visual system. After all, there are obviously numerous initial differences that affect the development of animals and humans. Nonetheless, the study of animals has been invaluable in discovering the structure of the human visual system (see, e.g., Bechtel and Mundale 1999).[4]

Finally, it is also worth noting that if Karmiloff-Smith were right about the import of potential cascading effects, this point would cut both ways—just as we wouldn't be able to use data from atypical development to learn about

[4] This is not to say that there are no relevant initial differences, or that the human visual cortex is identical to that of other primates. The point is simply that the existence of some differences does not in and of itself mean that no inferences between cases are possible, or that there cannot be common mechanisms or shared general features that are unaffected by the small differences.

neurotypical development, we wouldn't be able to use data from neurotypical development to learn about atypical development. Yet the study of neurotypical development has proven to be invaluable for understanding atypical development.[5]

What these considerations go to show is that Karmiloff-Smith's case against rationalists' use of data from atypical development doesn't hold up as a principled argument. There is no more of a general obstacle to learning about neurotypical development by studying spared abilities in instances of atypical development than there is by studying other neurotypically developing individuals, or by studying non-human animals such as cats and monkeys.

At the same time, it must be noted that it is still *possible* that in some specific case the seemingly spared abilities in a given type of atypical development are in fact different from their neurotypical counterparts, and this might undermine the argument from neural wiring in such a case. But if this is what the neuroconstructivist's argument comes to, it is important to recognize that the burden of proof has shifted. It is now up to neuroconstructivists to show that, in a given type of atypical development, the seemingly spared abilities in question are actually atypical in important respects and hence of little relevance to the understanding of neurotypical development.

Let's consider a case that Karmiloff-Smith has discussed extensively—Williams syndrome. As you may recall from Chapter 13, WS is a rare genetic disorder that is associated with a complex profile of cognitive deficits, including considerable difficulties with many aspects of spatial cognition combined with relatively spared language and face representation (Landau and Hoffman 2012). However, as Karmiloff-Smith (1998) points out, the genetic anomalies in WS have widespread effects on brain development. If one compares the fully formed brains of WS adults with neurotypical adults, the differences are extensive. WS brains are considerably smaller (80% the volume of neurotypical brains) with reduced cerebral grey matter and other atypical properties. For Karmiloff-Smith, such differences show that it is wrong to describe the WS brain as being typical in any respect. We should expect widespread repercussions affecting WS cognition. And this is exactly what is found, according to Karmiloff-Smith.

We will focus on face recognition, one of the flagship examples she offers in support of her claim that rationalists have mistakenly taken cognitive capacities

[5] See, for example, Landau and Hoffman's (2012) enlightening account of WS, which carefully examines many aspects of spatial representation, comparing neurotypical developmental (and its timeline) to the abilities found in children and adults with WS. Landau and Hoffman build a very strong case for the view that many of the signature peaks and valleys in WS cognition can be accounted for in terms of a developmental lag in which the overall pattern of development is the same as for neurotypical children yet is delayed and ceases at a point corresponding to the cognitive capacities present in neurotypical 4- to 5-year-olds.

to be preserved in the brains of WS individuals.[6] For Karmiloff-Smith, people with WS may be relatively good at recognizing faces, but not because they have a preserved face-specific rationalist learning mechanism. If there were such a mechanism that was spared relative to other WS cognitive capacities, it would function in the same way as it does in neurotypical control participants. But according to Karmiloff-Smith, WS individuals *don't* represent faces in the same way as neurotypical individuals—where faces are typically represented in a holistic manner, people with WS analyse faces largely in terms of their individual features.

The original basis for this claim is Karmiloff-Smith's (1997) investigation of the face recognition abilities of a small number of adolescent and adult experimental participants with WS. Her investigation reports two relevant experiments. The first included a standard face memory task (the Benton test) and an informal follow-up in which participants were shown the same faces again but upside down. (Recall that one of the key features of neurotypical face perception is that it is impaired for inverted faces and that this is generally interpreted as showing that neurotypical face perception is holistic; see Chapter 8.) Karmiloff-Smith found that the majority of the WS participants performed in the typical range for the initial task but differed from age-matched neurotypical controls regarding inverted faces. And when asked about how they remembered a face, they pointed to individual facial features, in contrast with the neurotypical participants, who "talked about the whole face looking the same" (p. 518).

It is doubtful that much can be gleaned from these results, however, since the report does not go into any detail regarding the procedures involved in the experiment and does not present any data or statistical analysis. Moreover, self-report regarding such processes is notoriously unreliable. People—neurotypical or otherwise—simply do not know how they accomplish basic cognitive tasks, whether the task involves recognizing faces, recovering the syntactic structure of a sentence, or inferring the edges of an object from differences in surface luminance. People's answers to questions about how they accomplish such tasks say far more about their theories of how their mind works or about their understanding of what the questioner wants to hear than about the actual mechanisms underlying the cognitive ability in question.[7]

The other experiment in Karmiloff-Smith (1997) has similar problems. The same participants were given a matching task, requiring them to compare faces in which stimuli varied as to how similar or different they were. The stimuli were designed so that correct answers were supposed to require more holistic

[6] See Chapter 8 for discussion of how rationalist accounts of the origins of face recognition abilities bear on concept nativism in light of different ways of drawing the conceptual-nonconceptual distinction.

[7] People's answers to these sorts of questions also give rise to illusory explanations, as discussed in Chapter 5.

processing in some cases and more feature-by-feature processing in others, and Karmiloff-Smith contends that the WS participants' performance was poor with stimuli requiring holistic processing. However, as with the previous experiment, the details regarding the nature of the stimuli and the procedures used are not reported, making it difficult to draw any substantive conclusions.

In subsequent work, Karmiloff-Smith et al. (2004) revisited the WS face-processing debate with three further experiments examining the developmental trajectory of face perception in people with WS. We will discuss just the first of these experiments, which we take to be the strongest of the three. In this experiment WS participants and age-matched neurotypical control participants were shown pairs of faces one at a time, with both faces in the pair displayed either upright or inverted, and had to determine whether the faces were the same or not. For the face stimuli that didn't match, this was done in two different ways. In one of these (the *feature condition*), a single feature differed between the two faces but the spatial relations among the features were the same (e.g., although the faces had different noses, these noses were matched regarding their distance to the eyes, mouth, and so on). In the other (the *configural condition*), the features were identical—same nose, same mouth, etc.—but the distances between some of these features was different (e.g., the nose was further apart from the mouth in one face than in the other). Karmiloff-Smith et al. analysed the results by looking at the cases where the faces matched separately from the cases where they didn't match. They found no significant difference in accuracy between WS participants and neurotypical control participants for the matching stimuli, but found that with the non-matching stimuli, the WS participants did worse than controls in the upright configural condition (i.e., the condition where the difference between the faces resided solely in the spacing of features). This led Karmiloff-Smith and her colleagues to conclude that people with WS don't process faces in the same holistic way as neurotypical controls.

What should we make of these results? The first point to note is that, as Karmiloff-Smith et al. themselves observe, in general, the WS participants performed as well as the neurotypical control participants. "At first blush, our overall result suggest…the clinical group was as accurate as the controls on both identity recognition and difference detection" (p. 1271). The difference between them only showed up when Karmiloff-Smith et al. restricted their analysis to the configural condition for upright non-matching stimuli.

In a critical discussion of Karmiloff-Smith et al. (2004), Landau and Hoffman (2012) point out that this analysis obscures whether WS participants are truly less sensitive to configural differences in non-matching faces or whether, instead, they are employing a different criterion for how certain they need to be before reporting that two face stimuli are the same. This distinction between *differences in sensitivity* and *differences in decision criteria* can be clarified using signal detection theory.

Signal detection theory is based on the idea that the response someone gives when asked whether they perceive a given property isn't just a matter of their sensitivity to the presence or absence of the property but also depends on what is called their *response bias*. Roughly speaking, a response bias is a feature of decision making that determines how likely we are to judge that a property is present when, as in most cases, there is at least some uncertainty. In some circumstances, it may make sense to adopt a liberal response bias. This would be one in which there is a greater tendency to say the property is present so as to maximize capturing all of the cases in which it is ("hits") even if this means increasing the number of false alarms. A liberal response bias would be warranted, for example, when deciding whether a bump on the ground is a landmine when passing through a minefield. In other circumstances, it may make sense to adopt a conservative response bias. This would be one in which there is a greater tendency to maximize avoiding false alarms even if this means capturing fewer hits. For example, a conservative response bias would be warranted when determining when to initiate the launch sequence for a manned spacecraft when initiating this sequence without certain conditions being met would lead to disaster. Importantly, response bias can vary independently of a person's sensitivity to the presence or absence of the target property. Even if two individuals are equally sensitive to the presence or absence of the property, if they differ regarding their response bias—if one is more liberal and the other more conservative—this would lead to their having different proportions of hits, misses, false alarms, and correct reports that the property isn't present.

Signal detection theory provides a way of separating out how sensitive people are to a signal from how liberal or conservative their response bias is. When Landau and Hoffman applied signal detection theory to Karmiloff-Smith et al.'s data for the non-matching upright configural condition, they found that WS participants had the same level of sensitivity to whether the faces were the same or not as the neurotypical control participants. The reason the WS participants responded differently was because they were adopting a more liberal response bias. Effectively, WS participants were more concerned to make sure that they got it right when the faces were the same than they were to avoid false alarms (i.e., saying "same" for non-matching faces).[8] When the data were adjusted in light of response bias, there was no significant difference between WS participants and neurotypical participants. On the whole, then, Karmiloff-Smith et al.'s (2004)

[8] In fact, their liberal response bias had the odd result of making them *more* accurate than the neurotypical controls when responding to matching faces. Since they were more inclined to report "same" in general, they captured more cases where the faces were the same (achieving 81% accuracy compared to 74% for the control participants).

study also fails to provide good evidence for her central claim that people with WS represent faces in significantly different way than neurotypical individuals do.

What's more, there is also evidence pointing to the opposite conclusion. Consider, for example, the following elegant study by Tager-Flusberg et al. (2003), which has found strong evidence that WS adolescents and adults *do* engage in holistic face processing, just like neurotypical individuals. This study used a part-whole face test procedure in which a sample face is followed by two other stimuli and participants have to determine which of the two matches the sample. In one condition (*whole face*), the sample face is followed by two other faces, while in another condition (*isolated-part*) it is followed by images of individual facial features (e.g., two different noses). The key finding was that WS participants, just like age-matched controls, performed better on the whole face condition but crucially only for upright faces. With inverted stimuli, there was no advantage to using whole faces for the test stimuli. Evidentially the WS participants were not representing the upright faces just in terms of their features. They were representing them in terms of how they are integrated as a whole—a very good indication that people with WS represent faces in the same holistic way as neurotypical individuals.[9]

In short, concept nativism stands up well to the neuroconstructivist critique. The objections regarding genetic determinism, environmental sensitivity, and neural plasticity turn on mischaracterizations of the commitments of rationalists. Neuoconstructivism has no advantage over rationalism regarding any of these concerns. There is also no principled reason to suppose that the argument from neural wiring is undermined by the fact that small initial differences might lead to comprehensive differences in the brains and cognitive capacities of neurotypical individuals and individuals with particular types of genetic disorders. Work in neuroscience has shown that brain development is robust, building the same anatomical structures and the same essential interconnections despite differences in initial conditions and environmental variation. Karmiloff-Smith is of course right that it is in principle possible that small changes in initial conditions could lead to widespread differences later in development in particular cases, which means that there is always the possibility of a neuroconstructivist challenge to some particular instance of a rationalist argument based on neural wiring. However, with this sort of challenge, the devil is in the details, and the burden is firmly on advocates of neuroconstructivism to show this for any particular cognitive capacity.

[9] Neuroimaging studies also suggest a noteworthy correspondence in the brain activity for faces in WS participants and neurotypical controls. The neurotypical response in the ventral visual cortex is for the fusiform face area (the FFA) to respond selectively to faces and with tell-tale signs of face-specific processing—responding less to inverted faces and responding to holistic features of faces only in the upright position (Kanwisher 2010). The FFA has also been found to be selectively responsive to faces in people with WS (Sarpal et al. 2008; O'Hearn et al. 2011).

While we cannot rule out the possibility that a challenge of this type might succeed at some point, we have argued that for the representative example of face recognition in WS—one of Karmiloff-Smith's flagship case studies—her challenge is unsuccessful.[10] Accordingly, the arguments for neuroconstructivism don't provide any grounds for calling into question the use of data involving genetic anomalies in arguing for some form of rationalism or the general case for concept nativism.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0020

[10] It is also perhaps worth reiterating that the rationalist case for any particular domain-specific rationalist mechanism is typically based on multiple complementary types of argument appealing to multiple sources of evidence, and that concept nativism does not purport to identify the rationalist basis for any given trait merely by locating spared abilities that accompany particular types of genetic disorders. Regarding face representation and recognition, we have also presented data that supports at least two further arguments—the argument from early development and the argument from animals (see Chapters 8 and 10). Together with the argument from neural wiring, this makes for a strong complementary package suggesting that face representation is indeed the result of a characteristically rationalist learning mechanism.

21
Perceptual Meaning Analysis

Many empiricists are opposed to theories that postulate any innate concepts, but as we noted in Chapter 2, empiricism (like rationalism) comes in stronger and weaker forms. In this chapter, we want to consider the prospects for accounts of conceptual development that are empiricist but that at the same time draw upon a small number of innate concepts—in other words, we want to examine the types of challenges to concept nativism that are raised by a relatively moderate form of empiricism.

One of the most important accounts of this type is the theory of conceptual development that has been championed by Jean Mandler in her influential book *The Foundations of Mind: Origins of Conceptual Development* and in an important series of related papers (Mandler 1988, 1992, 2004, 2008a, 2012). Since Mandler's theory is a paradigmatic example of such an account, examining its prospects as an alternative to concept nativism will serve to illuminate the prospects for moderate empiricist approaches more generally.

There are two key tenets of Mandler's approach to explaining the origins of children's earliest learned concepts. One is the assumption that the only characteristically rationalist psychological structures that such learning makes use of are drawn from a small stock of innate spatial concepts (i.e., concepts that only have spatial content).[1] The other is the hypothesis that the process that generates these earliest learned concepts is a domain-general process that performs what she calls *Perceptual Meaning Analysis*.

Perceptual Meaning Analysis, as Mandler understands it, is a process that transforms perceptual information into a more abstract, schematic format consisting of *image schemas*, which are analogue or iconic representations that only represent spatial information:

> This mechanism is an attentive process that extracts spatial information from perceptual displays and while retaining its analog character recodes it into a skeletal (somewhat topological-like) form. For example, the infant attends to an apple being put into a bowl, but Perceptual Meaning Analysis outputs something like *thing into container*. Redescriptions like this enable the concept formation that makes conscious thought possible. (Mandler 2008a, p. 212)

[1] Or in other words, for any local rationalism-empiricism debate about children's earliest learned concepts, Mandler holds that each of these concepts traces back to a local acquisition base whose only characteristically rationalist psychological structures are ones drawn from this small stock of spatial concepts.

According to Mandler, this process of redescription allows infants to treat perceptually distinct entities or events as being of a kind and to reason about them accordingly. For example, we saw in Chapter 19 that infants may represent superordinate categories before representing subordinate categories, as when 9-month-olds distinguish animals from vehicles but not dogs from fish in an object examination task (Mandler and McDonough 1993). Perceptual Meaning Analysis accounts for this phenomenon by maintaining that the image schema induced by seeing an animal doesn't retain information about the animal's body plan, whether it has fur or not, or most other perceptual details. All it represents is the spatial equivalent that the entity moves on its own and interacts with other things from a distance, a representation that is composed of the conceptual primitives THING, START PATH, –CONTACT, and LINK (Mandler 2012, p. 429) (we discuss the interpretation of these conceptual primitives below).

What makes Mandler's proposal an empiricist proposal is that Perceptual Meaning Analysis is a completely domain-general process. According to Mandler, only a very small number of relatively concrete (i.e., non-abstract) innate concepts are posited within this framework—all of which are spatial—and no domain-specific rationalist learning mechanisms at all. "Instead, a single domain-general mechanism and a few known attentional proclivities suffice" (2012, p. 445). Among these proclivities is an interest in movement. According to Mandler's account, this leads the mechanism for Perceptual Meaning Analysis to preferentially attend to object motion, which in turn guides the process of concept construction in early development.[2]

Once infants have acquired these image schemas, which redescribe their perceptual experience in purely spatial terms, the image schemas may be enriched in a number of ways (Mandler 2012). One of these is when an image schema is associated with a feeling and is used to interpret that feeling. For example, Mandler suggests that the feeling of pressure or resistance that accompanies pushing an object may become integrated with, and interpreted in terms of, one or more spatial representations, such that the combination of the feeling and spatial representations constitutes an initial representation of causation. This may then give rise to a conceptualization of causation in which certain entities are seen as possessing a force that accounts for their ability to initiate motion or to cause motion in other entities. Analogy is another source of representational enrichment

[2] We should note that although Mandler self-identifies as an empiricist in some of her writings (e.g., 2004, p. 61), in other writings she rejects the label and describes her position as "as a compromise between nativist and empiricist views" (2012, p. 443). While Mandler's view is less empiricist than some, we still consider it to be an empiricist view. Nothing really turns on the label, however, as our critique in this chapter can be seen as arguing that regardless of how her view is characterized, it is not rationalist enough. It is also worth noting that Mandler's image schemas will be counted as concepts on some but not all of the ways of drawing the conceptual/nonconceptual distinction (see Chapter 6 for different ways of drawing this distinction).

suggested by Mandler. She uses analogy to explain how spatial representations come to represent time along the lines we discussed previously in Chapter 12.

Mandler takes the idea that children's earliest learned concepts are spatial in nature to be a nearly inevitable consequence of the fact that what matters most about an object is its interactions with other objects. "What things do is the *core* of their meaning", hence infants should concentrate "on the way things move and how they move vis-à-vis each other" (Mandler 2004, p. 87). For Mandler, this fact is nicely accommodated by image schemas that represent the spatial relations among objects as events unfold, for instance, one object touching another (+CONTACT), the trajectory that an object takes when moving towards or away from another object (PATH), or the movements of objects interacting with one another (LINK).

We agree with Mandler about the importance of "what things do". But this is in no way incompatible with a rationalist perspective on conceptual development. In this chapter we argue that even a moderately empiricist framework like Mandler's has a number of major shortcomings that show why it is necessary to postulate concept nativism's richer acquisition base involving further characteristically rationalist psychological structures contributing to domain-specific rationalist learning mechanisms and (in all likelihood) a considerably larger stock of innate concepts. We will organize our discussion into three parts. We will start by looking at the conceptual primitives that are postulated in the Perceptual Meaning Analysis framework. Next we will look at the initial period of conceptual development and the types of concepts that are taken to be obtained simply from Perceptual Meaning Analysis without enrichment. Lastly, we will look at Mandler's proposal that these spatial representations are enriched in later development through their association with bodily feelings. As we proceed, it will help to have a way of referring to the period of development in which infants are supposed to only have concepts that are a product of Perceptual Meaning Analysis. We will refer to infants in this earlier stage of development as *younger infants* and infants who can take advantage of processes of enrichment as *older infants*. Mandler's estimate for when enrichment processes begin to take place is after the first six or seven months of life (Mandler 2012; Mandler and Cánovas 2014).

Let's begin by examining Mandler's claim that her conceptual primitives are purely spatial representations. We will focus on one of her major statements of her view, which offers the following list of primitives (Mandler 2012, p. 427):

PATH	THING ±MOTION
START PATH	BLOCKED MOTION
END PATH	±CONTACT
PATH TO	LOCATION
LINK	MOVE (BEHIND)
CONTAINER	MOVE OUT OF SIGHT (–SEEN)
(IN)TO	MOVE INTO SIGHT (SEEN)
(OUT)OF	

Although these are presented using familiar words and symbolic conventions, it is essential to remember that this is only meant to be for the convenience of the reader and that we shouldn't take them to have more content than her theory says they have. It is important that these representations are supposed to have spatial content and *only* spatial content—nothing more. For some, their spatial content is relatively unproblematic. LOCATION, for example, simply refers to spatial location. But for others, there are serious questions about whether the restriction to purely spatial content can be maintained.

Take the representation THING. As we noted above, THING is one of the primitives that Mandler takes to form the basis for the early concept of an animal (THING, START PATH, –CONTACT, and LINK), which she describes as the concept of a "self-starting interactor" (2012, p. 429). Likewise, THING is one of the primitives that forms the basis for the early concept of an inanimate entity (THING, –MOVE, and –LINK), which is intended to pick out entities that are not self-starting and don't interact with other things in the way that animates do. But what exactly is the content of THING in these and similar image schemas? Mandler says that "THING refers to any perceptually bounded cohesive object" (2012, p. 429). In other words, THING is supposed to function as an early object concept, picking out entities that maintain their boundaries and connectedness. On this account, a wooden block is a thing. When it moves, it holds together. If someone grabs it from one end, the whole block moves. In contrast, a pile of sand isn't a thing. If someone tosses it into the air, its boundaries quickly dissipate, with sand scattering in different directions.

The point we want to highlight about THING is that, while it is certainly true that objects are located in space, it is doubtful that even the earliest concept of an object is purely spatial. At the very least, the core representation of an object as an entity that retains its boundaries and cohesiveness isn't just about how such entities *are* behaving (or have behaved). Infants have associated expectations about how an object *could* behave or how it *would* behave in different circumstances, which cannot be understood simply in terms of representations of the object being in a particular spatial location. Consider once more how a paradigmatic object like a wooden block is represented. In seeing it as an object, there is an expectation that its boundaries will move together and that it won't crumble or break apart when it is pushed or pulled, an expectation that exists even if the block is *not* pushed or pulled and remains perfectly stationary. The expectation goes beyond what it is perceived as actually doing; it concerns how it would behave under these other sorts of conditions. It is also an expectation that appears in early infancy. When shown novel static three-dimensional objects, 3-month-olds look longer if only a portion of an object moves when it is grabbed from the top (Spelke et al. 1993).

Younger infants also represent non-spatial properties of objects and can use this information to individuate and identify objects. For example, 4-month-olds

can use functional information to individuate and identify simple tools (Stavans and Baillargeon 2018). In a standard *individuation and identity task*, infants see an object appear from behind a screen only to return behind it, and then another object with different properties appears from behind the screen and returns behind it. Then the screen is lowered revealing either one or two objects. If infants look longer at the one-object outcome, this shows that they expected two objects and that they were able to use the objects' different properties to individuate them as distinct objects and to keep track of their locations. In the experiment using functional properties of objects, 4-month-olds were presented with a variant on the individuation and identity task in which the first object that appeared from behind the screen had one functional property (e.g., infants had previously seen it used to mash sponges) and the second had another (e.g., infants had previously seen it used to pick up sponges). In a control condition, infants saw the same two objects undergo similar motions as in the experimental condition but with no indication of their functions. In both conditions, they were tested on an outcome in which just one object appeared when the screen was lowered. The result was that the infants looked longer at this outcome in the experimental condition than in the control condition, indicating that they inferred the presence of the two objects when representing them as having distinct functional properties.

Now consider for a moment the implications for Mandler's proposal that the content of THING (i.e., OBJECT) is purely spatial. One problem is simply that the properties that are used to individuate these objects aren't spatial—they are functional.[3] What matters to infants' representing the masher as one object and the tongs as another isn't their shape or even how they move, since these features were the same in both the (functional) experimental condition and the (non-functional) control condition. The critical property is what the objects are used to do. Another problem is that these functional properties aren't about currently visible movements. They are about what the objects have done and what they can or should do—none of which are *spatial* properties. So it is hard to see how Mandler's purely spatial understanding of THING can do justice to younger infants' representation of an object.[4]

What about the other primitives that Mandler posits? Motion is through space, so MOTION is spatial. But arguably to represent certain aspects of motion, one also needs to represent temporal properties and relations. Consider, for example, what Mandler says about INTO: "the primitive INTO is represented by an image-schema of something moving into the opening of an otherwise closed shape" (2012, p. 427). But if the image schema doesn't represent the temporal structure of

[3] Younger infants can individuate objects using other non-spatial properties too. For example, they can infer that two objects are present when they hear two distinctive sounds produced from behind a barrier, relying on their representation of the non-spatial property of *being capable of making such-and-such sound* (Wilcox et al. 2006).

[4] See also the discussion of how infants represent and reason about objects in Chapters 15 and 17.

this type of event, how can it represent the moving—a change in location over time—as opposed to the position of the object? In other words, how does the image schema framework distinguish between INTO and IN?

Notice that it doesn't help to take the representational vehicle for *into* to be something that has temporal structure, like a mental video clip as opposed to a static image.[5] This is because its temporal aspects would still need to be interpreted as such and interpreted correctly. Suppose that an image schema "video" of this sort begins with an object positioned outside of a container and ends with the object inside the container. It might seem natural to construe this as the object *moving into* the container. But that is only because we interpret the temporal properties of the image schema in temporal terms and rely on the convention in which the temporal structure of the represented event matches the temporal structure of the representational vehicle. A system that is entirely unable to represent temporal properties or that isn't sensitive to this convention would have no way of distinguishing the situation in which the object starts outside of the container and ends up inside the container from the situation in which the object starts inside of the container and ends up outside of the container.[6]

Of course, a purely spatial representation of the sort that Mandler has in mind could be interpreted to mean *into* by a suitably equipped cognitive system. The same is true of any representational vehicle if treated as an arbitrary discursive symbol, like a word in natural language. But Mandler's conceptual primitives aren't supposed to be discursive symbols that can have any type of content. They are supposed to be analogue representations whose interpretation is restricted to purely spatial properties and relations. So it looks like her primitives for various types of motion (e.g., START PATH, END PATH, PATH TO, (IN)TO, and (OUT)OF) are problematic too. Either they have purely spatial content and can't represent what they purport to, or else they incorporate temporal content and Mandler's restrictions on the content of her primitives isn't being enforced.

The same pattern holds for most of Mandler's other primitives. It should be obvious that MOVE OUT OF SIGHT (–SEEN) and MOVE INTO SIGHT (+SEEN) are not purely spatial, since they build in the idea of *being seen / not being seen*.[7] Or take LINK. This too covertly relies on temporal and modal forms of representation. Mandler says that "LINK refers to a variety of contingent interactions between objects as when a hand picks an object up, back and forth interactions of people,

[5] Representations can be understood in terms of what they represent and in terms of the item that does the representing. By *representational vehicle*, we mean the latter.

[6] As Wittgenstein (1953) argued with respect to arrow symbols, a symbol such as "⇒" does not intrinsically point to the right, but only does so for agents who know the convention to interpret it this way. A society of intelligent aliens might instead adopt the convention of interpreting arrows as pointing *away* from the direction of the arrow head rather than towards them, in which case the symbol "⇒" would be used to point to the *left*, rather than the right.

[7] Mandler acknowledges this exception, but still seems to view ±seen as a Perceptual Meaning Analysis primitive. See Mandler (2012, pp. 427–428).

or between paths as when one object chases another" (2012, p. 429). But to represent one object as chasing another, where the movements of one is contingent on the movements of the other, it is not enough to represent them as traversing similar paths. The chaser's path must be represented as responsive to the target's, such that a change in the chaser's direction typically follows (i.e., temporally follows) a change in the target's direction. Moreover, the chaser's path needs to be conceptualized as being dependent on the target's path to understand the contingency of the interaction—for example, to understand that even though the chaser's path happens to be straight at the moment, it would veer to the right if the target were to move in that direction. But again, such properties aren't spatial properties, so LINK, as Mander uses it, turns out not to be purely spatial either.

LINK is also problematic in a different way. A cognitive system employing a conceptual primitive has to interpret it in a consistent manner. It can't have a menu of possible meanings that we as theorists who are external to the system pick and choose from in order to make the interpretation fit the case at hand. But this is what seems to be happening with Mandler's characterization of LINK's content. If LINK truly has the broad interpretation Mandler says it has—involving a variety of different types of contingent interactions—then when infants witness a hand picking up a toy, they would have to understand this as meaning that two objects are in some relation of contingent interaction that encompasses events as diverse as grasping, chasing, and engaging in a game of peekaboo (all of which she takes to be represented with LINK). In other words, infants should represent and reason about grasping, chasing, and peekaboo in essentially the same way, which clearly isn't the case.

We needn't work through the rest of Mandler's proposed primitives. It should be evident by now that the sorts of problems we have noted are general enough to raise significant concerns about the extent to which Mandler is actually starting from primitives that are purely spatial. Let's move on to Mandler's claim that younger infants' concepts are derived from Perceptual Meaning Analysis and hence constructed just from her proposed primitives. If it could be shown that the concepts that infants have in this stage of development go well beyond those constructible from a set of primitives of the type her account is based on, this would present another major challenge to the Perceptual Meaning Analysis framework. We will argue that this is what a proper analysis of the data show. Younger infants have many concepts they couldn't possibly have acquired via Perceptual Meaning Analysis.

Our discussion will focus on one of Mandler's flagship examples of a concept that is supposed to be learned via Perceptual Meaning Analysis: infants' initial concept of a goal. According to Mandler, younger infants learn to represent goal-directed behaviour without a mentalistic concept of a goal, using instead a concept with purely spatial content:

> the first concept of goal is not derived from ascribing intentionality to an animate agent, but rather from two types of spatial patterns of action. These patterns consist of combinations of the primitives of START PATH, PATH TO, END OF PATH, and LINK. A self-starting object that moves on a direct PATH TO an object at its end and links with it gives rise to a concept of "move to." Similarly, an object that goes on different but linked PATHS TO (linked in that they are contingent on going around a blockage in a direct path) to a particular end point or object also gives rise to a concept of "move to." (2012, p. 435)

In one of the standard experimental paradigms for determining whether infants represent goals, infants view a scene where two toys are on different sides of a stage and an agent, with only the agent's arm visible to the infant, directly reaches for and grasps the same one of these toys a number of times in a row. Following this, the toys' locations are switched and the agent either continues to reach for and grasp the same toy as before (old object, new location) or reaches for and grasps the other toy (new object, old location). Infants look longer when the agent reaches for the new toy, even though this involves reaching for the same old location. This shows that they had interpreted the initial reaching events as having the goal of reaching for the old toy, and so they look longer when the agent changes goals and reaches for the new toy (Woodward 1998, 1999). Further studies have obtained similar results with infants as young as 3 months old (Luo 2011; Choi et al. 2018; Woo et al. 2021).

How could proponents of Perceptual Meaning Analysis explain these sorts of findings? They would have to claim something like the following: Infants extract a representation in which the agent's hand moves on a direct path that ends with the old toy and with the hand and toy being linked, and this creates the expectation that the hand will continue to move on similar paths, linking with this same object in different locations.[8]

Unfortunately, the situation is far more complicated than can be captured by this or any other purely spatial description. First of all, many of these studies employ a type of control condition where the agent involved in the goal-directed movement undergoes exactly the same pattern of movement as in the experimental condition, but due to contextual factors in the control condition, the infants observing these same movements form different expectations. In this type of control condition, infants are familiarized to events in which there is just a single object present, in contrast with the two objects in the experimental condition. So the infants still see the agent's hand (or in some experiments, the whole agent)

[8] Notice that this analysis helps itself to the representation of *distinct paths* and *different locations*. These would seem to require representations for quantifiers and non-identity and consequently representations that aren't spatial or fully analogue.

move repeatedly to one object in the familiarization trials. Indeed, the only difference between the experimental condition and the control condition is that, in the experimental condition, there is a second object nearby. The test trials are also the same in both conditions. The agent, the object that the agent's movement is directed towards, and the accompanying object are all present, and the agent moves towards one or the other of these objects. But in the control condition, infants no longer expect that the agent will continue to move towards the object it had repeatedly moved towards in the familiarization trials, contrary to the prediction made by Mandler's account.

Why do infants respond differently in these two conditions? The most plausible explanation is that infants are representing these situations not in terms of a goal conceptualized merely as a distinctive pattern of movement, but rather in terms of the agent manifesting a preference, or a disposition to pursue a particular goal (Luo and Choi 2012). In the one-object familiarization control condition, there simply isn't enough information to infer such a preference. The agent might be approaching the solitary object simply because it is the only one that is available. In contrast, in the two-object familiarization condition, the fact that the agent repeatedly moves towards one and not the other object is good evidence that this is a preference. Three-month-olds evidentially appreciate this difference. Thus they must be interpreting these events not by focusing exclusively on their spatial properties but by taking into account contextual factors that can be used to interpret the psychological meaning of the motion—whether the agent is exercising a preference. This suggests that children at this age have far richer conceptual resources than Mandler's purely spatial concepts.

In fact, the sensitivity to context is even more subtle than we have so far described. Three-month-old infants can observe an agent repeatedly grasping one object while a second object is present during familiarization trials and still not form the expectation that the agent will continue to selectively interact with the first object. This happens if infants can see that the second object is obscured from the view of the agent during the familiarization trials (Kim and Song 2015; Choi et al. 2018). To explain this further level of context sensitivity, we need to attribute to infants the ability to infer a disposition to pursue a goal in a way that takes into account the agent's visual perspective. Clearly this can't be expressed in the austere terms of a purely spatial system of representation.

These studies demonstrate that the patterns of movement that Mandler takes to be constitutive of the early representation of a goal are not sufficient for infants to attribute a goal to an agent. It turns out that they are not necessary either. Consider again the studies of helping and hindering discussed in Chapter 18. In this work, it is evident that infants represent the goals of the agents who are helped or hindered. If they didn't, they wouldn't interpret the helpers as helpers and the hinderers as hinderers. But crucially the infants infer the agents' goals

without Mandler's *direct motion towards an object* or *repeated movements towards an object via different paths*. For example, the goal in the case of the Climber climbing the hill wasn't to approach an object. It was to get to the top of the hill (Hamlin et al. 2007). The agents in this and related scenarios don't have goals that are inferable from their directly moving towards an object that they make contact with, or from their repeatedly moving via different paths towards an object that they make contact with. They don't move in these ways, and even if they did, this would not pick out the goal that the infants attribute to them.

Mandler's incorporation of *self-starting motion* into the early concept of a goal isn't necessary either. Studies where it is a human agent who has the goal typically display only a restricted portion of the human (e.g., an arm reaching into the scene from offstage), so infants aren't in a position to use Mandler's criteria to determine whether the agent's movement is self-propelled or due to contact with another object. And yet infants as young as 3 months old attribute goals to the agents in these circumstances (Choi et al. 2018). This strongly suggests that they have a concept for human agents and that they automatically treat humans as having goals, something that they don't do for novel non-human entities that move in a similar manner with the same ambiguity as to whether they are self-propelled (e.g., Woodward 1998; Luo and Baillargeon 2005).

What all this shows is that the patterns of movement that Mandler associates with the early representation of a goal are not constitutive of young infants' early concept of a goal; rather, they are merely taken to be *evidence* of the presence of a goal by infants. When infants respond to these patterns, they are relying on a far more abstract concept, one whose content captures that the agent is trying to bring about a state of affairs it wants to achieve. This problem with Perceptual Meaning Analysis isn't about the particular types of motion that Mandler has proposed for the early representation of a goal. The same difficulties would arise for any similar proposed patterns of motion, or any analysis that is restricted to spatial content. Goal-directedness simply isn't a matter of moving in a way that can be couched in spatial terms, nor is it understood in such terms by younger infants. They have rich expectations that can only be explained by attributing to them such concepts as AGENT, GOAL, PREFERENCE, TRYING (i.e., trying to achieve a goal whether it is achieved or not), BEING ASSISTED IN ACHIEVING A GOAL, BEING HINDERED IN ACHIEVING A GOAL, BEING IN A POSITION TO PERCEIVE AN OBJECT, and so on, and by attributing to them the capacity to view different types of behaviour as defeasible evidence regarding the applicability of these concepts.

We've gone through this example in some detail to show how Perceptual Meaning Analysis fails to do justice to even a flagship case that is supposed to illustrate this account's ability to explain how infants acquire their earliest learned concepts. There is nothing particularly unusual about this one example, however. The other case studies that Mandler discusses have similar problems. And there are many concepts that younger infants have been shown to have that Mandler

either doesn't discuss in detail or doesn't consider.[9] For example, younger infants differentiate objects (blocks, balls) from stuffs (sand, milk) and have different expectations regarding their behaviour (Hespos et al. 2016). They see the world in numerical terms (see Chapter 8) and can perform numerical comparisons for items as diverse as objects and actions (Kobayashi et al. 2004; Wood & Spelke 2005). And they are also able to track and enumerate groups of things (e.g., dots moving in groups) (Wynn et al. 2002). None of these concepts—STUFFS, GROUPS, ACTIONS, EVENTS, or concepts for numerical quantities—can be constructed from Mandler's spatial primitives. The inferences that they support are also a testament to how abstract younger infant's ways of conceptualizing the world can be. For example, the expectations that 5-month-olds have for solid objects but not for stuffs hold even when the solids and stuffs exhibit the same perceptual boundaries (Hespos et al. 2016).

Even more challenging for the Perceptual Meaning Analysis framework is to account for the myriad ways in which younger infants see the world in social terms. For example, as we saw in Chapter 9, they take language to be communicative (Vouloumanos et al. 2014). They also differentiate linguistic from non-linguistic sounds and use language as a guide to categorization (Ferry et al. 2010), treat melodies as having social meaning (Mehr et al. 2016), represent social affiliation using imitation as a cue for social categorization (Powell and Spelke 2018), and can infer which of two individuals is socially dominant by taking into account both the number of their allies and whether their allies are in a position to witness the conflict (Pun et al. 2022). Needless to say, it is difficult to imagine how the sorts of representations involved in these cases could reduce to a combination of purely spatial representations.[10]

[9] This may be because, at the time that she was writing, she didn't think there was enough data to establish that younger infants have the concepts in question. In Mandler (2008b) she says, "insofar as we have data, I know of no concepts that preverbal infants form that are not within image-schemas' scope. However, this research area is wide open with new data coming in all the time" (p. 273). This is certainly true, and our own general sense of the trend in infancy research is that, as researchers continue to devise more sophisticated experimental methods—particularly ones that simplify the demands on infants' memory and attention—infants reveal themselves to have far richer conceptual capacities than researchers were previously willing to entertain.

[10] Another challenge that Mandler's account faces concerns the explanation of how the primitives involved in image schema combine. The problem is not that analogue representations cannot combine, but rather that the type of combination required appears to involve predication and other semantic resources that fall outside the scope of purely analogue representations as opposed to the type of discursive symbols which Mandler's system is designed to avoid (e.g., quantification and negation). To represent that one of two objects moves into contact with a third object, for example, one needs several instances of the image schema THING—THING-1, THING-2, THING-3—where the things referred to are understood as sharing the property of being things, but as being distinct from one another, and where featural properties associated with other image schemas are predicatively linked to some of these but not others (e.g., THING-1 stands in the CONTACT relation to THING-2, but not to THING-3). To represent this, a purely analogue representation would need to be interpreted by a conceptually richer system, or would need to be augmented with word-like labels (see also footnote 8 above).

Let's turn now to the final part of Mandler's account, the part about conceptual enrichment. As we explained earlier, Mandler takes infants' first concepts to have exclusively spatial content but thinks that older infants and children can acquire concepts that aren't just spatial through processes of conceptual enrichment. The earliest form of conceptual enrichment that is supposed to take place is one in which bodily feelings become associated with an image schema. To illustrate some of the difficulties with this aspect of Mandler's theory of development, we will focus on her central example of conceptual enrichment: the development of an enriched concept of causation (Mandler 2012, pp. 432–434). Mandler holds that there is no innate abstract concept of causation that incorporates the idea that a physical cause has a force or power that brings about its effect. Instead, she offers a two-part theory of how infants come to learn that "force is needed to cause objects to move" (2012, p. 432).

The first part of her theory is an account of an important finding by Alan Leslie and Stephanie Keeble, who showed that 6-month-olds distinguish launching events (which, as noted in Chapter 12, look causal to adults) from non-launching events (which don't) (Leslie 1982; Leslie and Keeble 1987).[11] A paradigmatic launching event, you may recall, is one where an object A moves towards object B, stops at the point of contact, and B immediately (with no time lag) moves off in the same direction A had been heading in. A non-launching event is similar except that there is a spatial or temporal gap (i.e., A doesn't contact B before B moves, or if it does, there is a pause before B moves). Leslie and Keeble showed infants animations of a launching event or a non-launching event repeatedly. After the infants habituated to these events, the animations were reversed (that is, played backwards), and the infants' looking times were measured to determine whether their recovery of attention differed across these conditions. Notice that when the animations are reversed, there is an equal amount of change in the spatial-temporal properties of the events. But there is a crucial asymmetry nonetheless if the launching event is represented causally. If the launching event is represented causally, then when it is reversed, the cause and the effect also change—now B causes A to move, rather than A causing B to move; by contrast, when the non-launching event is reversed, the only change is the direction of movement of the objects.

What Leslie and Keeble found was that infants recovered attention far more to the reversal of the launching event, suggesting that they represented this event in causal terms.[12] Mandler's explanation of what is going on with infants' response to the launching event is that it creates a percept "of the motion of one object being

[11] See also Mascalzoni et al. (2013) for related findings with newborns.

[12] Leslie (1994) goes on to argue that this causal interpretation triggers an innate domain-specific mechanism that represents the forces involved in mechanical events. He calls this "ToBy"—short for "Theory of Body mechanism".

transferred into another" (2012, p. 432).[13] With the launching event, it's not just the direction of motion that changes, but also which object the motion is transferred to. Mandler claims that this doesn't involve the representation of a force being exerted, just the perception of a pattern of motion. As a result, Perceptual Meaning Analysis "can use the primitives of MOTION INTO to begin a conceptual interpretation" of these sorts of experiences (2012, p. 432).

The second part of Mandler's theory is that the understanding that force is involved in one object causing another to move emerges later as MOTION INTO becomes associated with the right sorts of internal feelings. This happens "only when infants begin to move themselves around and engage in the behaviors that result in feelings of force being applied to objects. As infants begin to push against things, they also begin to experience for the first time strong feelings of pressure and resistance" (2012, p. 432). The result is that "the existing spatial concepts enable the infants to interpret the bodily feeling" such that "a representation of the relevant situation can activate a remnant of the feeling at the time, and vice versa" (2012, p. 433). In other words, children learn to associate MOTION INTO with these internal feelings in their own interactions with objects and then generalize the association so that seeing something that is conceptualized with MOVE INTO equally activates these feelings.

Mandler's account of the enriched concept of causation is similar to White's (2009, 2012) account, which we discussed in Chapter 12. Since Mandler's account faces the same sorts of problems that we previously raised for White, we can be brief here. Mandler's account, like White's, faces the problem of initial representational access and hence what we called *the why problem* and *the how problem*.

In this case, the why problem asks why an infant would even form a general association between the critical internal feelings and the range of events in which one object causes another to move. Suppose for the sake of argument that Perceptual Meaning Analysis does produce a purely spatial representation of the movements that correlate with causal interactions (MOVE INTO). Still, why would infants form the needed association with the representation of these movements? First of all, most instances of the motions covered by MOVE INTO—all of the ones they passively observe—are *not* associated with any of these feelings. When an infant watches as one object hits another and causes it to move, the infant doesn't experience any feelings of pressure, resistance, or effort. So there are a massive number of counterexamples to the general association they are supposed to acquire. Second, even in the infant's own case these feelings are very inconsistently correlated with causation. Many instances of causation involve little

[13] The idea that what is seen in a launching event is a transfer of motion goes back to Michotte (1946/1963), who referred to this transfer as *ampliation of movement*. There are interesting questions about what is meant by seeing a transfer of motion as opposed to seeing one object causing another to move, but our main objections to Mandler won't turn on this issue.

noticeable effort, and infants would regularly experience feelings of pressure and resistance when they are not trying to cause a transfer of motion, for example, just from feeling the surfaces they are resting on. Taken together these two points underscore the fact that it seems highly unlikely that a general-purpose learning mechanism with no domain-specific biases to guide it would have the needed data to form the general association that Mandler's theory builds on.

Now the how question. Even if infants were somehow to develop a strong general association between the internal feelings and the pattern of movement, how would this constitute an understanding of causation in terms of forces? Why would this association amount to the conceptualization of a force being imparted rather than two things—a movement and a feeling—just occurring together? As we pointed out in connection with White's developmental theory, it is certainly possible to represent a cause-effect sequence as *merely* a sequence, for example, to represent a sensation of pressure when a toy is contacted followed by seeing the toy topple over. But to be an account of where an understanding of force comes from, something more is needed.

As adults, we naturally interpret the feeling involved as an instance of trying to make something happen. The problem Mandler faces is in explaining how we come to interpret the feeling in this way. Note that there are only two possibilities here. Either it is an intrinsic feature of the feeling that it is taken to be causal (that is, simply having the feeling automatically brings with it the sense that it is causal), or else one can have the feeling independently of this interpretation but by adulthood it is habitually interpreted to be causal. In the first case, infants would already have to have some way of representing causation simply in order to experience feelings of effort—the idea of causal force would be an essential constituent of the feeling—so feelings of effort could not explain how purely spatial representations come to be representations of causation. And in the second case, infants would have to have some way of representing causal force that is independent of the feeling and that is used to interpret such feelings in a causal way. (Without this, linking an uninterpreted feeling to MOVE TO is not going to generate the sense that one object causally acts on the other.) But then, of course, we would need an account of the origins of this independent way of representing causal force. So, either way, the enrichment account cannot explain the origins of the concept CAUSE. As we pointed out with White, the way around this problem is to suppose that infants have access to an innate concept CAUSE (or FORCE) that is part of a force dynamics mental model. Such a model would enable an infant to interpret the toy as an agonist and the object that contacts it (and that dislodges the agonist) as an antagonist.

Mandler's notion of conceptual enrichment is somewhat open-ended, as it also includes conceptual change that is supported by language and analogical reasoning. More would need to be said about these further types of conceptual enrichment if our focus were on Mandler's full developmental account. However, if we are just sticking to conceptual development in infancy, the main source of

enrichment is enrichment via learned associations with bodily feelings. And it is clear that Perceptual Meaning Analysis plus this type of enrichment is in no better position to explain many other conceptual capacities in infancy than it is to explain acquisition of CAUSE. For example, we saw earlier that infants can individuate and track objects by representing their functional properties. But it is hard to see how associating some type of bodily feeling with a spatial substitute for an object's function would allow infants to understand what a functional property is and that an object can possess this property whether or not it is currently undergoing any motion. Likewise, infants understand the abstract relations of possessing, giving, and taking (Tatone et al. 2015; Tatone et al. 2019), but these aren't constituted by any particular pattern of movement, and it is not at all clear how associated bodily feelings could provide the missing content.

There is also a whole range of examples that exceed the resources of Mandler's Perceptual Meaning Analysis account (enrichment processes included) that have to do with infants' understanding of social phenomena. For example, infants have a sophisticated understanding of social dominance. They expect smaller individuals to defer to larger individuals and, as we noted earlier, expect individuals with fewer allies to defer to individuals with more allies (Thomsen et al. 2011; Pun et al. 2016). They appreciate that when one individual loses to another in a zero-sum conflict (controlling territory), they will defer in a physically different type of conflict (obtaining a desired object) (Mascaro and Csibra 2012). Incredibly, infants not only represent dominance relations but also, like adults, distinguish between dominance based on fear and dominance based on respect.[14] When obedience is based on fear as opposed to respect, subordinates often disobey an order if the dominant individual who issues the order isn't present. When a dominant individual tells a group to go to bed, 21-month-old infants look longer if the group disobeys the order after she leaves the scene but only if her dominance has been shown to be based on respect (Margoni et al. 2018; see also Thomas et al. 2018). Here too it is not at all clear how the concepts in question—e.g., FEAR-BASED DOMINANCE, RESPECT-BASED DOMINANCE, and OBEDIENCE—can be acquired by combining spatial content with simple, unconceptualised bodily feelings.

Or consider concepts of social groups and group affiliation. A variety of experiments have shown that infants in their second year of life draw inferences about an agent's behaviour towards others based on their social group affiliations and, in particular, that infants expect an agent to behave differently towards in-group and out-group members. In one study, 16-month-olds witnessed two groups in which there was within-pair cooperation and some between-pair conflict. Later, when shown individuals from these different groups who hadn't previously interacted, they were surprised to see them cooperate with one another (Rhodes et al.

[14] In the literature on mechanisms of cultural transmission, these two forms of social power have been referred to as *dominance* (for fear-based dominance) and *prestige* (for respect-based dominance). See Henrich and Gil-White (2001) for more on this distinction.

2015). In another study, infants as young as 13 months old saw three agents interact—a bystander, a victim, and a wrongdoer (whose bad deed was to steal a toy from the victim). The infants subsequently expected the bystander to refrain from helping the wrongdoer—but only if the bystander and the victim belonged to the same group (Ting et al. 2019). These and other findings clearly go beyond the resources of Mandler's Perceptual Meaning Analysis, even with enrichment. The concepts SOCIAL GROUP, IN-GROUP, OUT-GROUP, and so on are a long way from purely spatial concepts or spatial concepts that are supplemented with associated bodily feelings.

To sum up, Mandler's theory of conceptual development is of special interest because, although it is clearly an empiricist theory, it aims to make use of a crucial (albeit relatively small and relatively concrete/non-abstract) stock of innate concepts, making it a moderate form of empiricism. We have argued that this account has a number of shortcomings that indicate the need for a richer acquisition base. One issue is that her conceptual primitives are presented as if they are far more austere than they really are, in that they aren't actually restricted to purely spatial content, as required by the Perceptual Meaning Analysis framework. But even if we put this issue aside, a further issue remains, namely that combinations of these primitives can't account for the conceptual capacities of younger infants (e.g., their understanding of goals). And adding associated bodily feelings is not enough to account for infants' acquisition of concepts that Mandler agrees aren't exclusively spatial (e.g., concepts like FORCE). So the Perceptual Meaning Analysis framework doesn't amount to a viable alternative to concept nativism.

Of course, none of this means that spatial concepts aren't important and that new spatial concepts can't be learned in part through constructions from innate spatial primitives. But we have argued in this chapter that it would be wrong to suppose that exclusively spatial content is the cornerstone of early conceptual development. If we are going to account for the complex, context-sensitive nature of infants' inferences, we need to attribute to them far richer concepts with the recognition that infants may use spatial information as evidence about the applicability of these concepts but that these concepts aren't themselves fundamentally spatial representations. While other moderate empiricist accounts with different highly limited sets of innate primitives and different domain-general learning mechanisms remain possible, Mandler's account and the difficulties it faces are indicative of the limitations of such accounts more generally. Empiricist accounts, even moderately empiricist accounts like Mandler's, simply do not have the resources needed to explain conceptual development. What the arguments in this chapter indicate is that a richer acquisition base—entailing some form of concept nativism—is needed.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0021

22
Embodied Cognition

Our final chapter in Part III focuses on the empiricist challenge stemming from the theoretical framework of embodied cognition. It would be hard to overstate the influence and impact of embodied cognition on philosophy and cognitive science. Many leading philosophers and cognitive scientists consider themselves to be embodied cognition theorists, and a great many more have strong sympathies with embodied cognition. And one of the key advantages of embodied cognition for many of its proponents is that it is seen as vindicating an empiricist picture of the mind by offering a novel and highly successful alternative to rationalist accounts of cognitive and conceptual development. Accordingly, no discussion of empiricist challenges to concept nativism would be complete without an examination of embodied cognition.

The term "embodied cognition" refers to a loose affiliation of research programmes in philosophy and cognitive science that are also widely seen as involving a radical departure from conventional accounts of cognition. Although proponents of embodied cognition generally describe their approach as drawing attention to the importance of the body to understanding the mind, this emphasis on the body can take a number of different forms (Shapiro 2019). Our discussion will address three important forms that are especially relevant to the evaluation of concept nativism. These are (1) the view that concept acquisition varies with a learner's body type, (2) the view that human action can often be explained without postulating rich internal representations and representational processes, and (3) the view that concepts are realized in sensorimotor and affective systems and shouldn't be understood as amodal representations. Our discussion here will again give us the opportunity to address further conceptual domains beyond those we have discussed so far.

Let's begin with the view that concept acquisition varies with body type. Body type in this context may refer to the sorts of major differences that appear across species (e.g., the body plan of a quadrupedal mammal versus the body plan of a limbless reptile) or to smaller differences among individuals of the same species (e.g., whether a person is left or right handed). Regardless, the claim by many proponents of embodied cognition is that these bodily differences lead to different types of experiences and that these differences in experience can have a

significant impact on the types of concepts that can be acquired. We will refer to this strand in embodied cognition as the *different-body/different-concepts* hypothesis.[1]

A standard example that has been used to illustrate the different-body/different-concepts hypothesis is the representation of space:

> The concepts *front* and *back* are body-based. They make sense only for beings with fronts and backs. If all beings on this planet were uniform stationary spheres floating in some medium and perceiving equally in all directions, they would have no concepts of *front* or *back*. But we are not like this at all. Our bodies are symmetric in some ways and not in others. We have faces and move in the direction in which we see. Our bodies define a set of fundamental spatial orientations that we use not only in orienting ourselves, but in perceiving the relationship of one object to another. (Lakoff and Johnson 1999, p. 34)

In this passage, Lakoff and Johnson note the tight connection between the type of body that an animal possesses and the spatial concepts that it comes to acquire. These concepts should be relevant to the particularities of the animal's sensory faculties and to its manner of locomotion. Since these, in turn, depend on features of the animal's body, spatial concepts are "body based".

Proponents of embodied cognition have drawn two conclusions from this sort of example. The first is that we should expect conceptual variation to accompany bodily variation. In particular, if spatial concepts are body based, organisms with different types of bodies should exhibit corresponding differences in how they represent space. Space itself doesn't dictate the types of spatial representations that a mind comes to possess; spatial representation is all about the contingent fit between a body and its environment. Second, body-based representations aren't supposed to be innate or dependent on domain-specific rationalist learning mechanisms. Since they are tuned to the way that an animal's body interacts with its environment, they should form as an animal begins to move in its environment and in response to the characteristic types of motion that its body and environment permit.[2]

[1] Shapiro (2019) refers to this view as the *conceptualization hypothesis*, according to which "the kind of body an organism possesses constrains or determines the concepts it can acquire" (p. 80). See also Casasanto (2014) for discussion of the related *body-relativity hypothesis*, according to which different types of bodies lead to different ways of thinking.

[2] The combination of cognitive variation and learning through body-environment interaction is a major theme in the embodied cognition literature. For example, in the opening chapter of a widely cited overview of how embodied cognition constitutes a "new science of meaning", Bergen (2012) mentions the different ways that readers are likely to conceptualize dogs and polar bears: "For you, the word *dog* might have a deep and rich meaning that involves the ways you physically interact with dogs—how they look and smell and feel. But the meaning of *polar bear* will be totally different, because you likely don't have those same experiences of direct interaction. If meaning is based on our experiences in our particular bodies in the particular situation we've dragged them through, then

For present purposes, we can grant the assumption that the representations that an agent possesses, in some sense, reflect the fit between its body type and its environment. The crucial question is whether the different-body/different-concepts hypothesis conflicts with and undermines concept nativism. To this question, our response is similar to one that we have pressed a number of times in Part III (e.g., in our discussion of artificial neural networks). It's that there is no inherent conflict between concept nativism and the theoretical insight that is central to this research. The key point to note is that, despite the scepticism about rationalist accounts of conceptual development that is common among proponents of embodied cognition, the view that different body types are associated with different concepts is entirely neutral as to whether these concepts are acquired via domain-general learning mechanisms or are instead acquired with the help of rationalist learning mechanisms or are even innate.

Suppose for the sake of argument that Lakoff and Johnson's uniform stationary spheres do differ cognitively from human beings in that the spheres invariably lack the concepts FRONT and BACK, whereas humans can be counted on to possess these concepts.[3] Now this species difference could be because individual spheres and individual humans, with their different types of bodies and different environments, end up having different types of experiences that are preserved by the concepts they acquire through general-purpose learning. As Lakoff and Johnson point out, humans generally move in the direction that their face is projected towards, while the hypothetical spheres don't have faces or any articulated external body parts. On the other hand, the conceptual difference could be because the bodily difference between humans and spheres is associated with differing evolutionary histories that have left humans and spheres each with innate systems of spatial representation that are well adapted to their differing ways of interacting with their environments. On this *rationalist* approach to spatial representation, the principle *different-body/different-concepts* would still apply. There would still be a correlation between body type and the presence or absence of the concepts FRONT and BACK. But this correlation would turn on the two species having differing acquisition bases—acquisition bases that are suited to their different body types.

Notice, too, in this particular case just how minimal the bodily constraints on representation happen to be. The sorts of body-based properties at issue have to

meaning could be quite personal. This in turn would make it variable across people and across cultures" (pp. 12–13).

[3] Despite the fact that the example was designed to illustrate a type of animal that would have absolutely no way of learning the concepts FRONT and BACK, it's actually not so clear that these spheres couldn't do so. As Shapiro (2019) points out, although the spherical beings don't themselves have bodies with a front and back, the concepts FRONT and BACK might nonetheless be important to them if objects or agents in their environment have significant functional differences in terms of their fronts and backs. "If the objects that approach them always do so with the same side 'forward'…then this may suffice to endow the spherical beings with the concept FRONT" (p. 120).

do with whether an organism has a front and back defined relative to how it perceives and moves. These and related body-based properties and relations (e.g., bodily symmetry/asymmetry, localized perceptual organs, and constraints on self-directed motion) are hardly *new* design features. They have been around for hundreds of millions of years. This means that at least some body-based constraints on the representation of space are among the conditions under which human—as well as mammalian, vertebrate, and much invertebrate—evolution has taken place, structuring the adaptive problems surrounding spatial representation far back in ancestral times. For this reason, one could argue in exactly the opposite direction of anti-rationalist advocates of embodied cognition. Instead of holding that children in each generation could in principle learn the fundamental way of representing space that suits the human body, one could hold that there is no need for children in each generation to have to do this. Evolution would have had ample time and resources to create innate systems of spatial representations that are well adapted to these core aspects of embodied living.[4]

To be clear, we aren't actually arguing for this more rationalist approach here. Our point is simply that, for each case of conceptual variability, it's a separate matter—to be investigated as a substantive issue—whether the variability should be explained in terms of experience-responsive domain-general learning, experience-responsive rationalist learning, or innate differences. The principle *different-body/different-concepts* is entirely neutral with respect to the rationalism-empiricism debate.

Proponents of embodied cognition who oppose concept nativism might object that there is no reason to postulate innate spatial concepts or domain-specific rationalist learning mechanisms for acquiring spatial concepts. Given that humans and spheres are bound to interact with their environments in different ways, these conceptual differences can emerge as a product of general-purpose learning. And if they can emerge in this way, then considerations of parsimony tell us to reject the rationalist explanation for being superfluous. Notice, though, that this is just another instance of an argument based on methodological empiricism, or the assumption that an empiricist learning model should be taken to be the default view when explaining how concepts are acquired. By now it should be clear that this assumption is entirely unfounded and that general-purpose learning should not be assumed when evaluating the prospects of competing empiricist and rationalist accounts (see Chapter 17). Where things stand, then, is that even on the

[4] Consistent with this idea, researchers have found that human newborns seem to have a basic understanding of the organization of their own bodies (Filippetti et al. 2013; Filippetti et al. 2015). In this work, infants who feel their face being touched look longer at a visual stimulus of a face being touched, but only when the seen and felt touching are spatially and temporally congruent and only when what is seen can be related to the infant's own body (e.g., the seen face is upright as opposed to inverted). Infants as young as 6 months old have also been found to spontaneously encode novel agents as having a front-back direction of orientation and to use this to predict their actions (Hernik et al. 2014).

assumption that the principle *different-body/different-concepts* holds, nothing directly follows regarding the status of concept nativism.

Let's turn now to a second strand in embodied cognition research that has been thought to undermine concept nativism. This is the idea that much of human action can be explained without postulating rich internal representations and the sorts of complex internal processes that are associated with rationalist models of various cognitive capacities. Clark (1999) expresses this view when suggesting that much action shouldn't be modelled in terms of a linear cycle "perceive, compute, and act". Instead, it can be coordinated as it unfolds through "real-time adjustments" that

> replace the notion of rich internal representations and computations, with the notion of less expensive strategies whose task is not first to represent the world and then reason on the basis of the representation, but instead to maintain a kind of adaptively potent equilibrium that couples the agent and the world together. (p. 346)

One often cited example of this sort of coupling concerns the type of navigation a baseball player uses to catch a fly ball. A cognitively rich account might help itself to calculations that predict the ball's final location based on the angle of its upward trajectory, its speed, and so forth. In contrast, an embodied cognition account might hold that there are simple perceptual cues that can be directly coupled with action, eliminating the need to postulate rich internal representations and complex internal processes, for example the heuristic of moving in whatever direction makes the ball appear to travel in a straight line in your visual field (McBeath et al. 1995).

Another widely cited flagship example involves the explanation of the so-called *A-not-B error* regarding object representation (which we discussed in Chapter 17). Recall that in the A-not-B error, an experimenter visibly hides an interesting object in one location (location A), lets the infant reach for it, and repeats this sequence a number of times. The experimenter then visibly places the object in a new location (location B). The surprising finding is that infants less than 1 year old continue to reach for location A even though they can see right in front of them that the object has been placed in location B. Embodied cognition theorists have claimed that this phenomenon is best explained in terms of "perceiving, moving, and remembering as they evolve over time, and that the [A-not-B] error can be understood simply and completely in terms of these coupled processes" (Thelen et al. 2001, p. 4). They argue that infants' performance on the A-not-B task varies in subtle ways depending on perceptual-motor factors that affect motor planning. For example, whether the infant reaches towards the A or the B location is affected by the number of times that the infant previously reached for the object in the A location and whether the infant's posture changes between the

A and the B trials (necessitating a different motor plan). This focus on perceptual-motor processes is taken to eliminate the need for central processes and even the possession of an object concept. "[T]here is no such thing as an 'object concept' in the sense of some causal structure that generates a thought or a behavior… There is only 'knowledge' of objects as embedded in the immediate circumstances and the history of perceiving and acting in similar circumstances" (Thelen et al. 2001, p. 34).

We will consider these two examples in turn. Regarding the first—the navigation involved in catching a fly ball—the main point to keep in mind is that, once again, there is no reason why a rationalist needs to reject the proposed explanation. Rationalists are not in any way opposed to action sometimes being controlled by simple, cognitively efficient processes. The heuristic-based strategy in this case is quite plausible, and similar heuristics may allow for computationally efficient processes in a wide range of cases. At the same time, however, there is no reason to suppose that this type of explanation can account for *all* human action.[5] Even in the domain of navigation, there is overwhelming evidence for a number of domain-specific rationalist learning mechanisms whose computations over internal representations mediate between perception and action.

As discussed in Chapter 1 and several subsequent chapters, one of these is a navigational mechanism that relies on geometrical representations of the environment to regain a correct heading after disorientation—to the exclusion of representations of other properties (e.g., landmark locations or featural properties like visual patterns or colours in the environment). A diverse range of findings argue for the existence of this innate domain-specific rationalist learning mechanism. Many of these findings are centred around the reorientation task in which (in the original version of the experiment) participants see an item hidden in one of four corners in a rectangular room and then are led to become disoriented. We noted in Chapter 1 that 18- to 24-month old children subsequently look for the item equally in both the correct corner and its geometrically equivalent opposite corner, and that they continue to behave in this way even when one of the short walls is covered with a blue cloth (and hence landmark information is available to help isolate the correct, unique location of the hidden item) (Hermer and Spelke 1996).

In addition, studies with a very broad range of animal species show that many animals behave in the same way, relying on geometry to locate the hidden object.[6] As we saw in Chapter 10, this is so even when they are reared in environments that prevent them from learning about the critical geometrical properties prior to

[5] For a forceful elaboration of this point, see Goldinger et al. (2016).

[6] Comparable results have been found in other primates (Gouteux et al. 2001), mammals (Cheng 1986), birds (Vallortigara et al. 1990), fish (Sovrano et al. 2003), and even insects (Sovrano et al. 2013). Some theorists have tried to explain these results without positing any geometrical representations per se. For a trenchant critique of these views, see Duval (2019).

being tested in the rectangular environment (Brown et al. 2007). We also saw in Chapter 1 that children and adults with Williams syndrome have specific difficulties with the reorientation task, despite being able to locate objects in a fixed location when they aren't disoriented (Lakusta et al. 2010). Finally, we saw that neurotypical children's behaviour showed quirky changes in response to variations of the reorientation task. In particular, children rely on geometry when the rectangular shape is instantiated by an arrangement of walls—even if the "walls" are a mere 2 centimetres high—but not when it is instantiated by a coloured patch on the floor or by four freestanding pillars.[7] Following Spelke and Lee (2012), we pointed out that this otherwise puzzling fact makes sense from a rationalist perspective when considered in an evolutionary context (Chapter 15). Arguably a well-designed system for navigation ought to incorporate an innate mechanism for representing the geometry of the contours of an environment, since, among other things, other location cues in natural environments are less reliable.

The geometry-based reorientation mechanism is only one of a number of innate specialized navigation mechanisms. As we discussed in Chapter 4, many species, including humans, also have a path integration mechanism for navigation, which computes the cumulative distance and direction that an organism has travelled from a point of origin (Gallistel 1990; Loomis et al. 1999; Wittlinger et al. 2006; Smith et al. 2013). Another navigational mechanism allows organisms to plot a path by reference to a cognitive map, which represents the overall structure of an environment with its major landmarks. This mechanism can be used when path integration isn't helpful, for example, when determining how to get from one's present location to a location that was visited on a previous occasion without first returning to the starting point of the current journey (Langston et al. 2010; Wills et al. 2010; Cheeseman et al. 2014). We have also seen that, in accordance with Evolved Navigation Theory, there is evidence that humans possess a psychological mechanism for representing (and misrepresenting) distances in navigation that takes into account the risk of falling (Jackson and Cormack 2007, 2008, 2010; Jackson and Wiley 2013; see Chapter 15).

None of these navigation mechanisms can be reduced to the type of direct coupling of perception and action that has been proposed for catching a fly ball. For example, path integration requires representing the critical variables of distance and direction for each change of direction in an extended journey and computing the cumulative distance and direction of the starting position relative to the current position. Likewise, a cognitive map represents the layout of an environment regardless of whether it is currently perceived, supporting computations for plotting novel paths. But even if some of the navigational feats that are

[7] This asymmetry holds even though the coloured patch and the four pillars are highly perceptually salient and form, or can be seen to form, the same geometrical shape as the diminutive walls (see Figure 1.2).

associated with these mechanisms could be achieved by directly coupling perception and action, a rationalist account of the origins of these mechanisms would still be called for given the wealth of evidence we have touched on.

It is perhaps also worth noting that these mechanisms provide further support for the idea that the principle *different-body/different-concepts* is perfectly compatible with concept nativism and that it often imposes only the weakest constraints on the types of representations an organism can possess. Suppose, for example, that the geometry-based mechanism for reorientation were, in some interesting sense, body based. Notice that this body-based mechanism would have to depend on extremely general features of our bodies. After all, it is a capacity that is also possessed by animals with very different body plans and that are evolutionarily distant from human beings. As we have seen, fish, for example, have been found to spontaneously use the geometry of an enclosure in attempting to locate a target corner, in much the same way that 18-month-old human infants rely on the geometry of the room to locate a hidden object. But of course fish don't have arms or legs, and have substantially more control over their up and down motion than humans do. So to the extent that bodily features are driving the attention to geometrical properties, the bodily features in question would have to be ones that are common to most or all vertebrate species (and to many invertebrate species too). And the collective weight of the evidence for this mechanism (pulling together the argument from early development, the argument from animals, the argument from neural wiring, and the argument from cognitive and behavioural quirks) would just go to show that this body-based capacity is rooted in an innate domain-specific mechanism.

Let's turn now to consider the A-not-B error. Here we can be quite brief. We can grant embodied cognition researchers that the A-not-B error is influenced by perceptual-motor factors, as rationalists like Susan Carey (2009) have been happy to do.[8] It is important to recognize, however, that even if this is right, such factors should not be thought to be the *only* factors that influence performance on these tasks. As we saw in Chapter 17, another important factor is whether the placement of the object occurs in a social context that elicits the expectation that this is an occasion in which generalizable information is being communicated. As Topál et al. (2008) showed, such expectations, linked to an innate system of natural pedagogy, play a large role in explaining infants' failure to reach for desirable objects in A-not-B tasks. More generally, the A-not-B error is only one small part of the development of children's understanding of objects. And, contrary to Thelen et al. (2001), there is an abundance of evidence for a rationalist account of the

[8] Although there is still a question about why the perceptual-motor factors have the impact they do. A key source of this impact may be that younger infants are more apt to choose location A over location B because their attention is directed to maintaining their balance and that this leaves them with fewer attentional resources to keep track of the fact that the object's location was switched to location B (Berger et al. 2019).

origins of object representation (see again the discussion of object representation in Chapters 10, 15, and 17). This just goes to show that, if the embodied cognition explanation of the A-not-B error is to be maintained, this will have to be in the context of a broader rationalist account.

We turn now to the third and final strand in the embodied cognition literature that is often taken to be in tension with concept nativism. According to this strand, conventional (non-embodied) accounts of concepts mistakenly assume that sensorimotor experience is recoded into a fully abstract format and that conceptual activity takes place in this amodal code in its own distinctive areas of the brain. In contrast, proponents of embodied cognition have argued that conceptual activity isn't separate from sensorimotor processes. It takes place in sensorimotor and affective systems by simulating aspects of what would be perceived and felt and how one might act in a given situation (Barsalou 1999). For example, within an embodied cognition framework, the concept KICK isn't an amodal representation that abstracts away from the sensory motor activity that takes place when seeing someone else kick something or when experiencing kicking something oneself. Rather, it is realized by the very same sensorimotor representations that are activated when perceiving or undertaking a kick.

Our discussion of this view will focus on a domain where the opposition to amodal representations enjoys a good deal of initial plausibility, namely, the domain of emotion concepts (concepts such as ANGER, JOY, PRIDE, and DISGUST). The embodiment claim in this case is that emotion attribution involves simulating being in the attributed emotional state (Niedenthal 2008). Emotion concepts, instead of being understood as abstract amodal representations that happen to refer to emotional states, are taken to be realized by sensorimotor activity and bodily changes that also occur when one is in the corresponding emotional state. Recognizing joy or anger in others, for example, would turn on the activation of many of the same representations and bodily reactions that underlie experiences of joy and anger in oneself.

Proponents of this view have cited a variety of findings in its favour (Niedenthal et al. 2014). The most exciting data in this area turn on interventions where facilitating or hampering bodily aspects of emotional activity seem to facilitate or hamper the use of a related emotion concept. For example, Havas et al. (2010) examined women who had Botox injections in the corrugator supercilii muscle to reduce frown lines. After receiving the injections, they were slower when reading sentences containing the words "sad" and "angry" but not when reading sentences containing the word "happy". According to the embodied cognition account of emotion concepts, this selective reduction in reading speed makes sense because the concepts SAD and ANGRY (but not HAPPY) incorporate motor processes that produce facial expressions with a furrowed brow. The Botox injections affected conceptual processing by impairing these women's ability to engage the needed motor activity. In related work, Niedenthal et al. (2009) asked participants to

determine if a given word was associated with an emotion, where some of the participants held a pen between their lips thereby interfering with both the muscles that pull the mouth into a smile and the muscles that curl the upper lip when exhibiting disgust. Holding a pen in this way was found to correlate with a reduced ability to categorize joy-related and disgust-related words as such, but had no effect on anger categorization.[9]

Let's assume for the sake of argument that emotion concepts are embodied in the way that embodied cognition researchers have claimed—that emotion concepts used in the attribution of emotions to others are realized by much of the same sensorimotor activity and affective changes that make up one's own experiences of the corresponding emotion. Still, just like the principle *different-body/different-concepts*, this fact in itself has no implications for the rationalism-empiricism debate. It is perfectly compatible even with a strong rationalist view that holds that certain emotion concepts are innate. A *rationalist embodied view* would just maintain that the sensorimotor and affective states that constitute these emotions are part of an innate categorization mechanism for recognizing these emotions and for engaging in other conceptual processes. For this reason, the question isn't whether to choose an embodied account of emotion concepts or a rationalist account. It is whether the origin of any given embodied emotion concept is best explained in empiricist or rationalist terms.

To illustrate this point and to show how an embodied rationalist account of the origins of certain emotion concepts is a serious possibility, consider the concept DISGUST. First of all, the emotion of disgust is arguably an innate feature of human psychology. There are numerous elicitors of disgust that are plausibly universal, including corpses, bodily products (e.g., faeces, vomit, and urine), breaches of the body (e.g., open wounds and sores), and outward signs of disease (Curtis et al. 2011; Kelly 2011). There is strong evidence of a universal facial expression for disgust (Elfenbein and Ambady 2002), and there are common features of the non-verbal vocalizations accompanying disgust that are even produced by congenitally deaf individuals, who have never heard how people react to disgusting situations (Sauter et al. 2019).[10] Pathogens have been one of the main selective pressures in human evolution (Fumagalli et al. 2011), and evolutionary theorizing about disgust suggests that disgust evolved in part as an adaptation for pathogen avoidance. It has been noted that the cost of infection places substantial selection pressure on all animals (not just humans), that this has led to adaptations for

[9] Work relating to these findings also illustrates the complexities of interpreting replication failures, with some findings in this area being supported by further research (e.g., Bulnes et al. 2019), and other highly influential findings failing to replicate (e.g., Wagenmakers et al. 2016; Morey et al. 2022).

[10] Though early work on the universality of emotions focused on recognition based on static facial cues, it is important to remember that emotions have complex multi-modal expressions (Keltner et al. 2019).

pathogen avoidance in other species, and that the pancultural core elicitors of disgust in humans pose a high risk of exposure to pathogens (Curtis et al. 2011; Kelly 2011).[11]

Why suppose that there is a rationalist account of the origins of the *concept* DISGUST as opposed to just the emotion of disgust itself? One reason is connected to the argument from prepared learning. Like pride, disgust is likely to have evolved as part of a communicative system.[12] And the communicative social functions of disgust would have required a reliable mechanism for interpreting disgust signals in other people. This argues for a concept of disgust that traces back to characteristically rationalist psychological structures in the acquisition base being part of an adaptive communicative system. Disgust isn't simply about responding to a fixed set of potentially harmful objects and substances when they are directly encountered. When manifested it helps onlookers to keep a safe distance from a potential hazard. Onlookers can also learn about new types of pathogenic hazards without having to engage in risky trial and error learning themselves. But of course both of these forms of socially mediated learning require being able to represent disgust in others.

Another reason in favour of adopting a rationalist account of DISGUST is connected to the argument from universality. A variety of evidence argues for the recognition of disgust signals being a human universal (Elfenbein and Ambady 2002; Sauter et al. 2010).[13] A particularly challenging test case for this claim of universality comes from societies where there is no word for disgust. In studying such a society, researchers compared the recognition of emotions in speakers of German (which has a word for disgust) to speakers of Yucatec Maya (which doesn't) (Sauter et al. 2011). When asked to label photographs of people exhibiting different emotional expressions (disgust, anger, and sadness), the Yucatec Maya speakers labelled the samples of disgust and anger in much the same way, unlike the Germans. However, when asked to match images in a non-linguistic task, the Yucatec Maya participants distinguished disgust from both anger and sadness just like the Germans. In fact, both groups showed categorical perception in that they did not treat blended images of emotional expressions as blends of two emotions but instead treated them as categorically falling under whichever

[11] Disgust's contribution to avoiding pathogens is most likely just one part of a suite of adaptations that have been described as the *behavioural immune system* (Schaller and Park 2011). Other components of the behavioural immune system include a particular sensitivity to social norms regarding behaviours that are associated with vectors for disease (e.g., sexual interaction and personal hygiene (Oaten et al. 2009)) and adaptations for avoiding and removing parasites that attach to the surface of the body (Kupfer and Fessler 2018). Innate preparedness for these connections may well involve articulation in the acquisition base among these elements, which would suggest a rationalist account of at least some further, related concepts.

[12] Our earlier discussion of the evidence for a rationalist account of the origins of PRIDE and other communicative emotions (see Chapter 14) further supports the case for a rationalist account of the origins of a range of embodied emotion concepts.

[13] The argument from universality here, as elsewhere, turns on how probable it is that the universal psychological trait isn't acquired solely via domain-general learning mechanisms (see Chapter 11).

emotion category contributed more to the blend (even if it was just over 50% of the blend).

In short, there is good reason for supposing that there is a rationalist account of the origins of the concept DISGUST, not just the emotion of disgust. On the assumption that DISGUST and other emotion concepts are embodied, then this would also be reason for supposing that there is a rationalist account of the origins of at least some embodied concepts. Being embodied is perfectly compatible with concept nativism.

We now want to consider a concept where it is not only true that a rationalist account and an embodied account are compatible, or where there is independent evidence for a rationalist account in a conceptual domain that is thought to be embodied, but where *the very same evidence* that supports embodiment also supports rationalism. The concept we will consider to illustrate this is the concept of physical formidability.

Physical formidability refers to the ability to resolve conflicts through physical force or the threat of such force. It's a graded and relative property. The same person may seem highly formidable to a smaller, weaker individual, but not particularly formidable to a larger, stronger individual. It turns out that people are quite good at assessing physical formidability. One study found that people are accurate when visually assessing men's fighting ability and that these assessments can be made just by looking at photographs of men's bodies (Sell et al. 2009). Other studies have found that people are even accurate when looking at just a photograph of a man's face (Sell et al. 2009) or listening to a recording of a man's voice (Sell et al. 2010).[14]

Although there has been little discussion of the concept PHYSICAL FORMIDABILITY in the rationalism-empiricism debate about the origins of concepts, physical formidability has undoubtedly been a major factor in determining who acquired and retained access to contested resources over the course of human evolution (Sell et al. 2009; Sell et al. 2012). Archaeological evidence indicates that human prehistory saw intraspecies violence on a scale far greater than is found in contemporary large-scale societies. In terms of sheer percentages of deaths, prehistoric violence exceeded even the most violent periods in recorded history (Keeley 1996). Physiological evidence also indicates that human hands and faces, especially in men, are adapted for combat (Morgan and Carrier 2013; Carrier and Morgan 2015). This and other work suggests that there was intense selection pressure in human evolution for the ability to make rapid and accurate assessments of physical formidability—to know when to use force and when to capitulate, and when to be cautious about entering into a dangerous conflict and when to recruit allies or adopt other means to augment one's defences.

[14] Women's physical strength can also be assessed from photographs of their bodies, though less accurately, and accuracy decreases even further when their formidability is assessed just from their faces or voices (Sell et al. 2009; Sell et al. 2010).

Of course, in dynamic real-world situations, physical formidability isn't just a matter of raw strength or fighting ability. It is also influenced by whether an individual has a weapon, the number of allies and kin an individual has, their age and health, the degree of group cohesiveness among allies and kin, and other factors. Keeping track of all of these variables and the potentially complex interaction effects among them could become computationally expensive—just like keeping track of the many variables that could affect the trajectory of a fly ball and using these to predict where it will land. However, assessments of physical formidability could be simplified with an *embodied summary representation* that is responsive to these differing variables. Daniel Fessler, Colin Holbrook, and colleagues have suggested that using a bodily representation of *size and strength* as a proxy for all these variables would be a particularly efficient way to summarize this information, where more formidable opponents are represented as being larger and stronger relative to one's own body regardless of the source of their formidability. This proposal is supported by some rather surprising findings.

In one study, participants were asked to estimate men's size and strength when shown just their hands holding either a weapon or a tool. If a man's hand was holding a gun or knife as opposed to a drill or handsaw, then the man was judged to be larger and more muscular (Fessler et al. 2012). In another study, participants were tested while they were alone or while there were potential male allies present. When judging the formidability of a "convicted terrorist" on the basis of his photograph, participants with potential allies present represented the terrorist as being smaller and less muscular (Fessler and Holbrook 2013a). In another study, participants were asked to estimate the size and strength of a "criminal" from his mug shot in one of two conditions—either making the estimate while walking in sync with another person or merely walking beside the other person. Synchronized movement is associated with enhanced cooperation and hence operates as a body-based signal of coalitional cohesion. In this case, the criminal was represented as smaller and less muscular when participants were engaged in synchronized walking than asynchronous walking (Fessler and Holbrook 2014). And in another study, participants were asked to make assessments of size and strength either in a condition that simulated quadriplegia or in one that simulated minor nerve damage—in the first case, by strapping the participant to a wooden chair; in the second, by attaching small metal caps to the participant's fingertips. In both cases, they were shown just the face of an angry-looking man and asked to indicate his size and muscularity. Participants in the first condition, who were effectively incapacitated, deemed the men to be larger and more muscular than participants in the second condition, who were still free to move their bodies (Fessler and Hobrook 2013b).

Taken together, this work makes a compelling case for an embodied cognition perspective on the concept PHYSICAL FORMIDABILITY. Assessments of physical formidability appear to be grounded in a representation of the relative size and

strength of an individual compared to a self-assessment of one's own size and strength, where both of these size and strength representations are sensitive to a wide range of factors that are relevant to fighting ability. Things like who is armed, supported by allies, or physically incapacitated are all folded into one concrete body-based summary representation.[15]

But notice that this very same work also makes a compelling case for a rationalist account of the concept PHYSICAL FORMIDABILITY. Since people with knives or guns are not actually bigger or stronger than people without them, a domain-general learning mechanism wouldn't be expected to arrive at this way of representing the situation. The same goes for the other effects we have mentioned. Walking in sync with others or having potential allies present doesn't actually make adversaries smaller or physically weaker, so a domain-general learning mechanism should not learn that they look smaller or weaker under such conditions. On the contrary, the more exposure it is given to potential adversaries and the more feedback it receives about their actual size and strength, the less likely it should systematically misrepresent their size and strength. And even if it did misrepresent size and strength, there would be no reason for it to do so in these particular ways, as opposed to systematically underestimating or overestimating people's size and strength as a function of one's general level of risk aversion.[16] Thus, the very same evidence that argues for an embodied representation of physical formidability also argues for a rationalist account of its origins.

One of the themes of this chapter has been that embodied cognition, in its different forms, isn't incompatible with concept nativism and that there is considerable room to explore rationalist proposals within an embodied cognition framework. Because the two are compatible, then given the overwhelming case for concept nativism (see especially Part II), future research on embodiment vis-à-vis the human conceptual system would be best served by exploring accounts that trace back to some of the many characteristically rationalist psychological structures in the acquisition base. That is, embodied cognition researchers should consider the possibility of distinctively rationalist forms of embodied cognition that take into account the sorts of considerations that factored into the seven

[15] For some additional factors affecting the summary representation, see Holbrook and Fessler (2013); Fessler et al. (2016); and Scrivner et al. (2020).

[16] This is not to say that there is no role for domain-general learning to play in all of this. You have to learn what a gun is and what it can do in order for it to have the impact that it does on the summary representation of size and strength, and this undoubtedly involves a certain amount of domain-general learning. Nonetheless, the input regarding modern weapons enters into the summary representation presumably because the formability assessment mechanism evolved to be specifically responsive to information about the possession of weapons. After all, ancestral humans would have had to learn about potential weapons too. Similarly, violent conflicts in ancestral times would have been greatly influenced by the presence of allies, their group's cohesion, physical liabilities, and so on. So it makes sense that a domain-specific rationalist learning mechanism would be responsive to all of these factors, supported by domain-general learning regarding the categories that meet its input criteria.

arguments for concept nativism from Part II. This isn't the place to revisit each of those arguments, but we want to illustrate how rationalist and embodied cognition research may fruitfully interact by briefly looking at an example linked to the argument from neural wiring. The example concerns constraints on the neural underpinnings of the conceptual representation of hands, feet, and tools.

Within an embodied cognition framework, there are no amodal representations, so concepts for hands and tools should be grounded in sensorimotor representations. Now neurological research has found that there are distinct specialized regions of the visual occipitotemporal cortex for representing different kinds of things and that the regions for representing hands and for representing tools normally overlap. Proponents of embodied cognition would claim that concepts for hands and tools are partly grounded in this circuitry—in the way that embodied concepts are always grounded in particular sensorimotor systems. We expect that many would also go on to explain this overlap as resulting from domain-general learning mechanisms operating on sensorimotor experience, particularly visual experience of seeing things like a carpenter pounding a nail with a hammer. However, this speculative developmental account can't be right, since congenitally blind individuals (who have never seen any hand-tool manipulations) have the very same hand-tool overlap in the visual occipitotemporal cortex (Peelen et al. 2013). What's more, it's unlikely that the overlap in such cases results from personal experiences of using tools with one's hands, since it also occurs in individuals who were born without any hands and consequently have no motor experience of their hands manipulating tools (Striem-Amit et al. 2017). Given all this, the overlap between hand and tool regions of the visual occipitotemporal cortex is likely to be a further case of constrained neural plasticity of the sort that we highlighted in the argument from neural wiring in Chapter 13. And this suggests that if an embodied cognition account of hand and tool concepts is adopted, this ought to be a rationalist embodied cognition account.[17]

[17] Related work has looked at representations for human actions in what is known as the *action observation network*, comparing the development of this network in individuals born without hands or upper limbs and individuals with intact arms and hands. Participants were shown actions that could naturally be performed in different ways, with either a hand or foot movement (such as fully opening vs. fully closing a partly open door either by pushing or pulling it with a hand or foot). No significant differences in the action observation network for representations of actions or the body parts involved in them were found between the groups. And both the large-scale organization of the action observation network and the abstract representation of actions independent of the body part involved in their performance were essentially the same across the groups. This work indicates that there are equally strong constraints on the development of this network and the associated representations for human actions, suggesting that if an embodied cognition account of the visual representation of observed actions is adopted, this should be a rationalist embodied account as well (Vannuscorps et al. 2019).

To sum up, research in the embodied cognition tradition has uncovered many important findings that should be factored into how we think about the workings of the mind. However, embodiment does not in itself vindicate an empiricist view of cognitive and conceptual development. On the contrary, we have argued that the role of the body in shaping cognition is perfectly compatible with concept nativism. Indeed, as with a number of other insights that have been associated with empiricist theorizing and that were discussed earlier in Part III, the core idea behind embodied cognition not only is compatible with rationalism but is substantially improved when developed in the context of an overall rationalist framework.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0022

23
Conclusion to Part III

Part III has examined a representative selection of some of the most important and influential empiricist accounts of concept acquisition, along with some key conceptual and methodological challenges grounded in empiricist theorizing. These alternative approaches and objections have been thought to show that concept acquisition can and should be explained wholly in terms of empiricist learning mechanisms without having to postulate many (or even any) characteristically rationalist psychological structures in the acquisition base. We have argued that none of these empiricist proposals or objections undermine concept nativism and that a critical evaluation of concept nativism's empiricist opposition only reinforces the explanatory appeal of concept nativism.

Some of the empiricist challenges to concept nativism that we have touched on are based on misunderstandings of what concept nativism claims. For example, in Chapter 20 we saw that neuroconstructivists accuse rationalists of holding that parts of the brain are "hardwired" in such a way that they are unaffected by the environment or by prior states of development. As we have repeatedly emphasized, however, this charge is based on a serious mischaracterization of concept nativism. Concept nativism wholeheartedly embraces both environmental influences on concept learning and developmental processes that are sensitive to prior neural and cognitive changes. And it does not entail that cognitive traits are determined, fixed, or unchangeable. Rather, it holds that these neural, cultural, and environmental influences should be understood as being mediated not solely by empiricist learning mechanisms but by a combination of different types of learning mechanisms which notably includes rationalist learning mechanisms that trace back to characteristically rationalist psychological structures in the acquisition base. This means that the acquisition base contains many characteristically rationalist psychological structures, but it doesn't in any way preclude development, change, or variation. It only entails that these structures aren't themselves acquired via more fundamental psychological structures.

Other major empiricist objections to concept nativism that we have discussed aim to show that rationalist interpretations of some influential test cases in recent rationalist research are unsubstantiated. For example, proponents of methodological empiricism have argued that there are simple empiricist explanations of data associated with rationalist views of object representation, and neo-associationists have argued that there are low-level perceptual explanations of data that rationalists have taken to demonstrate early forms of sociomoral

representation in infants. Of course, in all such cases the devil is in the details. But we have argued that prominent empiricist criticisms of this kind fail either because they wrongly focus on the mere possibility of an alternative explanation or because they focus too narrowly on a small set of experimental results, disregarding crucial further considerations that decisively undermine these alternative empiricist explanations. When these considerations are given their due, important test cases—like the representation of physical objects or the representation of agents' helping or hindering others—not only fail to undermine rationalism but actually provide further support for it. In this way, our discussion of empiricist alternatives in Part III has also served to substantially supplement the positive case for concept nativism that we gave in Part II.

While several of the influential empiricist objections that we have discussed turn on misunderstandings or problems of these sorts, empiricist theorizing about the origins of concepts should not simply be dismissed. One of the main morals in Part III is that many of the most important resources and insights to be found in empiricist theorizing can and should be integrated with rationalist elements. In Part III we have argued with respect to a number of case studies that these kinds of resources and insights are not only perfectly compatible with concept nativism but also capable of making a far greater contribution to our understanding of conceptual development when incorporated into a rationalist framework. For example, we saw in Chapter 22 that an embodied cognition approach to conceptual development is consistent with concept nativism and that there is good reason to suppose that flagship cases involving the A-not-B error in object cognition, bodily concepts like FRONT and BACK, navigation-related concepts, and emotion concepts, among others, are all better understood in these terms. Likewise, in Chapter 19 we saw that artificial neural network models of semantic memory needn't assume that semantic memory starts out as a largely undifferentiated system of representation. Instead, such models can build in innate high-level category-specific dimensions and settings—for example, ones that innately prepare the network to single out animals and plants and to have distinctive expectations involving these categories. Such a rationalist artificial neural network model would be far better equipped to explain the actual course of conceptual development in these domains. Similar points could be made about the role of positive and negative reinforcement learning in the domain of moral cognition (Chapter 18) or the role of spatial thinking in conceptual development (Chapter 21).

All this is to say that concept nativism doesn't entail a categorical rejection of all the insights and resources of empiricist theorizing. Empiricists have been wrong in maintaining that conceptual development is mediated *entirely* (or nearly entirely) by domain-general learning mechanisms. But empiricists have been right to emphasize that domain-general learning makes an important contribution to understanding development. Since domain-general learning is fully compatible with rationalist approaches, concept nativists can and should maintain

that cognitive and conceptual development traces back to a rationalist acquisition base containing not just characteristically rationalist psychological structures but also domain-general learning mechanisms, and that development is mediated by a rich set of rationalist learning mechanisms working alongside and together with domain-general learning mechanisms.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0023

PART IV
FODORIAN CONCEPT NATIVISM

24
The Evolution of Fodor's Case against Concept Learning

Rationalism encompasses a broad spectrum of views. But when it comes to rationalism about concepts, people often lose sight of this fact and unwittingly identify rationalism with *radical concept nativism*, Jerry Fodor's notorious view that virtually all lexical concepts are innate. Since Fodor's radical concept nativism is closely connected to his allegation that theories of concept learning are deeply problematic, it is also common to assume that what divides rationalist from empiricist views of the conceptual system is the role they assign to learning. Empiricist models are supposed to favour learning, while rationalist models are supposed to be categorically opposed to learning.

It should be absolutely clear by now that we reject this construal of the rationalism-empiricism debate. As we argued at length in Part I, and repeatedly illustrated throughout Parts II and III, rationalists and empiricists don't disagree about whether learning should be taken to figure prominently in conceptual development but about whether, and to what extent, such learning depends upon characteristically rationalist psychological structures in the acquisition base.[1] Although it is a common mistake to take the rationalist viewpoint in general regarding the origins of concepts to just be Fodor's view, his view really is an extreme outlier. For this reason, we refer to Fodor's view as *radical concept nativism* to highlight how extreme it is—a point that is not disputed, even by Fodor—and to contrast it with our own view and with the broader class of rationalist views that we refer to as *concept nativism*.[2] This broader class covers a range of views. Some of these take a rationalist approach to only a handful of conceptual domains, while others, like our own approach, take a rationalist approach across many conceptual domains.

[1] For readers not reading the chapters in order, there are a number of technical terms that were introduced and explained earlier that we will continue to rely on in Part IV, including "acquisition base", "rationalist learning mechanism", "characteristically rationalist psychological structures", "articulation", and "alignment". Brief summaries of how we are using these and other terms may be found in Boxes 1–7 in Chapter 2 and in Box 8 in Chapter 6.

[2] Radical concept nativism is still a type of rationalist account. But it is useful to partition rationalist views in a way that excludes it from the mainstream rationalist views that we collectively refer to using the term "concept nativism". In addition to highlighting the radicalness of Fodor's view, this also provides a relatively easy way to refer to the collection of rationalist views that are not extreme in the way that Fodor's is without having to repeatedly use a cumbersome expression like "non-extreme versions of concept nativism" to collectively refer to these rationalist views, which are the rationalist views that are primarily at issue in the rationalism-empiricism debate.

Different rationalist views will also strike different balances between postulating concepts as innate or as acquired via rationalist learning mechanisms. And, of course, different rationalist views will posit different types of rationalist learning mechanisms—differing along the various dimensions that we highlighted in Chapter 2, such as complexity, articulation, and alignment. But one thing they all have in common is that they take learning to be critical to the development of the human conceptual system.[3]

While Fodor's views about learning are undoubtedly extreme—and rightly rejected by virtually all philosophers and cognitive scientists—it is also true that many leading theorists in cognitive science have found it to be enormously valuable to engage with Fodor's views and arguments. We wholeheartedly agree with them about this. Fodor's discussions of these issues are among the richest and most interesting in contemporary philosophy and cognitive science—a fact that can be easy to overlook given the extreme implausibility of Fodor's conclusions. And some of the most important and creative approaches to human conceptual development—such as Susan Carey's proposal about bootstrapping (which we touched on briefly in Chapter 2)—have without a doubt been directly motivated by the challenge that Fodor presents.

Much of our discussion in Part IV will be focused on an immensely important and influential view about learning that many take to be at the heart of Fodor's arguments—a view that imposes a fundamental constraint on any theory of concept learning. While rejecting Fodor's radical concept nativism, these theorists agree with Fodor's claim that primitive concepts and representations (ones that are not composed of simpler representations) cannot be learned and so must be innate. In fact, this view about conceptual structure and the limits on what can be learned lies behind a nearly universally accepted model of concept acquisition—endorsed in different ways by rationalists and empiricists alike—the *Acquisition by Composition model* (*ABC model*) of concept acquisition, which was introduced in Chapter 5. According to this model, concept learning requires that the learned concept be a complex concept which is formed from a compositional process that builds the new concept out of its semantic constituents.

One of the key morals of Part IV of the book is that this model is mistaken. By carefully analysing Fodor's arguments, we show precisely how they go wrong, which in turn shows why the ABC model of concept acquisition should also be rejected. The rejection of this model opens up a range of new possibilities for

[3] We use the terms *our version of concept nativism* and *our concept nativism* in the way they were introduced in Part I to refer to our own view, a form of concept nativism according to which many concepts across many conceptual domains are either innate or acquired via rationalist learning mechanisms. As we emphasized earlier, the arguments we have given in support of our concept nativism are not intended to argue for a single fully specified theory of conceptual development, but rather are compatible with a variety of rationalist theories with detailed commitments regarding the characteristically rationalist psychological structures in the acquisition base which current research is, unsurprisingly, unable to decide among.

explaining how concepts can be learned which we explore in relation to a variety of different types of concepts and different theories of the nature of mental content (in Chapter 25).

In his final treatment of these issues, Fodor rejected radical concept nativism, while standing by the core of his earlier arguments that concept learning is not possible (Fodor 2008). Exploring Fodor's alternative to radical concept nativism also turns out to be fruitful, particularly in highlighting the importance of culture to the origins of concepts—which we argue Fodor's alternative is unable to account for. Our discussion of these issues (in Chapter 26) also explores the depth of the connection between our own rationalist account of the origins of concepts and cultural learning.

In short, there is a great deal to be gained by carefully examining Fodor's arguments, despite the highly counterintuitive nature of their conclusions. The first step is to get a sense of the evolution of Fodor's thinking. Fodor has been a persistent and outspoken critic of theories of concept learning ever since he first broached the issue in the 1970s. Though Fodor's case against learning, and the extreme rationalist views that have been closely connected with it, are among Fodor's most distinctive philosophical commitments, many of his readers have been baffled by this ongoing theme in his work. How could anyone deny that concepts like XYLOPHONE and KANGAROO are learned? In this chapter, we explain Fodor's case against concept learning and how it has evolved over the years, culminating in his most recent and trenchant treatment of these issues in *LOT 2: The Language of Thought Revisited*.

24.1 *The Language of Thought* (1975)

Fodor first presented his case against concept learning in his landmark book *The Language of Thought*. He argued that any account of the acquisition of concepts—if it has any hope of making good on its claim that concepts are learned—must take the process to involve framing and testing hypotheses about the concept to be learned. From here he went on to claim that this apparently simple observation leads to a big problem.

The problem is that their reliance on hypothesis testing makes theories of concept learning *viciously circular* because representing the hypotheses and the evidence needed in order to learn a concept requires being able to use that very concept:

> What has been argued is, in effect, this: If the mechanism of concept learning is the projection and confirmation of hypotheses (and what else *could* it be), then there is a sense in which there can be no such thing as learning a new concept. For, if the hypothesis-testing account is true, then the hypothesis whose

> acceptance is necessary and sufficient for learning *C* is that *C* is that concept which satisfies the individuating conditions on Ø for some other concept Ø. But, trivially, a concept that satisfies the conditions which individuate Ø *is* the concept Ø. It follows that no process which consists of confirming such a hypothesis could be the learning of a *new* concept (viz., a concept distinct from Ø). (Fodor 1975, pp. 95–96)

In other words, if C is the target concept, then to learn it, the learner must first be in a position to frame a hypothesis to the effect that C is this concept. And if the learner can do that, then she must already have C—in which case there is nothing to learn. (We should note that Fodor had a broad and inclusive notion of a concept in which most mental representations would count as concepts, and his circularity argument doesn't turn in any way on how the conceptual/nonconceptual distinction is drawn. He could equally have said that if C is the target representation to be learned, then the learner must already have this representation C—and hence wouldn't really be learning it.)

But why think that concept learning takes this form? One reason is very general. As Fodor noted, it's not enough that a concept be acquired via causal interactions with the environment for it to be learned. Maturation involves a certain amount of interaction with the environment. For that matter, so do causal flukes (e.g., if one were to miraculously acquire a concept upon being hit on the head, or through some type of futuristic neurosurgery). The point is that, to be worthy of the name, concept learning must involve something akin to a rational process in which the experience that occasions a concept justifies its adoption, as hypothesis testing does.[4]

A second reason for taking concept learning to have this form is that, as Fodor argued, experimental work on concept learning in psychology has employed a model in which concept learning amounts to inductive extrapolation, which essentially involves taking concepts to be learned via hypothesis testing. In a typical experiment, subjects are given stimuli and the task of sorting them in accordance with an unknown rule while receiving feedback about whether their responses are correct or not. Suppose the criterion of success is to pick out the red items from among many different shape and colour combinations. To succeed in the

[4] When we introduced the term *learning mechanism* in Chapter 2, we deliberately characterized learning mechanisms in a way that encompassed all cases in which a psychological trait is acquired via psychological processes, including processes that aren't necessarily rational. We did this because, in the end, the rationalism-empiricism debate is about the character of the psychological structures in the acquisition base, and so we needed a single term to cover any case in which a psychological trait is acquired via psychological processes as opposed to being in the acquisition base. This loose use of the terms *learning* and *learning mechanism* were appropriate for our purposes in Parts I–III of the book. However, in Part IV we will employ a more restrictive notion of learning, which differentiates cases of learning from other cases of psychological-level acquisition and which, as in Fodor's usage, is intended to be broadly in keeping with the ordinary usage of the term "learning".

task requires keeping track of the relevant feedback (e.g., that the response *yes* to a red circle is rewarded) and using this to ultimately settle upon the correct criterion. In other words, in each trial as the experiment proceeds, the learner must entertain a hypothesis about the sorting rule (*Is it the red things? Maybe the red circles?*) and represent the evidence that pertains to its evaluation. But then the trouble kicks in. To be able to represent the hypothesis that the correct sorting rule picks out the red items requires that one already have the concept RED. So we are left with the troubling view that you need the concept RED in order to learn it.

This circularity argument makes a powerful prima facie case for the radical conclusion that no concepts can be learned. But, if concepts can't be learned, presumably they must be innate, which as Fodor noted, constitutes "a very extreme nativism" (Fodor 1975, p. 96). The radicalness of the conclusion led Fodor to remark in a discussion from the same period that "what is puzzling about the [circularity] argument… is exactly that it requires only these fairly banal assumptions to arrive at the wildly paradoxical conclusion that all concepts are innate" (Fodor 1980, p. 328). And in *The Language of Thought*, he expressed sympathy with the protest that something must be wrong if concepts like OBOE and SILICON turn out to be innate. Is there any way to avoid this conclusion? One possibility he noted was that lexical concepts—concepts corresponding to words in natural language—might decompose into more basic concepts. Then perhaps learning could be "reconstructed as a process in which novel complex concepts are composed out of their previously given elements" (Fodor 1975, p. 96). This possible loophole became the focus of his next major discussion of these issues.

24.2 "The Present Status of the Innateness Controversy" (1981)

Fodor's (1981) "The Present Status of the Innateness Controversy" is perhaps the richest and most detailed philosophical discussion of the rationalism-empiricism debate concerning the origins of concepts in the twentieth century. It undertook an extensive exploration of Fodor's earlier circularity argument, the potential loophole to this argument cursorily noted in *The Language of Thought*, and the general theoretical landscape concerning concept acquisition shaped by these considerations. In the course of this discussion, Fodor formulated a way of thinking about the rationalism-empiricism debate about the origins of concepts that was embraced in much of cognitive science. In effect, Fodor argued that rationalists and empiricists are equally committed to some form of the ABC model of conceptual development and that the focal point of their disagreement comes down to competing views about the set of semantically primitive representations that constitute the basis from which all complex concepts are formed. On this

picture of the debate (which we should note is very different from the one we presented in Part I), while empiricists suppose that the stock of primitive representations is relatively modest, rationalists suppose that it is quite rich.[5]

What's more, Fodor also argued in this paper that the ABC model, though theoretically cogent and critical for understanding concept learning on all accounts (rationalist and empiricist), isn't a real possibility for most lexical concepts. He did this by arguing that most lexical concepts are semantically primitive and hence lack the needed semantic structure to be composed from simpler representational components. The result was that by closing the loophole from *The Language of Thought*, "The Present Status of the Innateness Controversy" presented Fodor's fullest and most powerful case for his radical concept nativist thesis that virtually all lexical concepts are innate.

The central argument in Fodor's paper is somewhat complicated, so let's take a careful look at each of its two stages. The first is the claim that only complex concepts can resuscitate the idea that concepts are learned via hypothesis testing. To see why, suppose that a learner finds herself in the sort of concept learning experiment mentioned in the previous section and that the correct rule has to do with whether a stimulus is RED (i.e., the concept to be learned is RED). In that case, to succeed in the task, the learner has to make and keep track of observations that would help her confirm this hypothesis. What sorts of observations? Observations of red stimuli, presumably. Notice, however, that if red stimuli need to be categorized as such in order to even collect evidence for the hypothesis at issue, then the learner would have to already have and be using the concept RED. By contrast, if the correct rule involved a more complex concept—for example, if it had to do with whether a stimulus is a red circle—then in principle the target concept (RED CIRCLE) might be confirmed by observations that are framed solely in terms of its constituents (RED and CIRCLE); RED CIRCLE itself wouldn't have to be used.

The second step in Fodor's argument is the more controversial one. Many theorists are comfortable with the suggestion that only complex concepts can be learned because of the widespread assumption that ordinary lexical concepts generally have the needed semantic structure. In contrast, Fodor claimed that most lexical concepts are primitive and so can't be learned. Moreover, as in *The Language of Thought*, he thought that the natural inference was to conclude that these concepts

[5] Fodor is explicit that both rationalists and empiricists are committed to the ABC model: "Both sides [in the rationalism-empiricism debate] assume that the space of concepts potentially available to any given organism is completely determined by the innate endowment of that organism. This follows from the assumptions that (a) the set of potentially available concepts is the closure of the primitive concepts under the combinatorial mechanisms; (b) the set of potentially available primitive concepts is innately fixed; and (c) the combinatorial mechanisms available are themselves innately specified" (1981, p. 277). As we noted earlier, Fodor is using a very broad and inclusive notion of a concept in which most mental representations would count as concepts. So the assumptions highlighted here should be understood as stating that the set of all possible concepts for an organism is fixed by the set of its primitive representations, and that these and the mechanisms for combining them are innate.

must be innate. The result was Fodor's notorious radical concept nativism, the view that there are at least as many innate concepts as words in language.

Fodor's argument against lexical concepts having hidden semantic structure turned on a mix of empirical and theoretical considerations. One of these cited the persistent difficulties that philosophers have had in providing satisfactory definitions for the terms they have attempted to analyse. As Fodor noted, philosophers have sought definitions of terms like "knowledge", "truth", and "goodness" since antiquity and have reached no consensus on how they should be defined. Purported definitions are nearly always subject to counterexamples. Fodor argued that this fact begins to make sense on the assumption that the underlying concepts lack definitional structure. In a related argument, Fodor noted that psychological experiments corroborate the lack of definitional structure for lexical concepts. The concepts that are the best candidates for definitional analysis (e.g., causatives such as BREAK or KILL) show no effects of their purported complexity relative to concepts that are presumably less complex (Fodor et al. 1980).[6] This too makes sense once it is accepted that the former don't have definitional structure after all. Finally, Fodor criticized what he took to be the main alternative proposal about the kind of structure that concepts might have apart from definitional structure, namely prototype structure. A concept has prototype structure when it decomposes into simpler concepts that collectively specify an abstractly represented central tendency or best example. For example, the prototype for BIRD incorporates representational elements for such things as feathers, beaks, and flying even though it's understood that not all birds fly. Fodor cited a number of problems with prototype structure, but the main one was the objection that prototypes aren't compositional and that this is evident given the fact that most complex concepts don't have prototypes. "[T]here may be prototypical *grandmothers* (Mary Worth) and there may be *prototypical properties of grandmothers* (*good*, *old* Mary Worth). But there are surely no prototypical properties of *grandmothers most of whose grandchildren are married to dentists*" (Fodor 1981, p. 297).

What makes the position in "The Present Status of the Innateness Controversy" so radical is this second part of the argument, which addresses the potential loophole in the circularity argument that Fodor had identified in *The Language of Thought*. Since this loophole applies *only* in the case of complex concepts, Fodor's argument that virtually all lexical concepts are primitive has an enormous significance. By Fodor's circularity argument, the argument that lexical concepts are primitive means they can't be learned, and this in turn leads to the outrageous conclusion that virtually all lexical concepts are innate—including a multitude of concepts that are associated with specific cultural, technological, and scientific developments (DOLLAR, HUBCAP, JAZZ, BACTERIA, ELECTRON, etc.).

[6] The standard analysis of a causative defines the causative word or concept in terms of the cause of a given type of event. For example, KILL is analysed as CAUSE TO DIE.

Without Fodor's argument that virtually all lexical concepts are primitive, the circularity argument only gets us as far as the ABC model. This is a model that is widely accepted by both rationalist and empiricists.[7] But with Fodor's argument that lexical concepts are primitive, the circularity argument goes much further, leading to an extraordinarily radical view about the scope of innate concepts. This explains both why "The Present Status of the Innateness Controversy" has been taken to contain the definitive statement of Fodor's radical concept nativism, and why this paper was so important in framing much of the contemporary debate about the origins of concepts. It made clear that the circularity argument on its own might have quite modest consequences regarding the stock of innate concepts and representations, while highlighting the importance of getting clear about which representations are primitive and especially whether lexical concepts are primitive.

As Fodor (1981) left things, the question of how many concepts and representations are innate amounts to a question about the size and character of the set of the primitive (i.e., unstructured) representations that constitute the basic elements that can combine to form complex concepts. This way of looking at the matter is now widely accepted in much of cognitive science and has set the agenda for many theorists who hope to preserve the common-sense idea that ordinary lexical concepts are learned.[8] The focus for such theorists has been to identify the internal semantic structure of different types of concepts. Thus many of Fodor's critics have come to think that he had the overarching dialectic right. If few scientists have gone on to embrace Fodor's radical concept nativism, this is because they have held that Fodor was wrong about the structure of lexical concepts, maintaining that they *do* have the necessary structure to be learned.

24.3 *Concepts: Where Cognitive Science Went Wrong* (1998)

On the surface, Fodor's 1998 book *Concepts* appears to present a major shift in his thinking on these issues. This can make it easy to overlook the fact that the core elements of Fodor's position in his 1981 paper are all still in place. In particular, Fodor continued to endorse the circularly argument that it's not possible to learn semantically primitive concepts, and he continued to endorse his arguments that virtually all lexical concepts cannot be learned because they are primitive. So, *Concepts* is just as strongly anti-concept-learning as the 1981 paper.

[7] Which is not to say that this thesis is correct; below we argue that it isn't.

[8] For example, in the introduction to an influential volume on lexical semantics, Levin and Pinker write: "Psychology…cannot afford to do without a theory of lexical semantics…Whether or not one agrees with Fodor's assessment of the evidence, the importance of understanding the extent to which word meanings decompose cannot be denied, for such investigation provides crucial evidence about the innate stuff out of which concepts are made" (1991, p. 4).

But in *Concepts*, Fodor began to explore the possibility that there is a way to take primitive lexical concepts to be *neither learned nor innate*. Is there an overlooked alternative to both learning and innateness? Fodor's proposal in *Concepts* was that the way to escape the confines of the learned versus innate dichotomy is to provide a general alternative for the acquisition of lexical concepts that shifts the story about their acquisition from the psychological level—the level of representational states and processes—to the neurological level:

> [T]hough there has to be a story to tell about the structural requirements for acquiring DOORKNOB, intentional vocabulary isn't required to tell it. In which case, it isn't part of cognitive psychology. Not even of "cognitive neuropsychology"…(as opposed, as it were, to neuropsychology *tout court*). (Fodor 1998, p. 143)[9]

In other words, the reason why lexical concepts aren't learned isn't that they are innate but rather that their acquisition isn't owing to anything akin to an inferential process. On this view, the general story regarding the origins of primitive concepts doesn't involve psychological-level processing; their acquisition is explained directly and entirely in neurological terms. This is not simply the view that ultimately it is the activity of the brain that explains the acquisition of such concepts (just as it is the activity of the brain that explains cognition generally). Rather, it is the view that these concepts are not acquired via any type of psychological process, so the only kind of process involved is a biological one.

On the face of it, this may seem like a major change in Fodor's thinking. But, it is important to remember that *Concepts* still holds lexical concepts aren't learned. Moreover, from the point of view of the account of innateness we argued for in Part I—in which a psychological trait is innate if isn't acquired via a psychological-level process—Fodor's view in *Concepts* that concept acquisition can only be explained in biological terms just amounts to another way of saying that these concepts are innate. If concepts are not acquired via any psychological process, they thereby are part of the acquisition base (and so on our understanding of innateness, they are innate). So even though *Concepts* purports to offer an alternative to taking primitive concepts to be innate, from our point of view the account that it offers isn't a meaningful alternative to taking them to be innate. Moreover, because *Concepts* takes this biological account to be a *general* account of the origins of lexical concepts, Fodor's (1998) view is tantamount to holding that virtually all lexical concepts are innate. In other words, despite initial

[9] Terminological note: The term "intentional" in this context is used as a technical term along the lines that we discussed in Part I. It doesn't refer to ordinary intentions (as when someone intends to go on a diet) but to items (e.g., mental states) that have the property of representing or being about other things (see Chapter 6). In proposing that intentional vocabulary isn't needed, Fodor is suggesting that the states and processes involved in the acquisition of DOORKNOB aren't representational.

appearances, the view in *Concepts* really is no different than the radical concept nativism articulated in Fodor (1981).

24.4 *LOT 2: The Language of Thought Revisited* (2008)

In his most recent treatment of these issues, in *LOT2*, Fodor concludes, of all things, that his previous discussions showed "a failure of nerve" for not going far enough (2008, p. 138). While previously he held that lexical concepts can't be learned because they are primitive, he at least allowed for the possibility that manifestly complex concepts can be learned (e.g., ones that are expressed by phrases in language, as opposed to individual words—concepts like LARGE BROWN COW). Though far stronger than the argument in *The Language of Thought*, Fodor's argument in *LOT2* shares its simplicity. The main difference is that *LOT2* claims that there are no exceptions to the earlier circularity argument, not even for manifestly complex concepts. Consequently, whether lexical concepts are primitive or not is irrelevant to the question of whether they can be learned. "I must confess that I have come to agree with my critics that there is something wrong with the argument as [*The Language of Thought*] presented it; namely, that the conclusion is too weak and the offending empirical assumption—that quotidian concepts are mostly primitive—is superfluous" (2008, p. 130).

Fodor's new argument is thus a far more powerful version of the earlier argument. It draws much the same anti-learning conclusion but without the need to take a stand on whether any given concept has semantic structure or not. Just as you need RED to entertain and test the hypothesis that RED applies to all and only red things, so you need the concept RED OR SQUARE to entertain and test the hypothesis that RED OR SQUARE applies to all and only red or square things. The result is that lexical concepts can't be learned, not because they are primitive, but because the circularity argument shows that *no* concept can be learned via hypothesis testing. The whole idea of concept learning is simply confused.

Here is our reconstruction of the *LOT2* argument:

1. Concepts (whether primitive or complex) cannot be learned via hypothesis testing.
2. There is no other way that a concept could be learned.
3. Therefore, concepts can't be learned.

In support of the first premise, Fodor argues in a way that is highly reminiscent of *The Language of Thought*, claiming that hypothesis testing (HF) models of concept learning are inherently circular (2008, p. 139)[10]:

[10] "HF" is Fodor's shorthand for the view that "concept learning is a process of inductive inference; in particular, that it's a process of projecting and confirming hypotheses about what the things that the concept applies to have in common" (Fodor 2008, p. 132; italics removed).

> Now, according to HF, the process by which one learns C must include the inductive evaluation of some such hypothesis as 'The C things are the ones that are green or triangular'. But the inductive evaluation of that hypothesis itself requires (*inter alia*) bringing the property *green or triangular* before the mind as such…Quite generally, you can't represent anything as *such and such* unless you already have the concept *such and such*. All that being so, it follows, on pain of circularity, that 'concept learning' as HF understands it *can't* be a way of acquiring concept C…Conclusion: *If concept learning is as HF understands it, there can be no such thing*. This conclusion is entirely general; it doesn't matter whether the target concept is primitive (like GREEN) or complex (like GREEN OR TRIANGULAR).

In other words, to test and confirm the hypothesis that the concept to be learned is the concept C, you must use the concept C. But if you are already using the concept prior to learning it, then you aren't really learning it; you already have it. So no concept, not even a complex concept, can be learned in this way.

Of course, the problem with hypothesis testing models wouldn't be so bad if other approaches to concept learning were viable. The burden of Fodor's second premise is to exclude all other approaches in one fell swoop. Fodor argues for the need for hypothesis testing, as he did in *The Language of Thought*, by noting the intuitive contrast between learning and instances where a concept is acquired through wholly non-rational processes (Fodor 2008, p. 135):

> [T]he experience from which a concept is learned must provide (inductive) evidence about what the concept applies to. Perhaps COW is learned from experiences with cows? If so, then experiences with cows must somehow witness that it's cows that COW applies to. This internal connection between concept learning and epistemic notions like evidence is the source of the strong intuition that concept learning is some sort of rational process. It contrasts sharply with kinds of concept acquisition where, for example, a concept is acquired by surgical implantation; or by swallowing a pill; or by hitting one's head against a hard surface, etc.

Fodor argues that hypothesis testing is the only proposal in which concept acquisition is rationally constrained. "[I]f we are given the assumption that concept learning is some sort of cognitive process, HF is de facto the only candidate account of what process it might be" (Fodor 2008, p. 139). Granted, there may be non-rational processes that eventuate in a concept's being acquired, but *learning* must involve the formulation and testing of hypotheses.

In sum, though we need hypothesis testing for a concept's acquisition to count as learning, hypothesis testing is itself a non-starter as an explanation for concept learning since it presupposes that the concept to be learned is already possessed.

The upshot, as Fodor puts it, is that "there can't be any such thing as learning a concept" (Fodor 2008, p. 139).

What about the question of whether lexical concepts (or any concepts for that matter) are innate? Here *LOT2* largely retains the view adopted in *Concepts*. Fodor claims that it needn't follow from the fact that a concept isn't learned that it is innate. Concept acquisition might instead be the product of predominantly non-representational neurological processes. While *LOT2*'s account of concept acquisition does involve a representational component, it turns out to play only a minor role in a largely a non-representational theory much like the one in Fodor (1998), as we'll see in Chapter 26.[11]

24.5 Conclusion

While Fodor's views on concept learning and innateness have evolved over the years, many of his central commitments have remained constant. From *The*

[11] Georges Rey offers a very different account of Fodor's argument in *LOT2*. As Rey sees it, in *LOT2* Fodor takes himself to show that all concepts are innate. Rey interprets Fodor as holding that there are only two options, concepts must be either learned or innate. And since Fodor argues that concept learning is impossible, Rey sees Fodor as concluding that all concepts must be innate (Rey 2014, p. 109). However, Rey rephrases this point using a new technical term, "innately possessed", saying that "one might usefully think of the set of concepts that are innately possessed as being just that set that can come to be manifested by learning and bootstrapping" (Rey 2014, p. 125). What "all concepts are innately possessed" amounts to is the claim that all concepts are in principle capable of being acquired through some form of learning given the conceptual system's semantic structure. So the view that Rey attributes to Fodor is that (1) all concepts are innate, and (2) all concepts being innate (i.e., "innately possessed") effectively means that all concepts are in principle capable of being acquired though some form of learning. Rey's interpretation of Fodor faces two main problems. The first problem is that it does not fit with the actual text of *LOT2*. Rey does not provide any textual evidence that Fodor holds these views, and there is clear textual evidence against Rey's interpretation. For example, in *LOT2* Fodor writes that "Minds like ours start out with an innate inventory of concepts, of which there are more than none but not more than finitely many" (p. 131). By contrast, the set of concepts that, in Rey's terms, are innately possessed is infinite, since the conceptual system is recursive (e.g., this set includes AMY'S OLDEST CHILD; AMY'S OLDEST CHILD'S OLDEST CHILD; AMY'S OLDEST CHILD'S OLDEST CHILD'S OLDEST CHILD; and so on…). Likewise, in *LOT2* Fodor states that from the fact that concepts can't be learned "it doesn't quite follow that any concepts are innate… 'learned' and 'innate' don't exhaust the options" (p. 130). If the impossibility of concept learning for Fodor doesn't even entail that there are *any* innate concepts, then it doesn't make sense to hold that Fodor claims that *all* concepts are innate. The second problem for Rey's interpretation is that, whether or not the interpretation is textually accurate, it is difficult to see the value of labelling all learnable concepts as "innately possessed". On such an interpretation, extreme empiricist views and extreme rationalist views turn out to be *equally rationalist*. For example, the extreme empiricist view that all concepts are acquired by a single domain-general process of operant conditioning and the absurdly strong rationalist view that every lexical concept is innate in the sense that it appears in the acquisition base would have exactly the same concepts be innately possessed provided that the theorists who adopt these clearly opposing positions agree about which concepts can be acquired. In effect, Rey's interpretation of *LOT2*'s claim that "learning is impossible" amounts to nothing more than the claim that any concept that is ostensibly learned depends upon the learner's innate psychology—a very odd claim to take the phrase "learning is impossible" to express. Moreover, though, the idea that concept learning depends on one's innate psychology is something that essentially all parties to the rationalism-empiricism debate agree on. So the view that Rey attributes to *LOT2* turns out to be an uncontroversial truism that is of no real interest as far as the rationalism-empiricism debate is concerned.

Language of Thought to *LOT2*, Fodor has consistently held (1) that concept learning requires hypothesis testing, (2) that hypothesis testing models are circular in that they presuppose the concepts whose learning they are supposed to explain, and (3) that primitive concepts cannot be learned. Prior to *LOT2*, Fodor thought that complex concepts may offer some consolation—in effect, that the ABC model can underwrite a certain amount of learning—while going on to argue that the vast majority of lexical concepts lack the needed structure. But in *LOT2*—which arguably presents Fodor's most radical position on the origins of concepts, as well as his most trenchant critique of the possibility of concept learning—the view is that no concepts can be learned and that the detour into the question of whether lexical concepts have semantic structure was a misstep on his part. In short, Fodor has never wavered in his opposition to concept learning. With this overview of Fodor's influential and important arguments and positions on the origins of concepts, we are now ready to critically examine the arguments behind Fodor's views. Seeing exactly where they go wrong will allow us to better understand the limits and possibilities for how concepts can be learned.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0024

25
Not All Concepts Are Innate

It should be clear by now that our version of concept nativism should not be confused with Fodor's radical concept nativism or with his view that concepts cannot be learned. But it is one thing to distinguish our view from Fodor's and quite another to pinpoint where his case against concept learning goes wrong. In this chapter, we respond to Fodor's most recent and most comprehensive critique of the idea of concept learning—the argument in *LOT2*.[1] Seeing exactly how this critique goes wrong also helps to clarify how the different strands of Fodor's evolving views about concept acquisition interact and helps to disentangle the ones that turn out to be red herrings from the ones that raise important issues that need to be addressed. Carefully examining the elements of Fodor's critique will also give us an opportunity to further highlight the limitations of the Acquisition by Composition (ABC) model of concept acquisition, which has been widely endorsed by both rationalists and empiricists in debates about the origins of concepts, and to draw attention to further ways in which learning is central to our rationalist approach.

25.1 Learning Complex Concepts

Although *LOT2* doesn't differentiate between complex and primitive concepts, it will be helpful to consider them separately as they turn out to raise interestingly different sets of issues. We will begin in this section with complex concepts and will discuss primitive concepts in section 25.2.

In his earlier work, Fodor had little to say about complex concepts. But in *LOT2*, he explicitly and emphatically argues for the claim that it is impossible to learn any complex concept at all because such learning would presuppose that the learner already has the concept that is supposed to be learned. This circularity argument, which we outlined in the previous chapter, draws on two critical premises about learning (updated here to reflect the current focus on complex concepts):

[1] For a detailed critique of Fodor's earlier arguments against concept learning and his radical concept nativism, see Laurence and Margolis (2002).

1. Complex concepts cannot be learned via hypothesis testing.
2. There is no other way that a complex concept could be learned.
3. Therefore, complex concepts can't be learned.

The conclusion to this argument is, of course, utterly implausible. On the *LOT2* view, not only are all primitive concepts unlearnable, but so are all complex concepts. This means that it is no more possible to learn the concept NINETEENTH-CENTURY RED AND YELLOW STRIPED TEAPOT than it is to learn the concept RED. The fact that Fodor no longer infers that a concept is innate from its being unlearnable does little to mitigate the outrageousness of this conclusion. But again, it's not enough to point out that Fodor's argument is implausible. What is needed is an understanding of why the argument doesn't succeed and what this can tell us about the acquisition of concepts—in this case, the acquisition of complex concepts. We will argue that, in fact, *both* of the premises of the argument are false. Hypothesis testing models for learning complex concepts are perfectly viable, so complex concepts can be learned according to Fodor's preferred understanding of what learning requires. But complex concepts can be learned in other ways as well, so proponents of learning shouldn't feel restricted to the hypothesis testing framework. Our presentation of this part of the argument, which focuses on complex concepts, will have three parts: (1) *Hypothesis testing defended, 1*; (2) *Hypothesis testing defended, 2*; and (3) *Beyond hypothesis testing*.

Hypothesis testing defended, 1: A case of hypothesis testing where there isn't even the appearance of circularity. Let's start with the first of these premises. Fodor's case that complex concepts cannot be learned via hypothesis testing turns on his claim that hypothesis testing models are circular in that they presuppose the very concepts whose learning they are supposed to explain. Fodor's discussion of this claim in *LOT2* is organized around the example of someone trying to learn the concept GREEN OR TRIANGULAR. He argues that in order for someone to learn this concept via hypothesis testing, they would have to entertain the hypothesis that the concept to be learned is GREEN OR TRIANGULAR before the process of hypothesis testing and confirmation *can even begin*. And in order to be able to entertain this hypothesis, they would have to already have and be using the concept GREEN OR TRIANGULAR, since this concept occurs as part of the thought that expresses this hypothesis. But if a prospective learner must already have and be using the concept she is aiming to learn *before* she can even begin to learn it—before she can test and confirm the critical hypothesis—then she can't really be learning the concept through hypothesis testing. In sum, Fodor argues that hypothesis testing models of concept acquisition are circular since they require a prospective learner to already have and be using the concept whose acquisition they are supposed to explain.

We think the artificial nature of Fodor's example—learning the concept GREEN OR TRIANGULAR in an unspecified learning context—ends up obscuring

important features of the type of learning process it is supposed to exemplify. Fodor doesn't give *any* details about the learner's aims, the context in which she finds herself, or how she proceeds. We are just told that in order for her to learn GREEN OR TRIANGULAR via hypothesis testing, she would first have to formulate the hypothesis that GREEN OR TRIANGULAR is the concept to be learned and then go on to confirm that this hypothesis is correct. Moreover, it is hard to see how any learning process is even needed to acquire such a simple disjunctive concept. If you already have GREEN, TRIANGULAR, and OR, it seems as though they can simply be put together to form GREEN OR TRIANGULAR as soon as the need arises.

Accordingly, we will switch examples to a concept and a learning context that is representative of the sorts of real-world situations in which someone aims to learn a new complex concept. We will use the example of the concept of a new dance which a dance student tries to learn—and on the face of it, *does* learn—by representing to herself the component movements in the dance. We will start with a case where the concept is learned in such a way that there isn't even the appearance of circularity. In this type of case, Fodor's circularity argument can't even get a foothold. Later (in the section "Hypothesis testing defended, 2"), we will present a case where there is the appearance of circularity, but we will show that, in such cases, this appearance is misleading; when properly analysed, this type of case doesn't require learners to paradoxically already have and be using a concept prior to acquiring it via a learning process any more than cases without even the appearance of circularity do.

Consider, then, the situation in which a dance student is trying to learn a new dance. One way to do this (though by no means the only way) is for the student to try to master a complex concept that describes the sequence of moves that make up the dance. To make the situation as concrete as possible, suppose that our dance student is enrolled in a course by the Royal Scottish Country Dance Society but happens to miss the class that covers Maxwell's Rant, a dance that involves the following sequence: *reflection reels of three on opposite side, followed by reflection reels of three on own side, followed by crossing with right hands, followed by casting off, followed by a half figure of eight, followed by leading down the set, followed by casting up, followed by turning with right hands*. In discussing the student's attempt to learn the complex concept describing this sequence of dance moves, we will highlight the fact that *concept learning via hypothesis testing is a process that unfolds over time*, a fact that we take to be crucial to the proper analysis of Fodor's circularity argument.

Now the first case of concept learning we want to consider is one in which the question about whether the learner has the concept before learning it doesn't come up—there isn't even a hint of circularity in the learning process. For this case, imagine that the student watches from the sidelines while her classmates practice the dance the next time they meet. The dance is complicated. Not surprisingly, the student's first attempt at representing the sequence to herself isn't quite right and she is well aware of having made some mistakes. This leads her to

watch the dance again, to make some corrections, and then repeat the process, making further corrections and filling in the gaps in her representation of the sequence. Gradually she builds a more thorough and accurate representation and eventually is down to a single omission. Then she watches one last time and says to herself "Of course! Before the final turning with right hands, I need to cast up". Crucially, at this point she still hasn't explicitly entertained the final representation of the dance. But the realization that she has only missed this one step causes her to confidently formulate and adopt the complex concept that captures the full sequence of moves: REFLECTION REELS OF THREE ON OPPOSITE SIDE, FOLLOWED BY REFLECTION REELS OF THREE ON OWN SIDE, FOLLOWED BY CROSSING WITH RIGHT HANDS, FOLLOWED BY CASTING OFF, FOLLOWED BY A HALF FIGURE OF EIGHT, FOLLOWED BY LEADING DOWN THE SET, FOLLOWED BY CASTING UP, FOLLOWED BY TURNING WITH RIGHT HANDS.

The thing to notice about this example is that it is completely immune to Fodor's charge of circularity. This is because all of the confirmation of the relevant hypothesis takes place *before* the full and final concept is even explicitly formulated. The student in this example is justifiably confident about the concept she is to learn having gradually improved her representation of the dance sequence in response to successive viewings, all of which happens *before* she finally explicitly entertains the full concept. After the hard work of formulating and rejecting her earlier hypotheses, she knows exactly how to formulate the concept being learned, and it only remains for her to explicitly construct the concept and be done with it. But if all of the justification involved in learning the concept occurs *prior to constructing the concept*, then the concept doesn't have to be in place before it is learned. On the contrary, it is because of the justification that came before its appearance that the concept even enters the student's mind.[2]

What this example shows is that it is perfectly possible to learn a new complex concept through a hypothesis testing procedure. There is absolutely no threat of circularity in the account because the learned concept appears on the scene only after the justification occurs.

Of course, there are other possible ways in which the dancer might have come to possess the concept and where it might make sense to say that although the concept is acquired, it isn't really learned. For example, though it is highly implausible, the concept could have been innately specified—not in the trivial sense of merely being composed out of more basic representations which are innate, but in the highly substantive sense of the prefabricated complex concept

[2] Notice that the hypothesis confirmation in this example also doesn't proceed via enumerative induction (as Fodor suggests it must—see section 24.4 and the section "Beyond hypothesis testing" below). Fodor's focus on enumerative induction is another way in which his case is illicitly biased to make the circularity claim seem more plausible. Since many of the most interesting hypotheses in science and everyday life aren't confirmed by simple enumerative induction, this focus is clearly inappropriate.

being in the acquisition base.[3] Or the learner could have formulated the correct concept simply as a lucky wild guess, completely independent of the evidence, before she even saw the dance being performed. Neither of these situations is terribly likely for the complex concept in our example, but that isn't the point. What matters is just that, had they occurred, there would be little reason to say that the concept was learned. By contrast, in the example we are considering, the concept has a very different kind of origin with all the hallmarks of learning. It is assembled as the need arises, and the process that is responsible for its occurrence, as well as its persistence in the agent's mind, is one that is sensitive to the agent's observations and her previous attempts to accommodate them. It is hard to see what more learning would require.

Finally, it is worth adding that there is nothing particularly special about the concept that is learned in the example. The same considerations apply, at the very least, to any complex concept that describes a sequence of events in terms of a more basic stock of event types (e.g., complex concepts representing a new chess strategy, a new cooking recipe, a new type of knot, or a new chord change). In *LOT2*, Fodor characterizes his argument against learning as an a priori argument and claims to have located a confusion that is inherent to the hypothesis testing framework. But the example we have given in this section, and the range of cases it illustrates, shows that the claim that it is impossible to learn new complex concepts via hypothesis testing is false. And, of course, if complex concepts can be learned via hypothesis testing, then it can't be true *a priori* that they *can't* be learned via hypothesis testing. And it can't be that the very idea of concept learning is confused.

Hypothesis testing defended, 2: Defusing the appearance of circularity. The critical feature of the example we have just given is that the complex concept being learned isn't explicitly represented until after the justification that prompts its formulation and adoption has already taken place. However, many instances of concept learning don't work that way. Take, for instance, a simple modification to this example in which the student is unsure about the final correction to her representation of the dance sequence. She might formulate the correct hypothesis that captures the full and complete dance sequence but feel compelled to seek further evidence regarding its accuracy. In a case like this, which we take to be fairly commonplace, it would appear that when the student entertains this final hypothesis, she is thereby using the very concept she is supposed to be learning. But if she is already using the concept, isn't Fodor right to say that she doesn't really learn it?

[3] On the trivial understanding of "being innately specified", the concept PURPLE CAT is innately specified if the concepts PURPLE and CAT are innate. On the substantive understanding, PURPLE CAT itself would have to be innate.

We will use this second case to show how Fodor's charge of circularity can be defused even in instances where, on the face of it, there is the appearance of circularity since further confirmation is sought after a learner explicitly formulates the correct hypothesis about the concept she is trying to learn. The first step to seeing why the worry about circularity is misplaced even in this type of case is to take a closer look at the learning process in such cases.

The concept learning process in this case begins with a series of attempts by the student to learn the dance by learning a complex concept that represents its sequence of moves (just as in the example in the section "Hypothesis testing defended, 1"). During the course of these initial attempts, the student is guided by her observations of her classmates performing the dance as she considers and evaluates various hypotheses about this sequence. The formulation, testing, and rejection of a number of plausible but ultimately mistaken hypotheses is an integral part of the concept learning process as it plays a crucial role in the student's arriving at the correct hypothesis. We will call this initial phase of the concept learning process the *exploratory search phase*. The exploratory search phase naturally leads to the next phase of the concept learning process, in which the learner explicitly formulates a representation of the correct complex concept and hypothesizes that it is the concept she is aiming to learn. We will call this second phase of the concept learning process the *concept-formulation phase*. In the case we are now considering, the concept-formulation phase happens before the student has determined that the concept she suspects is the correct concept is indeed correct. She still needs to verify that it is the concept she is trying to learn. This verification, we are supposing, occurs subsequently in what we will refer to as the *final verification phase* of the concept learning process.

With this analysis of the structure of the concept learning process, we can return to the apparent circularity and the fact that it seems like learners must already possess a concept in order to learn it. The key to seeing where Fodor's argument goes wrong is to properly understand the relationship between the formulation of the correct complex concept (the concept the agent is trying to learn) in the concept-formulation phase and the entire learning process.

Recall that in Fodor's discussion of learning the complex concept GREEN OR TRIANGULAR, he focuses on the part of the learning process that occurs after the learner formulates the correct concept (GREEN OR TRIANGULAR) as a part of a hypothesis to be tested, and he effectively identifies the learning process with the subsequent verification of this hypothesis. This makes it seem as if the learner must already have the concept before the learning process could even begin—making it impossible to truly learn the concept. But this is very misleading precisely because a lot more happens in an ordinary instance of learning via hypothesis testing. When we take into account all aspects of the learning process, we can see that the formulation of the correct concept doesn't occur *prior to the*

learning process. Rather, it is something that occurs *during the learning process* and that is an essential *part* of that process.

This is all easier to see in a case where the concept learning context is better specified and some of the details of the concept learning process are fleshed out, like the situation we are imagining with the dance student in which all three phases of the learning process clearly contribute to the student's arriving at and adopting the right concept. The initial exploratory phase (in which the student entertains and rejects a number of concepts, getting closer and closer to the right one) is what allows the student to eventually construct the correct complex concept in the concept-formulation phase. The formulation and rejection of incorrect hypotheses and further observations during the initial exploratory phase provide justification for the correct hypothesis, and this in turn is a key factor leading to the formulation of this hypothesis. The final verification phase (in which she ultimately confirms that the correct concept is in fact correct) is a further important part of the overall learning process (see Figure 25.1).

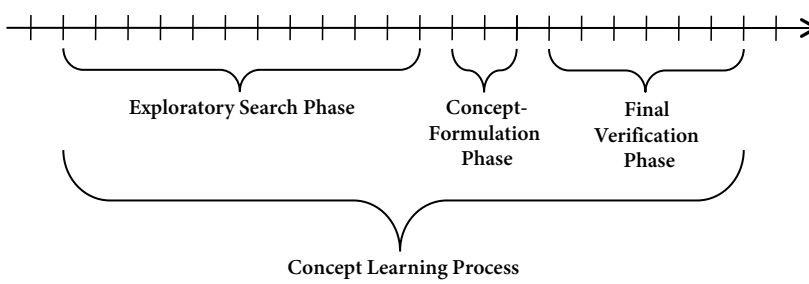

Figure 25.1 The typical structure of the concept learning process for learning a complex concept via hypothesis testing. The concept learning process has three phases. (1) In the exploratory search phase, the learner considers and rejects a number of hypotheses about the complex concept she is trying to learn. (2) In the concept-formulation phase, she explicitly represents the correct hypothesis but has yet to confirm it is correct. (3) In the final confirmation phase, she verifies that it is correct and accordingly retains the complex concept.

Once it is recognized that the concept learning process encompasses all three of these phases—that the learning process is not restricted to the final verification phase—the appearance of circularity completely disappears. It may be true that the student must acquire the concept prior to confirming that it is correct. But this doesn't mean that the concept is paradoxically "acquired before it is learned".

Instead, the correct way to see things is that the student acquires the concept *during the course of the concept learning process*—that acquiring the concept is a part of what happens when it is learned. And while the concept *is* still acquired prior to the completion of the learning process—how could it be otherwise?—this is in no way problematic, much less paradoxical.[4] When all aspects of the learning process are taken into account, Fodor's charge of circularity can be completely defused, even in cases where it may initially appear as though there is a problem of circularity.

What makes the acquisition of the target concept seem problematic in Fodor's presentation is that Fodor effectively identifies the entire concept learning process solely with the final verification phase, the point in which it is confirmed that the concept being considered is the correct one. But while this may seem plausible when the learning context is entirely unspecified, as it is in Fodor's minimal description of how GREEN OR TRIANGULAR might be acquired, it is not at all plausible when the learning context is fully spelled out, as in our example of the dance student learning the concept of a new dance.

Here is another way to see the force of our response to Fodor's charge of circularity. Consider how a hypothesis testing model of learning would apply to learning not just complex concepts but also complete thoughts, for example, the thought that *desert ants navigate using dead reckoning*. In learning how desert ants navigate, scientists relied on hypothesis testing.[5] Of course, in order to test the hypothesis that desert ants navigate using dead reckoning, they would have first needed to entertain this hypothesis in thought. Does this mean that these scientists didn't really *learn* that ants navigate using dead reckoning? Following Fodor, one could claim that they couldn't have learned it on the grounds that they had to already possess the thought prior to being able to carry out the hypothesis testing process that was needed to learn it. But no one would argue in this way because, in this case, it is obvious that the learning process is not confined to the final verification phase. Entertaining the hypothesis that desert ants navigate using dead reckoning is *part* of the learning process (something that happens *during* the learning process, not prior to it). The fact that the final confirmation of the hypothesis doesn't happen before the hypothesis is formulated in no way shows that it cannot be learned.

Fodor's circularity argument as it applies to complex concepts turns out to be grounded in a relatively simple confusion in which Fodor mistakenly identifies

[4] In an earlier paper, we presented several other complementary objections to Fodor's circularity argument (Margolis and Laurence 2011). We still endorse these further objections to Fodor's argument, but to simplify our discussion, we will not go into them here.

[5] We mentioned some of the evidence that confirmed the hypothesis about dead reckoning in Chapters 4 and 10.

the learning of a concept with just one part of a larger process. This is highly surprising given the enormous influence that Fodor's case against concept learning has had and the fact that the circularity argument has been so central to Fodor's scepticism about the possibility of concept learning since he first broached these issues in *The Language of Thought*. Why, then, has it been so hard to see this confusion?

In our view, a large part of the explanation is that the confusion is easier to see for the claim that complex concepts are unlearnable than it is for the claim that primitive concepts are unlearnable, and it is only in his discussion in *LOT2* that Fodor extended the circularity argument to cover complex concepts. The argument doesn't actually work against primitive concepts either.[6] But Fodor's discussion of the argument prior to *LOT2* was linked to further considerations—which are independent of the circularity argument itself—and which provide real insight into the problem of concept learning. In light of these further considerations, it *is* genuinely harder to see how a primitive concept—a concept that has no internal semantic structure—could be learned via hypothesis testing. In the case of a complex concept, one can imagine finding evidential support for elements of the concept without first having to represent the concept to be learned. This might allow the learner to represent a number of related hypotheses about the concept prior to representing it fully and correctly. But when the concept to be learned is a primitive concept, where would the correct hypothesis come from? It can't be pieced together in this way. Even the ability to make and remember observations that would point to the right hypothesis or provide evidential support for it would seem to depend on using the very concept that is supposed to be learned.[7]

These considerations have seemed to many to support the ABC model and the claim that primitive concepts can't be learned. Later, we will show that primitive concepts can be learned after all and that the ABC model isn't necessary for concept learning (see section 25.2). But we nonetheless think that there is an important kernel of truth in Fodor's circularity argument when applied to primitive concepts. This is that there is a real difficulty in explaining *how* and *why* learners would come to correctly hypothesize that a given primitive concept is the concept

[6] Although the flaw that we have identified in Fodor's circularity argument is easier to see when it is directed towards complex concepts, our criticism of the argument doesn't turn on whether the concept being acquired is complex or primitive. Rather, it turns on the fact that the learning process is not limited to the final verification phase, which is equally true in the case of primitive concepts. This means that Fodor's circularity argument leaves completely open the possibility that, in principle, even primitive concepts could be learned via hypothesis testing. But as we note in the text below, there is a different sort of problem about how primitive concepts might be learned via hypothesis testing, a problem that turns out to be closely connected to the argument from initial representational access.

[7] Recall from section 24.2 that this was part of Fodor's motivation in "The Present Status of the Innateness Controversy" to hold that the question of whether a concept can be learned turns on whether it has internal semantic structure (Fodor 1981).

to be learned. This problem is most pressing when a new primitive concept isn't part of a conceptual domain to which the learner already has some representational access. Essentially, then, the kernel of truth to Fodor's argument against the learnability of primitive concepts is the problem of initial representational access, which we discussed at length in Chapter 12.[8]

Returning to Fodor's circularity argument as applied to complex concepts, it seems that the lingering prominence of primitive concepts in Fodor's case against concept learning has led Fodor to greatly simplify what is going on when hypothesis testing is used to learn a complex concept. Since it is much harder to see what, if anything, the exploratory search phase of hypothesis testing might consist in when hypothesis testing is applied to primitive concepts, it is easy to overlook the significance of this phase of the process of concept learning when considering a complex concept in a purely schematic manner (as Fodor does in *LOT2*). But neglecting the exploratory search phase leads to a highly misleading picture of the concept learning process, lending an unwarranted degree of plausibility to Fodor's circularity argument against the possibility of learning complex concepts.

Beyond hypothesis testing. At this point, it should be clear why we should reject the first premise of Fodor's argument against the possibility of learning complex concepts. Not only is it possible to learn such concepts via hypothesis testing, but this type of concept learning is likely to be ubiquitous. Let's now turn to the second premise in Fodor's argument against the possibility of learning a complex concept—his claim that concept learning requires hypothesis testing. We will continue by arguing that Fodor is wrong about this point too. As with his first premise, seeing why we should reject it is illuminating about how concept learning actually works.

Fodor has always been emphatic that concept learning requires hypothesis testing. In *The Language of Thought*, Fodor goes so far as to claim that "there is only one kind of theory [hypothesis testing] that has ever been proposed for concept learning—indeed, there would seem to be only one kind of theory that is conceivable" (Fodor 1975, p. 36). In *LOT2*, he says that hypothesis testing is "the only candidate account" of concept learning (Fodor 2008, p. 139) and that there is a consensus in cognitive science that concept learning is based on hypothesis testing. However, he also notes that "though this consensus is pretty general, it's much more often than not inexplicit. There are very, very many theorists who

[8] Crucially, however, unlike Fodor's circularity argument, the argument from initial representational access (as we develop it) is not intended as an a priori deductive argument, and it isn't intended as a general argument applicable to all primitive concepts. While it does argue that some primitive concepts are not learned but innate, it in no way aims to show that it is impossible to learn primitive concepts—which is a good thing since, as we will argue, it's not impossible to learn new primitive concepts.

accept HF without fully realizing that it's HF that they accept. I imagine, indeed, that that's the usual case" (p. 132).[9]

Whether Fodor is right about the cognitive science community's unvoiced commitments, we ought to ask why Fodor thinks that hypothesis testing models are all but inevitable. After explaining a hypothesis testing model in *LOT2*, Fodor considers the question, "What's the evidence that children (for example) actually do learn concepts by some sort of induction?" (p. 136). He responds to this question as follows (p. 136):

> Fair question…as far as anybody knows, there is simply no alternative. The only reliable way to infer from a batch of singular beliefs (*this instance of* EMERALD *is green; that instance of* EMERALD *is green; that other instance of* EMERALD *is green;* etc.) to general conclusions (EMERALD applies to *green things*) is to take the truth of the former as evidence for the truth of the latter. So either concept learning is what HF says it is or there isn't any such thing.

Unfortunately, this response isn't very helpful. The claim that there is no alternative "as far as anyone knows" is controversial at best—as evidenced by the fact that, by Fodor's own admission, few in the cognitive science community see themselves as advocating hypothesis testing models of concept learning. And the remainder of Fodor's response simply *presupposes* that concept learning has the structure of an enumerative inductive inference. It may be that the only way to infer a general conclusion from a batch of singular beliefs is to regard the singular beliefs as evidence for the general conclusion. But why should we suppose that concept learning takes this form to begin with?

Still, there is another, more general, consideration that Fodor cites in *LOT2* (which we noted in Chapter 24), and this seems to be his driving motivation. He points to the need to distinguish genuine cases of learning from cases where a concept is acquired yet not learned, for instance, where a concept is acquired through "surgical implantation", "swallowing a pill", or "hitting one's head" (p 135). The problem that Fodor sees here is that whether a concept is learned or merely acquired isn't just a matter of whether the agent enters into a causal interaction with the environment. Causal interactions occur in both sorts of cases. So what distinguishes the instances where concepts are learned? For Fodor, it's that learning is a rational process in which the interactions with the environment provide the agent with evidence concerning the concept that she acquires. In contrast, surgical implantation and the like are, as Fodor likes to put it, *brute-causal processes*.

To this consideration, our initial reply is to note, with several of Fodor's other philosophical critics, that while Fodor is certainly right that learning must be distinguished from brute causation, there is a considerable amount of logical

[9] Recall from the previous chapter Fodor uses "HF" for the view that concept learning involves hypothesis testing and confirmation and that it amounts to an inductive process.

space between brute-causal processes and explicit hypothesis testing (Samet and Flanagan 1989; Sterelny 1989). As these earlier discussions point out, neither the ordinary common-sense notion of learning nor its application in psychological research is restricted to cases of hypothesis testing. A cursory look at any introductory psychology textbook provides a wealth of examples where this is evident. Take, for instance, rote learning or the learning of facts. In these cases, information is recorded and cognitive processes such as rehearsal ensure that the information is retained for future use. But there is absolutely no reason to suppose that hypothesis projection and confirmation is required. In learning someone's phone number, it's not as if one first has to hypothesize what the number is and then seek evidence to confirm that this hypothesis is correct.

We can see, then, that Fodor doesn't have much to back up the strong claim that hypothesis testing is mandatory for learning. But what isn't clear yet is how any of this translates into an alternative model of *concept* learning, as opposed to learning in other domains (e.g., fact learning or skill learning). In the remainder of this section, we will sketch a few ways in which complex concepts, in particular, can be learned that do not involve hypothesis testing. But first it is important to note another crucial though implicit feature of Fodor's argument.

In characterizing what is required for learning, Fodor moves swiftly from the need for a learner's activities to be rational in some minimal sense that contrasts with acquisition via brute-causal processes to the claim that her observations must count as evidence for what she acquires, and from here to the thought that the learner has to register the evidence by computing its bearing on an explicitly entertained hypothesis. While each step in this chain of inferences might be questioned, the point we want to call attention to is that Fodor is presupposing an *internalist* as opposed to an *externalist* approach to justification. The difference between these is that while an internalist approach requires an agent to represent and grasp the reasons that justify a judgement, an externalist approach doesn't—it only requires that the agent's judgement is arrived at by means of a reliable process.[10] An externalist approach to the justification involved in concept acquisition would be one in which what matters isn't so much that the agent recognizes and explicitly represents the evidential value of what she observes, but rather that she employs cognitive mechanisms that deliver new concepts through processes that reliably reflect appropriate environmental contingencies. Whatever one thinks about epistemic justification more generally, such justification is sufficient to distinguish cases of rational concept acquisition (ordinary cases of concept learning) from the cases in which a concept is acquired through Fodor's brute-causal processes (a bump on the head, surgical implantation, and so on). In any case, it is useful to consider alternatives to Fodor's hypothesis testing that adopt an

[10] See Goldman (1986) and Kornblith (2002) for examples of externalist accounts of epistemic justification.

externalist criterion. With this preliminary point in mind, let's consider a couple of examples of different ways in which complex concepts can be learned without hypothesis testing.

Perceptual learning. Consider what happens when someone forms a new complex concept as a result of perceiving an object or event that manifestly exhibits the combination of properties that the concept picks out. For instance, someone who encounters a black swan for the first time is likely to form the concept BLACK SWAN. She needn't have had the concept prior to her encounter, and it might never have occurred to her that black swans are a real possibility. But if she has the concepts BLACK and SWAN and sees a black swan with her own eyes, she will come to possess BLACK SWAN and will be prepared to record and organize new information about these unexpected creatures. Learning a new concept in this way is largely a matter of perceiving one's surroundings and being open to the new arrangements they present. Moreover, the designation *learning* is perfectly apt. Acquiring BLACK SWAN upon seeing some black swans isn't anything like receiving a surgical implant or miraculously acquiring the concept by being hit on the head in just the right way. If we adopt the externalist approach to these matters, we can even say that the process is rational in the relevant, attenuated sense. After all, the various perceptual and cognitive processes that support the concept's acquisition reliably yield accurate descriptions of the objects and events with which they causally interact. In fact, this is exactly what such systems are supposed to do. Perceptual-based learning of complex concepts is a paradigm of acquiring new concepts through the operation of psychological operations that have the function of extracting categorical information from the environment.

Communication-based learning. Concept learning can also be supported by explicit verbal instruction and communication. Think about what happens in a university classroom. In a logic course, the instructor might convey the definition of validity by saying that a valid argument is one in which, if the premises are true, the conclusion must be true as well. A good attentive student might thereby *learn* the concept VALID ARGUMENT (or the concept ARGUMENT IN WHICH, IF THE PREMISES ARE TRUE, THE CONCLUSION MUST BE TRUE AS WELL). Of course, this learning will later be reinforced through examples and embedded into a broader knowledge of logical concepts, but the verbal communication alone can be sufficient for the student to have learned the concept from the teacher. What's more, learning via communication isn't by any means confined to the classroom or to concepts that are especially difficult to master. Think about the ordinary situation in which a friend describes how to make a new pasta sauce that you happen to be interested in. Hearing the verbal description all by itself allows you to construct a complex concept that represents all the steps in the recipe. And as with instances of learning by perceiving, these sorts of cases aren't anything like acquiring a concept through brain surgery. The language processing mechanisms that direct the concept's acquisition have the function of extracting information from the

linguistic environment and of delivering new concepts corresponding to the linguistic input they are given.

What can Fodor say about these sorts of cases? Because he maintains that there are no alternatives to hypothesis testing for explaining how concepts are learned, he only has two options. (1) He can accept that these are examples of concept acquisition that don't involve hypothesis testing yet go on to deny that they count as genuine examples of learning. Alternatively, (2) he can insist that, despite appearances, our examples covertly involve hypothesis testing after all. Neither of these responses is especially plausible, however.

We have already touched on the first. The problem for Fodor is that the examples we have given are nothing like the brute-causal processes that he contrasts with genuine cases of learning. In most of the brute-causal scenarios that he mentions, it's little more than a lucky coincidence that the outcome is a new concept, much less one that is relevant to the agent's situation. (In cases such as miraculously learning a concept by being hit on the head in just the right way, the coincidence isn't just lucky; it's so far-fetched that it's hard to take seriously. But we won't press that point here.) By contrast, consider once again what goes into learning a new concept by means of explicit instruction and verbal communication. The agent may be fortunate to have access to the right teacher or the right conversational partner, yet once this other person says what she has to say, luck drops out of it altogether. It's because the learner has the right cognitive equipment—including language processing systems—that she is able to arrive at the needed concept. This type of acquisition involves cognitive and perceptual processes that have the function of producing new complex representations on the basis of relevant information that is reliably extracted from the environment.

The more interesting response, and the one that Fodor shows some sympathy for, is to claim that our examples of learning hide a critical dimension in which hypothesis testing is going on.[11] Recall Fodor's remark that "there are very, very many theorists who accept HF without fully realizing that it's HF that they accept" (Fodor 2008, p. 132). The problem with this response, however, is that there is no need for any hidden hypothesis testing, and no reason to believe that it occurs in the sorts of cases we have discussed.

Consider again perception-based concept learning of the kind that is illustrated by the black swan example. In cases like this one, learning a new complex concept is a matter of assembling the concept that corresponds to a perceived object or event. Now in principle it could be that what happens is that the agent initially formulates a hypothesis concerning the identity of the concept in question—*that the concept that is instanced before me is the concept* BLACK SWAN—and then proceeds to test this hypothesis against further

[11] If Fodor took this route, he'd surely add that our examples don't count as genuine cases of learning in the end, even with the covert hypothesis testing, since he maintains that hypothesis testing models are circular. But we have already dealt with the circularity charge.

observations. But there is no need at all for this extra layer of reflection. What is far more plausible is that the visual system simply registers the presence of a black swan, automatically resulting in construction of the mental representation BLACK SWAN. This representation would then serve as a nexus for organizing further incoming information about the encountered animal or others like it. Or consider the sort of case where someone learns a new concept through verbal communication. The learner hears her friend produce a verbal description and as a result manages to learn a concept of interest right there on the spot. Now our learner could, in principle, explicitly represent and entertain various different possibilities regarding the concept to be learned and then seek to test these possibilities through further observation. But it's unlikely that anything like this goes on in the normal case since the new concept would become available to the learner anyway as soon as she understands the speaker's words. Nothing further needs to be done.[12]

This concludes our critique of Fodor's argument as it applies to complex concepts. We have shown that, contrary to what Fodor has claimed, hypothesis testing models needn't be—and typically aren't—circular. There are plausible cases of concept learning via hypothesis testing where there isn't even the appearance of circularity because the full content of the hypothesis is confirmed prior to the explicit formulation of the complete hypothesis. In other cases, although there may be the superficial appearance of circularity, this is readily defused once it is recognized that the concept being learned isn't (and doesn't have to be) acquired *prior* to the start of the concept learning process (as Fodor implicitly supposes), but instead is acquired *during* the learning process as a key part of that process. We have also shown that hypothesis testing is not necessary for learning complex concepts. In particular, we have outlined a number of processes that don't involve hypothesis testing but that are well suited to explaining how certain complex concepts are learned.

25.2 Learning Primitive Concepts

We have seen that Fodor's argument against concept learning doesn't work for complex concepts, but what about primitive concepts—concepts that do not have

[12] What if Fodor were to concede that the agent doesn't *explicitly* engage in hypothesis testing, but were to insist that she does so *implicitly?* This won't help. Note that "implicit" here can't simply mean that it is done unconsciously. Our discussion has been completely neutral on whether the acquisition processes happen at the level of consciousness or not. Rather, "implicit" would have to mean that the learning process, however it works, produces the same result *as if* hypothesis testing were to take place. But if the process merely has to produce the same results as hypothesis testing and doesn't have to take the explicit form of hypothesis testing, then there is no reason to suppose that the concept must be entertained in the course of the concept learning process. And if the concept needn't be explicitly represented in the course of concept learning, there is no reason to suppose that there is any circularity in the learning process at all, so Fodor's argument would collapse immediately.

any internal compositional semantic structure? Though few theorists have endorsed Fodor's general anti-learning argument, a great many have agreed with his claim that primitive concepts at least can't be learned. This is because of the widespread commitment to the ABC model. As discussed earlier, according to this general approach, when a new concept is learned, it must be assembled from its semantic constituents, and while these may in turn be assembled from yet more basic representations (and so on), eventually all learned representations are dependent upon an innate stock of primitive representations from which they are composed. In this section, we will show that Fodor's scepticism about learning primitive concepts, and the ABC model that goes with it, are both mistaken. Primitive concepts can be learned even though (being primitive) they can't be assembled from simpler concepts.

To see how learning a primitive concept is possible, the place to begin is with a theory of what makes primitive concepts have the semantic content they do.[13] This is because we need to be clear about the conditions that must be met for possessing these concepts. Given an explicit specification of these conditions, we can then ask how an agent's mind comes to satisfy them and whether there is any way that it could be done through learning. In earlier work, particularly in Margolis (1998), we developed a model of just this sort, based on Fodor's own theory of content—his *asymmetric-dependence theory* (Fodor 1990a, 1990b). The model was designed to explain how an agent could learn natural kind concepts (e.g., concepts for different types of animals), treating these as primitive concepts. We chose Fodor's theory of content because it is specifically intended to apply to primitive concepts, as Fodor devised the theory after coming to believe that most lexical concepts are primitive. Putting aside many of the details, the core idea of Fodor's theory of content is that a concept expresses the property that it is causally dependent upon, in that instances of the property reliably cause the concept to be activated. For example, the concept ZEBRA expresses the property *zebra* because zebras reliably cause the activation of ZEBRA. That's not enough, of course, because things that aren't in the extension of a concept may reliably cause it to be activated too (e.g., horses in particular perceptual conditions). A second component of the theory deals with these cases by saying that the regularities that subsume them are dependent upon the causal link between zebras and ZEBRAS, and not the other way around. As Fodor would put it, the horse/ZEBRA regularity is asymmetrically dependent upon the zebra/ZEBRA regularity.

One important feature of Fodor's theory of content is that it analyses content in terms of mind-world causal relations, abstracting away from the internal processes that occur within an agent's mind. So it allows for the possibility that two

[13] As explained in Chapter 6, such theories are known in the philosophical literature as *theories of content*. They aim to explain such things as what makes the concept ZEBRA be about zebras and the concept SHOE be about shoes.

people can both possess ZEBRA and yet have very different beliefs and correspondingly different inferential dispositions regarding zebras. At the same time, it is important to remember that, even on such an account, the internal processes aren't irrelevant. This is because the mind-world relations that are constitutive of content have to be brought about somehow, and this will typically happen because of the various inferential connections that are associated with a concept even if these inferences are only contingently associated with the concept. We will continue to use the terminology we introduced in Chapter 5 and refer to these systems of inferential connections as *sustaining mechanisms*. Sustaining mechanisms establish and preserve the mind-world relations that constitute a concept's content.

Given the need for sustaining mechanisms, the question of how a primitive concept is acquired can be recast as the question of how one (or more) sustaining mechanism for the concept is acquired. While there are many different types of sustaining mechanisms, a useful point of orientation is to focus on one particularly interesting type, which we call a *syndrome-based sustaining mechanism*. This type of sustaining mechanism supports the possession of a natural kind concept by combining two fairly general sorts of cognitive dispositions. One tracks certain readily observable properties of a kind, while the other embodies the tendency to view the kind as having an essence or underlying reality that is shared by all of its members and that causally explains its observable properties. Together these dispositions allow for the construction of a sustaining mechanism for a concept like ZEBRA. In the standard case, encounters with zebras would cause an agent to record salient features of their appearance (their shape, motion, colour markings, etc.) and associate these with a new representation (ZEBRA) that would become activated by things with much the same appearance. Appearances can be deceiving, however. Under some conditions, an ordinary horse might suffice to activate ZEBRA. But this is where the essentialist disposition does its work. The agent would be disposed to withdraw the judgement that something falls under the concept ZEBRA upon learning information that would indicate the absence of the right essence, making the horse/ZEBRA connection less basic than (and thus asymmetrically dependent on) the zebra/ZEBRA connection.[14]

There are two key points to notice regarding this proposed model. First, it doesn't make use of any Fodorian hypothesis testing. To acquire ZEBRA, the agent needn't explicitly formulate and confirm a hypothesis *that the concept to be learned is the concept* ZEBRA. Indeed, she may never come to explicitly represent a hypothesis of this type at any point in the process of learning the concept.

[14] For example, even if they have no explicit views about what the essence of a zebra is, they may suppose that this essence is transmitted from parents to offspring at birth, in which case they might suppose that this animal doesn't have zebra essence after all when they discover that its parents are horses. For research on this sort of inference in children, see Keil (1989) and Gelman (2003).

Second, despite the lack of hypothesis testing, there is every reason to believe that the model is still a learning model. For one thing, it doesn't imply that ZEBRA and the like are simply part of the acquisition base—psychological processing is still critical to their acquisition. It's also not as if the model says that a sustaining mechanism for ZEBRA is all wired up in advance and simply waiting for an innately specified triggering condition to cause it to become activated. Far from it. What is innate, according to the model, is a general cognitive organization for creating a range of syndrome-based sustaining mechanisms in response to new natural kinds. The reason this organization leads to the creation of a sustaining mechanism for ZEBRA, and thereby explains the acquisition of ZEBRA as opposed to some other concept, is because of the particular sorts of cognitive processing that this organization initiates given causal contact with zebras. If the same organization were brought to bear in encounters with other animals (e.g., lions or giraffes), then the result would be concepts for these other kinds of animal instead. As with the learning models for complex concepts discussed in the previous section, the mechanisms involved in the acquisition of the concept have the *function* of producing new representations on the basis of relevant information that is systematically and reliably extracted from the experienced environment. And again, like the learning models for complex concepts discussed earlier, the construction of a syndrome-based sustaining mechanism, as we are envisioning it, stands in clear contrast with Fodor's brute-causal processes (futuristic neurosurgery, a miraculous blow to the head of just the right sort, etc.).

In a past discussion, we were content to argue that a model along these lines suffices to show that even primitive concepts can be learned (Laurence and Margolis 2002). But in *LOT2*, Fodor responded with a number of objections, arguing that our model has some crucial hidden flaws and consequently that there is still no hope for the idea of learning a primitive concept. We are also aware that some readers who don't share Fodor's particular worries but who are partial to the ABC model of concept acquisition might wonder how much our response to Fodor really shows. Is it limited to just natural kind concepts? And does it require a commitment to Fodor's asymmetric-dependence theory of content? We will address the questions about whether our account is limited to natural kind concepts and to Fodor's asymmetric-dependence theory of content in the next section. But first, we need to address Fodor's objections, which we will do in the remainder of this section. Working through his objections will involve taking a more detailed look at the question of how sustaining mechanisms relate to concept acquisition.

Fodor's first objection challenges our use of his asymmetric-dependence theory and its commitment to an externalist treatment of conceptual content (which he calls *semantic referentialism*).[15] Fodor charges that we "overstate the case for

[15] "Externalism" has (at least) two distinct meanings in philosophy. In our critique of hypothesis testing models above, we made reference to externalist theories of *justification*. (These are

semantic referentialism" (p. 141), as his argument against concept learning is neutral about the type of semantic content that concepts have. The general form of argument that Fodor claims to be using comes out just a bit earlier in the text. In a discussion couched in the form of a dialogue, he asks what could make it the case that an agent learns one of a pair of coextensive concepts rather than the other. Fodor answers as follows (2008, pp. 134–135):

> I mean, mustn't it be something like this: in consequence of their respective experiences, one learner comes to think to himself 'All those things are Cs' but the other comes to think to himself 'all those things are C*s'?…And is not inductive inference the process par excellence by which one proceeds from representing some things as Cs to representing all such things as Cs?…And is not the formation and confirmation of hypotheses the very essence of inductive inference?…So, does not the fact that it is possible to learn one but not the other of two distinct but coextensive concepts show that concept learning is indeed some kind of hypothesis testing?

Fodor's argument in this passage is that the only thing that could make it the case that an agent learns one of a pair of coextensive concepts rather than the other—C as opposed to C*—is that an inductive inference supports a process of learning C rather than C* via hypothesis testing. We agree with Fodor that for an agent to learn C rather than C*, the learning process involved needs to lead to the acquisition of C rather than C*. But there is nothing in what Fodor says here that amounts to an argument that concept learning requires *hypothesis testing*. Fodor is simply presupposing that this is all it can be—despite the fact that the whole point of the model that we introduced was that it presents an *alternative* to learning via hypothesis testing. It's no good replying to a proposed alternative by insisting that learning must take the form of hypothesis testing.

Still, the attention Fodor gives to coextensive concepts suggests a more pointed criticism of our model. This is that our sustaining mechanism account of concept acquisition is problematic since it can't distinguish C from C*. In response to *this* worry, we should first mention that in earlier work we were at pains to emphasize that our use of Fodor's asymmetric-dependence theory is only for illustrative purposes and that we are not committed to this particular theory of content

epistemological theories that hold that a person can be justified without explicitly representing and grasping reasons in support of a held view.) In contrast, the type of externalism that is relevant to the present issue—*semantic* or *content externalism*—has to do with the nature of mental content. An externalist theory of mental content holds that the content of a mental state is (or is largely) constituted by the relation between the state and aspects of the world (e.g., a mind-world causal relation). By contrast, *internalist* theories of mental content focus exclusively on mind-internal relations that characterize a mental representation's role in cognition.

(Laurence and Margolis 2002).[16] But regardless, the challenge pertaining to the learning of coextensive concepts is one that advocates of the asymmetric-dependence theory (or other forms of content externalism) can readily handle without having to capitulate to Fodor on the question of whether learning requires hypothesis testing. All that an externalist needs to do to address the problem of coextensive concepts is to maintain that concept identity isn't solely a matter of extension. When Fodor isn't arguing about whether concepts can be learned, he himself has been clear that externalists can avail themselves of the notion of a mode of presentation. For Fodor, modes of presentations are realized by the formal (i.e., non-semantic) properties of the mental representations in which thinking takes place (Fodor 2008, ch. 3). Alternatively, externalists can say that a concept's identity is partly constituted by its conceptual role whether or not conceptual role is taken to be part of a concept's content (Margolis and Laurence 2007a). Thus it's open to an externalist to say that what makes it the case that one acquires C as opposed to the coextensive C* is that the process of acquisition in the first case results in a representation that is partly constituted by the C-formal or C-conceptual-role properties (as opposed to the C*-formal or C*-conceptual-role properties) together with C's extension. Assuming that the acquisition involves a process akin to the construction of a syndrome-based sustaining mechanism, Fodorian hypothesis testing needn't come into it.[17]

[16] For example, preceding our demonstration that Fodor's asymmetric-dependence theory can be used in a learning theory, we said the following: "To see how this general strategy plays out in concrete terms, we will discuss how primitive concepts might be acquired under a specific theory of content. We want to emphasize, however, that the specific theory of content and the particular account of acquisition that we discuss are simply illustrations of our strategy for addressing Fodor. The sample theory of content which we will use is Fodor's own theory" (Laurence and Margolis 2002, p. 36). For a sketch of how similar learning models can be developed for other theories of content, see section 25.3 below.

[17] Rey (2014) raises a related argument against our account, charging that it overlooks significant ways in which concept acquisition is underdetermined by the information that is available to a learner. He claims that our account is unable to explain why a learner acquires ZEBRA when encountering zebras as opposed to a more general concept such as MAMMAL or ANIMAL (this charge is a version of what has come to be known as the "*qua* problem"). Likewise, Rey claims that our account can't explain why the learner acquires ZEBRA as opposed to ZEBRELEPHANT, which refers to all and only zebras encountered before one's 16th birthday and elephants thereafter (an example inspired by Nelson Goodman's extremely influential new riddle of induction; Goodman 1954). In making these charges, Rey isn't giving full due to the ample psychological resources that our account can draw upon. For example, it's easy enough to see that the syndrome-based sustaining mechanism for ZEBRA would differ from likely sustaining mechanisms for MAMMAL, since the features that the two mechanisms would encode would be systematically different, and consequently the inferential dispositions that control the activation of these concepts (and the properties they respond to) would be correspondingly different. To the extent that underdetermination poses any residual difficulty, we see this as a general problem that affects all theories of conceptual content and hence all theories of how concepts are acquired—both rationalist and empiricist. Moreover, any such residual difficulties would seem to apply equally to innate and learned concepts, since the content of concepts stands in need of explanation equally whether the concepts are innate or learned. Accordingly, despite Rey's suggestion to the contrary, there is no reason to suppose that underdetermination of this sort is a particular problem for our theory. Rey provides no account of how innate concepts (or concepts composed of innate concepts) would be immune to such problems, nor does he show why if there were such an account, it couldn't be extended to learned primitive concepts.

Fodor's second objection goes right to the heart of our earlier model by challenging whether our sample account of how these primitive concepts are learned really counts as a *learning* theory. As Fodor puts it,

> If the sort of referentialist/atomist story about conceptual content that Margolis and I like is true…then learning a theory can be (causally) sufficient for acquiring a concept. But it doesn't follow that you can learn a concept. So here we are, back where we started; we still don't have a clue what it might be to learn a concept. (Fodor 2008, p. 144)

This objection requires a bit of unpacking. In *LOT2*, Fodor cites a pair of related considerations that are supposed to show that acquiring a sustaining mechanism and learning a concept are entirely different things. The first is that even if a sustaining mechanism is learned, there is no guarantee that the concept it gives rise to is learned as a result (Fodor 2008, p. 144):

> "You can learn (not just acquire) A" and "Learning A is sufficient for acquiring B" just doesn't imply "You can learn B". For, the following would seem to be a live option: If you acquire a concept by learning a theory, then something is learned (namely, the theory) and something is (merely) acquired (namely, the concept); but what is learned isn't (merely) acquired and what is (merely) acquired isn't learned. To acquire the concept C is to lock to the property that Cs have in common; and such lockings may be mediated by theories. The theory that mediates the locking between the concept and the property that the concept is locked to may be, but needn't be, rational, or coherent, or well evidenced, to say nothing of true. That's why Ancient Greeks, who thought stars were holes in the fabric of the heavens, could nevertheless think about stars.

Fodor uses the term "theory" as shorthand for what we are calling a sustaining mechanism. In the sort of case that he is imagining, although the sustaining mechanism is learned, it is just a lucky break that the sustaining mechanism provides semantic access to the property that the concept expresses. Much of the information in the sustaining mechanism is false and the sustaining mechanism is irrational or lacking in coherence and evidential support. The problem Fodor sees this consideration raising is that if it is, as it were, an accident that the sustaining mechanism mediates access to the property, how can the agent be said to have learned the concept?

The second, related consideration Fodor raises is meant to increase the gap between acquiring a sustaining mechanism and learning a concept. It does this by reintroducing a familiar worry (Fodor 2008, p. 144):

> [Y]ou can [also] acquire a concept by *acquiring* a theory (i.e., by acquiring it but not learning it). I'm dropped on my head and thereby acquire the geocentric theory of planetary motion, and thereby become linked to, say, the property of being a planet. In such cases, neither the theory I've acquired nor the concept I've acquired has been learned.

This time the problem isn't that it's odd and unexpected that the sustaining mechanism provides semantic access to the property its concept expresses. It's that the way that the sustaining mechanism itself is introduced is the result of a causal fluke. But if the sustaining mechanism isn't learned, if it just pops into someone's head, how can the concept that it supports be learned as a consequence?

Taken together these two worries are intended to show that primitive concepts can't be learned by acquiring a sustaining mechanism and consequently that the attention we have given to sustaining mechanisms is wrongheaded.

Neither of these considerations succeeds at undermining our account as a learning model. Indeed, whatever plausibility they have as counterexamples to our model of concept learning rests on a misunderstanding of our model and the dialectic it is meant to address. Recall that Fodor's argument against concept learning is intended to establish that it is *impossible* to learn any new concept. So in response to our model of concept learning, it is no good for Fodor to argue that some cases of acquiring a sustaining mechanism fail to constitute learning a concept. He needs to show that *every case* of acquiring a sustaining mechanism fails to constitute learning a concept—that is, *that there aren't any cases at all* where acquiring a sustaining mechanism would count as learning a concept. One cannot possibly show that it's impossible to learn a concept through acquiring a sustaining mechanism by showing that a few highly idiosyncratic instances of this sort fail to count as learning. One might as well argue that since penguins can't fly, it's impossible for birds to fly.

Our argument against Fodor's case for the impossibility of concept learning centred on a specific instantiation of our general sustaining mechanism account. (We never claimed that acquiring a sustaining mechanism for a concept qualifies as an instance of concept learning no matter what the sustaining mechanism is like or how it is acquired). Accordingly, for Fodor to rebut our argument, it is incumbent upon him to show that a concept could not be learned *in the sort of case we outlined*—a case much like the one we described above in introducing the idea of a sustaining mechanism. Here the learner acquires the sustaining mechanism for a new concept (e.g., ZEBRA) directly through ordinary perceptual causal contact with instances of the concept, accurately recording relevant observable properties of the kind, and all of this takes place under the direction of a general intention to learn about this new kind. It is not an improbable and fortuitous circumstance that the information in the sustaining mechanism manages to

establish a nomic dependence between the concept and its instances, nor is it a causal fluke that the sustaining mechanism is acquired at all. The information recorded is relevant information that is acquired through perceptual and cognitive processes that have the function of recording such information and of organizing it into new representations of natural kinds. To address our challenge, Fodor must demonstrate that it is impossible to learn a new concept when a syndrome-based sustaining mechanism is acquired in *this* sort of paradigmatic case; Fodor's fluky cases are simply irrelevant given a proper understanding of the dialectic.

Fodor's third and final objection directly challenges our claim that the processes involved in our syndrome-based sustaining mechanism account could truly count as *learning*. Throughout his discussion in *LOT2*, Fodor maintains that hypothesis testing is the only conceivable model of concept learning. Accordingly, Fodor should be expected to resist our claim that concepts can be genuinely learned, as opposed to merely acquired, via a process of this kind. In fact, in a conference that was dedicated to Fodor's views on concept acquisition, Fodor not only sounded as if he took it to be an a priori truth that learning requires hypothesis testing but also as if he thought that our model should be rejected barring an alternative *definition* of learning (i.e., a definition that addresses the motivations that originally prompted the hypothesis testing analysis and that can serve as a principled guide for identifying when learning occurs).[18]

We confess that we don't have a definition to offer. But it would be deeply ironic if Fodor, of all people, held that against us. A central theme throughout Fodor's work has been that the search for definitions is almost always futile. Fodor has been especially critical of the idea that lexical concepts, in particular, can be defined.[19] But if most ordinary lexical concepts can't be defined, it's hardly fair to ask us to provide a definition of LEARNING. Instead, the assumption ought to be that LEARNING *cannot* be defined.

What's more, a brief inventory of different types of learning suggests that LEARNING picks out a rather heterogeneous set of phenomena. Consider the diversity that is manifestly associated with rote learning, learning a language, learning a complex manual skill (e.g., how to play the violin), learning the contents of a room, learning a novel route to an old location, learning algebra at school, learning what an avocado tastes like, learning which kinds of animals are dangerous, and learning an implicit cultural norm. These diverse phenomena are all natural to describe as cases of learning, but there is no reason to suppose that the underlying processes share a common defining set of features. (They certainly

[18] Symposium on *Solutions to Fodor's Puzzle of Concept Acquisition*, Annual Cognitive Science Society (2005). The transcript is available online: https://www.academia.edu/12356657/Solutions_to_Fodors_Problem_of_Concept_Acquisition_-_Transcript

[19] In "The Present Status of the Innateness Controversy", Fodor says, "I once heard Professor Gilbert Harman remark that it would be *surprising* if 'know' were definable, since nothing else is. Precisely" (Fodor 1981, p. 285). See also Fodor et al. (1980), which is aptly titled "Against Definitions".

don't all involve hypothesis testing, which they should if Fodor's views about learning were correct.)

As we emphasized in earlier chapters, to the extent that the scientific community is beginning to understand what goes on with these and other cases of learning, it is becoming increasingly plausible that specialized learning mechanisms are often responsible in good part for our abilities in different task domains. For example, much of language acquisition depends on domain-specific learning mechanisms, so there is no reason to suppose that learning English is accomplished in quite the same way as, say, learning a new route home. Even within a given task domain, the same outcome can depend upon rather different mechanisms. One could learn the contents of a room by opening the door and looking inside, but also by seeking a knowledgeable person's testimony. Though the result may be much the same, the mechanisms involved are strikingly different. Or to take another example, one could learn a route home by exploiting memorized landmarks but also by relying on dead reckoning—again, very different underlying mechanisms. The more we discover about the mind, the more we have to face the fact that there may be very little that all cases of learning invariably have in common.

Still, perhaps something can be said about some of the characteristics of typical instances of learning. This isn't to give a definition of learning, just to note a few of the features that implicitly guide the recognition of certain clear-cut cases. In an earlier paper, we suggested three (Margolis and Laurence 2011).[20] The first and most basic is that learning generally involves a cognitive change as a response to causal interactions with the environment. Of course, not all changes that trace back to an organism's environment will count as learning. That is the point of Fodor's examples where concepts are acquired through futuristic neurosurgery or through a miraculously fortuitous hit on the head. And not all cases of learning will involve environmental sensitivity, as some learning may be wholly a priori. Nevertheless, one important feature of learning is that it often, perhaps typically, involves a sensitivity to the environment.

The other two features we wish to suggest are ones that highlight aspects of the causal interactions that occur in paradigmatic cases of learning. One is that learning often involves a cognitive mechanism that isn't just altered by the environment but that, in some sense, has the function to respond as it does. For example, learning facts about the locations of various objects when entering a room isn't just a matter of having your mind altered upon perceiving the situation. The changes are of the sort that our perceptual systems and related belief-fixation mechanisms are designed to subserve. In contrast, when you get hit on the head, as in Fodor's example, your mind might be miraculously altered in a useful way,

[20] See Weiskopf (2008) for a similar proposal.

but the intervening mechanisms don't have the function of subserving these sorts of changes; it's just a matter of blind luck.

Finally, learning processes are ones that connect the content of an experience with the content of what is learned. The two aren't merely causally related. They are *semantically* related. Hypothesis testing exemplifies one type of semantic relatedness, but it hardly exhausts the possibilities. For example, the rote learning of a list of numbers may involve reciting the numbers several times out loud and chunking the numbers in thought. By any reasonable standard, the processes that are integral to these activities are ones in which the outcome is semantically related to the preceding experience. It's not as if the list enters into memory through sheer coincidence. The cognitive processes that bring about the change in thought are ones that undoubtedly turn on the contents of the mental states involved in the transition.

With this brief characterization of learning in hand, we can return to the claim that paradigmatic instances of acquisition involving the syndrome-based sustaining mechanism model are worthy of being described as learning. The model clearly has all three of the features that we have identified. In these sorts of paradigmatic cases, the learner gathers information about the kind based on her perception of the members of the kind that she encounters, so there is no question that the change is grounded in relevant causal interactions with the environment. Also, the information that is presented in experience isn't capable of directly creating a syndrome-based sustaining mechanism all by itself. The gathering of this information is guided by expectations about natural kinds in general and by expectations that derive from previous experiences with members of the kind in question and with members of similar kinds. It is filtered and processed by cognitive operations that take perceptual information and use it to control the application of a new concept, including its application in the context of similar future experiences. Here we have about as clear a case of semantically relevant processes as one could want. Finally, it's perfectly reasonable to suppose that the cognitive mechanisms and dispositions that support the formation of syndrome-based sustaining mechanisms have the function of building them as they do. It's even plausible that the essentialist disposition is owing to a biological adaptation for interacting with natural kinds, or a suite of adaptations for different types of natural kinds (animals, plants, natural substances, etc.). In short, our model of concept learning manifestly exemplifies the pretheoretic understanding of learning, satisfying the characteristics of typical instances of learning extremely well.

We conclude that the case that primitive concepts can be learned is very strong and that Fodor's responses to our model don't raise any substantive difficulties. None of Fodor's responses to our model are successful. First, though our model isn't wedded to an externalist theory of content, it wouldn't matter if it was, since coextensive concepts can be teased apart via their differing modes of presentation. Second, the existence of exotic cases involving causal flukes where acquiring

a sustaining mechanism doesn't suffice for learning a concept could not possibly show that there are no cases where acquiring a sustaining mechanism *does* suffice for learning a concept. More to the point, such exotic cases are simply irrelevant to showing that concepts cannot be learned in the sorts of cases of acquiring a sustaining mechanism that we based our argument on, which did not involve any type of causal fluke. Third, there is no need for us to provide a definition of learning in order to claim that our model is a learning model. It is enough if the model satisfies the features of paradigmatic cases of learning, which it clearly does.

Earlier we established that complex concepts can be learned (section 25.1). We can now add that primitive concepts can be learned too. Taken together these conclusions thoroughly undermine Fodor's scepticism about learning new concepts.

25.3 More Alternatives to the ABC Model

In the course of arguing against Fodor's anti-learning views, we have shown that primitive concepts can be learned. This is an important result not only because it helps to show why Fodor's argument against the possibility of concept learning fails, but because it speaks directly to the ABC model—which unlike his radical concept nativism, a great many theorists have followed Fodor in endorsing. Contrary to this tempting picture of the mind, it *is possible* to learn primitive concepts. Still, the previous section only mentioned one way of learning a new conceptual primitive—our syndrome-based sustaining model as it applies to natural kind concepts in the context of Fodor's asymmetric-dependence theory of content. In this section, we will briefly sketch some ways in which this account can be adapted to cover other types of concepts and other theories of content, and will mention a few other types of acquisition models that we consider to be complementary proposals.

Domain-specific conceptual templates. Our model for acquiring natural kind concepts worked on the assumption that while there may be a great deal of variation in the sustaining mechanisms for natural kinds, there is nonetheless a standard type of sustaining mechanism that has the function of supporting the acquisition of natural kind concepts in response to ordinary perceptual encounters. This standard type of sustaining mechanism can be thought of as involving a conceptual template that is filled in for each new kind when it is recognized as such. The template asks for information about the syndrome of the new kind and combines this with the essentialist disposition to create a sustaining mechanism that always has the same overall form. Earlier, we saw one of the big advantages of a theory positing this sort of conceptual template. It entailed that no particular natural kind concept has to be innate, while providing the basis for a learning model that is tuned to the contours of experienced natural kinds. Were the template to be

activated in Kenya, it would be well suited to acquire concepts for lions and gazelles. Were it to be activated in Australia, it would be just as capable of acquiring concepts for koalas and kangaroos.

We have seen that there is a good deal of evidence for specialized capacities directed to natural kinds (see Chapters 10, 13, 14, and 19; see also, e.g., Keil 1989; Gelman 2003; Atran and Medin 2008). But it is also reasonable to suppose that there may be comparable capacities for other types of conceptual domains and that these are associated with their own templates that underwrite a similar learning process—one based on the formation of a standard form of sustaining mechanism.

Another promising domain is the class of artefact concepts. Human life is deeply tied up with the manufacture and use of artefacts. Though there is some debate about how far back in the evolution of the species this goes, analyses of the archaeological record suggest that artefact use long predates the beginnings of *Homo sapiens* and that early humans possessed a rich and impressive range of artefact types (McBrearty and Brooks 2000). Undoubtedly, some of the facts that explain the success of human material culture have to do with the general capacity to faithfully receive and transmit cultural knowledge, allowing for incremental improvements from generation to generation (Basalla 1988; Richerson and Boyd 2005).[21] But it's likely as well that there have been adaptations that have given rise to domain-specific capacities for thinking about artefacts. This suggestion is not tantamount to Fodor's earlier, highly implausible view that humans are specifically endowed with innate concepts for each of the many types of artefacts that we surround ourselves with (e.g., the concepts SHOE, PENCIL, RADIO, CYCLOTRON, to name just a few). Rather, it is the far more modest and tenable suggestion that humans possess a few innate cognitive *templates* that contribute to rationalist learning mechanisms that explain the acquisition of artefact concepts.[22]

Though work on artefacts points in a number of directions (Margolis and Laurence 2007b), one general approach has received a good deal of empirical support and is likely to be at least part of the explanation of the cognitive basis of artefact concepts. This view claims that artefact concepts have an affinity with natural kind concepts. The connection isn't that artefact concepts activate the same essentialist disposition as natural kinds but rather that the categorization of artefacts, like natural kinds, involves an explanatory process and doesn't merely

[21] It should be emphasized that rationalists can and should accept the enormous importance of culture (and of gene-culture co-evolution) in accounts of human cognition and in explaining the origins of a multitude of concepts (Boyd and Richerson 1985; Richerson and Boyd 2005; Henrich 2016). We return to the way that rationalist theories account for the importance of culture in Chapter 26.

[22] As with natural kind concepts, it is plausible that there are a number of such systems, since there is great diversity in the types of artefacts that human beings create and use, including weapons, containers, clothing, ornaments, toys, shelters, tools for producing other tools, etc.

amount to ticking off a prescribed set of properties (Bloom 1996; Matan and Carey 2001; Barrett et al. 2008). According to this approach, something is deemed an artefact of a particular type because the best explanation of its salient observable properties is in terms of some deeper fact that accounts for these properties.

A prominent suggestion along these lines is that the deeper fact has to do with the intentions of the artefact's designer, particularly her intention that it serve a particular function. If this account (or one in the same general vicinity) is right, then a standard type of sustaining mechanism for artefacts may incorporate something analogous to the syndrome for natural kinds and combine this with the disposition to explain an item's perceived features in term of a designer's intent. Exposure to an artefact that exhibits the recorded features would activate the concept, while the activation would be modulated and perhaps retracted upon learning more about the intent of the designer who made the item (or upon learning that the item did not have a designer at all but was instead naturally occurring). Given this type of sustaining mechanism, artefact concepts could be learned in a way that is analogous to the way that natural kind concepts are acquired on our earlier account. When a learner encounters a new artefact, this would initiate a process that fills in the information specified by a domain-specific template and that associates it with the disposition linking artefact identity with the creator's intent. Which artefact concepts are acquired would then be a matter of the artefacts that the learner experiences, just as which animal concepts one acquires is a matter of which kinds of animal one encounters.

We are assuming, for purposes of argument, that artefact concepts are primitive concepts (i.e., that they aren't complex concepts).[23] Given this assumption, it seems that they could be learned by learning their sustaining mechanisms. The activation of domain-specific rationalist learning mechanisms for artefacts would guide the construction of new sustaining mechanisms (hence new concepts) in accordance with a stock of conceptual templates. The more templates, the more efficient the acquisition. What this goes to show is that there need be nothing special about natural kind concepts. Artefact concepts, and others that have their own conceptual templates, can be learned, thereby increasing the stock of primitive concepts.

Alternative theories of content. So far we have helped ourselves to the simplifying assumption that the content or meaning of a primitive concept is determined in accordance with Fodor's asymmetric-dependence theory. This has been a convenient assumption given that the theory handles primitive concepts so well. But it's important to see that the prospects for learning primitive concepts aren't tied to Fodor's asymmetric-dependence theory, since that would make our account

[23] Whether artefact concepts are primitive or not remains controversial. For arguments that they are, see Kornblith (2007); for an opposing view, see Thomasson (2007).

very precarious. The nature of mental content continues to be hotly disputed, and it would be an understatement to say that there is no consensus in support of the asymmetric-dependence theory (or in support of any other theory of content, for that matter). It would be useful then, if other approaches to content could be used as the backdrop for explaining how new primitive concepts can be learned. In fact, we think that all of the major theories of content, insofar as they cover primitive concepts, are encouraging in this regard.

Consider the family of theories that fall under the heading of the teleosemantics approach to content, which we briefly mentioned in Chapter 5 (e.g., Millikan 1984; Papineau 1984; Dretske 1988; see also Neander 2017, Shea 2018). Though there are important differences among the theories that are grouped under this heading, one of the foundational ideas of the teleosemantics approach is that a concept expresses the property that it has the function of detecting. For learned concepts, such functions might be acquired as mental representations are recruited by cognitive systems that depend upon the information they provide. Notice that there would still have to be sustaining mechanisms that accounted for how such representations are able to carry information about the properties they represent. So just as before, acquiring the concept is a matter of acquiring a sustaining mechanism, and we can begin to explain how concepts are learned by positing cognitive systems that underlie the construction of standard types of sustaining mechanisms.

In fact, in this case, much the same story that we told for natural kinds and artefacts within the framework of Fodor's theory of content can be imported with few modifications. This shouldn't be surprising since both approaches to content are centred around the occurrence of reliable mind-world causal dependencies. The main difference is that, on teleosemantics approaches, there is the added constraint that these are grounded in systems with the function of exploiting these informational states, where the function of a system is itself a matter of a history of selection.

Another prominent (and promising) theory of content is the causal-historical theory (Kripke 1972/1980; Putnam 1973). Though the theory was originally developed as a theory of reference for expressions in natural language (particularly for proper names and natural kind terms), it can be extended to the corresponding types of concepts (Burge 1979). The causal-historical theory is also highly relevant given our aims, because name concepts are widely thought to be primitive concepts (for reasons that trace back to Kripke's critique of the description theory of reference), and because there is little to be said for the idea that individual name concepts are innate.[24]

[24] According to the description theory of reference, a name refers to an individual through its association with a represented description which is uniquely true of that individual (or, on other versions of the theory, a description that expresses a cluster of properties which is mostly true of that individual). For example, the name "Aristotle" might be associated with the description "student of Plato and

Name concepts are rarely thought about in this context but are a useful case to consider in calling the ABC model into question. As with natural kind concepts and artefact concepts, we will focus on a single central way in which these concepts might be acquired. On the causal-historical approach, in the simplest case (where the name is being introduced for the first time), learning a new name concept begins with the perception of a new salient individual. Under the direction of the mechanism for acquiring name concepts, this perception causes the production of a new mental representation with a functional role characteristic of names.[25] This new representation would facilitate the storage of notable information about the encountered individual, and would become linked to object tracking processes that aim to reidentify it both in the short term and upon subsequent encounters. Now according to the causal-historical theory, a name concept's reference is the individual that is the causal source that led to the concept's initial production. This means that acquiring a new name concept may involve the activation of a conceptual template of sorts, but the significant difference between any two completions of the template has less to do with the newly recorded information than the identity of the individual that the concept causally leads back to. One way of picturing the situation is that the name template generates a different "mental file" for each name concept, but the referent of the name concept isn't determined by which individual the descriptions in the file are true of, but rather by the individual that stands in the appropriate causal-historical relation to the name concept.[26]

Suppose this general approach to name concepts is basically correct. Then we can say that specific name concepts aren't innate; they are acquired by systems that have the function of creating new name concepts upon exposure to salient new individuals. Moreover, because of the considerable amount of cognitive processing that goes into the creation of these new name concepts and because this processing is meant to reflect the details of experience, there is every reason to say that the resulting concepts are learned.

teacher of Alexander", in which case, it would refer to the man Aristotle, as he is the unique individual who satisfies this description. See Kripke (1972/1980) for a trenchant critique of this general approach.

[25] The functional role would involve, among other things, having these representations combine with predicates to form representations that express whole propositions, with other name representations to form conjunctions, and so on.

[26] Moving beyond the simplest case, other name concepts will often be learned on the basis of *reference borrowing*. In reference borrowing, the causal relation to the individual the name concept refers to is mediated by a complex causal chain that encompasses other individuals and their relations to further other individuals, and so on, and eventually to the referent. For example, someone might hear the name "Aristotle" being used, and this causes a new name concept to be created in their mind which is embedded in a psychological organization in which the referent is taken to be the individual that other speakers who use this name mean it to refer to, where other speakers defer to previous speakers, and so on, in a complex causal chain that stretches all the way back to Aristotle. The key point here is that the psychological states that are associated in the learner's mind with her new name concept aren't anything like a definition that describes the referent. Rather, they function as part of a different kind of sustaining mechanism that mediates the causal relation between the concept and its referent, making it a primitive concept.

Up to this point, the theories of content we have discussed have been externalist theories. But prospects for learning new primitive concepts are just as good for internalist theories. Consider a version of the conceptual role semantics approach to content (briefly discussed in Chapter 6) according to which content is not constituted by mind-world causal relations but rather by a concept's computational role in cognitive and perceptual processes.[27] In this case, it's no longer possible to rely on processes whereby new sustaining mechanisms are established. Nonetheless, it is possible to develop an internalist model that closely mimics this type of account. This would involve the same types of learning mechanisms centred around a template for acquiring new animal concepts. In this case, however, filling in the information in the template would allow one to learn a new primitive concept for a type of animal by creating a representation that, in virtue of this learned information, has the internal conceptual role in thought and perception that is constitutive of possessing this animal concept according to the internalist theory. For example, for someone who doesn't yet have the concept ZEBRA, encounters with a zebra would register the presence of a new animal and would initiate a process that produces a new symbol with a conceptual role appropriate to animals, filling in the specific details that this role requires by noting the relevant properties of the experienced animal.

Beyond templates. We have been suggesting that a good strategy for explaining how a primitive concept can be learned is to posit innate domain-specific mechanisms that specify one or more conceptual templates. A template guides learning by focusing the learner on the information it requires for concepts in its domain. This results in concepts that are very much a product of experience—concepts that embody both a record of the entities that have been encountered and a synopsis of the relevant properties these entities are perceived to possess. But must the learning of a primitive concept always proceed in this way? Certainly not. For one thing, much the same story could be told for templates that are not innate, but rather are themselves learned. Ultimately, there would have to be innate constraints on what types of templates could be learned, but the very possibility of learning a new template introduces an important dimension of flexibility. There are also interesting possibilities that make use of characteristically rationalist learning mechanisms pertaining to different domains in ways that combine them or have them interact with peculiarities of general cognition, thereby creating new concepts that don't quite fit into any single domain. A good example of this

[27] For a general overview of the conceptual role semantics approach, see Block (1986). However, we should note that Block himself defends a two-factor theory, which incorporates a causal-externalist component to content and hence doesn't represent the fully internalist version of the theory that we mean to consider in the text.

kind can be found in Pascal Boyer's theory of how certain religious concepts are formed (Boyer 2001).

Boyer proposes that religious concepts draw upon a small number of rationalist learning mechanisms based on intuitive theories (e.g., folk psychology, folk biology, and folk physics), and that new religious concepts arise when unexpected deviations from these theories are considered. The fact that the deviations are limited to just a few significant changes allows for the inferential structure inherited from the intuitive theories associated with rationalist learning mechanisms to remain largely intact. For example, if one were to hear about trees that talk, this would allow access to most of the inferences regarding trees (they still grow, need water, etc.). Yet the fact that the deviations are unexpected makes them memorable and hence likely to become established in the mind and to be conveyed to others who in turn are likely to remember them and pass them along to still others.

If Boyer is right, religious concepts depend upon experience. It's because members of a community hear about the curious talking trees that they develop concepts of these supernatural beings. But the cognitive mechanisms that mediate this learning aren't ones that simply fill in the details of, say, a *supernatural being* template. Rather, a considerable amount of innate differentiated structure comes to interact with quirks about human memory and with other general cognitive resources when supplied with what happens to be the right input. There is no guarantee that any old description of a supernatural being will take hold. If it isn't memorable, or deviates too much from the intuitive theories that are engaged by the acquisition process, then the corresponding concept will have little chance. But for the concepts that are competitive, they are readily learned because the mind has an organization that is receptive to them. In the typical case, the experience leading to such a concept's being learned involves hearing a story about the supernatural entities. We may assume that this would activate a new representational vehicle—an initially uninterpreted representational structure—that would gain its particular content through being assigned an appropriate conceptual role, that is, a role that captures the most salient details in the story, while also drawing upon information contained in the relevant intuitive theories. The new representation would then serve as a focal point for elaborations on this conceptual role as further stories and embellishments are traded in the community.

Though Boyer's treatment of religious concepts stretches the conceptual system beyond concepts that are the direct product of an innate template, there is still a residual trace of the template approach in his theory in that the concepts it covers retain much of the inferential structure associated with the intuitive theories they draw upon. Does this mean that primitive concepts can only be learned if a template of sorts is at the root of the process? Not at all. In Chapter 5, we sketched another type of model that is capable of explaining how certain new primitive concepts can be learned—our neo-Quinean model of abstraction. Recall that the model had three components: fine-grained general

representations, a similarity space for organizing these representations, and a selection process to isolate regions within the similarity space. There was a question about whether the output of this process is a primitive or a complex representation, but we argued that the natural way of looking at the matter is that abstraction produces new primitive concepts (see section 5.5). If this is right, then new primitive concepts (e.g., concepts for colours, shapes, and other similar categories) can be acquired without a conceptual template.

There are other approaches that don't rely on templates that also accommodate the learning of new primitive concepts. A particularly interesting one is Susan Carey's theory of conceptual bootstrapping (Carey 2009). The main idea behind this theory is that new primitive concepts can be learned by first learning a system of external symbols and the rules for manipulating them. The inferential patterns that this system embodies needn't be understood in the early stages of the learning. What's important is that the patterns in the uninterpreted (or partially interpreted) symbol system are later mapped to systems of representation that are meaningful for the agent and that a process of analogical reasoning relates the two in ways that result in the external symbol system taking on a whole new significance for the learner.

For example, as we saw in Chapter 2, Carey suggests that concepts for the positive integers are acquired by initially learning by rote both the counting sequence ("one, two, three…") and the counting procedure (in which each item is labelled by one and only one count term, using the count terms in a fixed order). At this point, counting for the child may have no more meaning than the sequence of handclaps in a game of patty-cake. But Carey claims that learning the positive integers takes place when children recognize an analogy between the counting procedure and operations that occur in a system of representation by which children represent and compare small sets of objects. If Carey is right, similar feats of bootstrapping account for how we learn the rational numbers, how we learn to differentiate the concepts of weight and density, and how we learn many of the uniquely human concepts that single us out as a species.

Much more could be said regarding these and other alternatives to the ABC model. However, we think that it should be clear enough by now that there are a variety of different ways in which primitive concepts might be learned. The rejection of the ABC model (and of Fodor's claim that primitive concepts cannot be learned) doesn't turn on the specific features of any particular model of concept acquisition, or of any particular conceptual domain, or any particular theory of content.[28]

[28] There is a further question about what type of limits there might be for the learning of new primitive concepts and representations. Could any kind of primitive representation be learned? Could *all* primitive representations be learned? Perhaps unsurprisingly, it turns out that this question

25.4 Conclusion

Though our concept nativism and Fodor's radical concept nativism are both rationalist accounts of conceptual development, in some ways they couldn't be more different. Learning is central to our account of where concepts come from, whereas radical concept nativism maintains that lexical concepts are simply innate. While Fodor's views of concept acquisition have evolved over the years, his opposition to the possibility of learning primitive concepts never wavers. In *LOT2*, he goes even further, arguing that the very idea of concept learning is simply incoherent. However, we have argued in this chapter that each strand in *LOT2*'s case against learning is mistaken. Learning doesn't require hypothesis testing—there are perfectly viable alternatives. But even if learning did require hypothesis testing, concepts could still be learned, since learning concepts via hypothesis testing is not in fact impossible. Contrary to Fodor and to the many advocates of the ABC model of concept acquisition, even new primitive concepts can be learned (and in a variety of different ways), thereby increasing the expressive power of the conceptual system.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0025

isn't a simple, straightforward one. One thing that is clear is that it is unlikely that all primitive representations can be learned. To get the process of learning going at all, *some* representations will be required to provide essential input to the learning process. The issue here is related to the ones that we highlighted earlier in Chapter 5 in relation to Locke's view that all general representations are learned via abstraction. Likewise, the considerations in Chapter 12 having to do with the questions of how and why learners would even entertain certain primitive representations argue that at least some primitive representations must be innate or acquired via rationalist learning mechanisms. At the same time, however, it should be noted that once certain primitive concepts and representations are in place, others will be learnable.

26
Fodor's Biological Account of Concept Acquisition—and the Importance of Cultural Learning

The previous chapter showed that Fodor's argument against concept learning doesn't work. Even primitive concepts can be learned. Still, there is one part of Fodor's most recent treatment of these issues that we have, to this point, left to the side. This is his claim in *LOT2* that concept acquisition might be the product of predominantly non-representational neurological processes. This biological account of concept acquisition—which aims to show how concepts are acquired without being either innate or learned—is independent of Fodor's argument against concept learning, and so requires separate consideration. In this chapter, we will begin by arguing that this alternative approach to concept acquisition is not viable and that the difficulties it faces serve to further highlight the central role of learning to any reasonable account of concept acquisition. We then step back to consider how our own view compares with Fodor's. It turns out that we are committed to there being a rationalist account of far more concepts than we have suggested so far. To some, this may make our account seem as implausible as Fodor's. But we will see that not only does this aspect of our concept nativism not burden our view with the dire consequences Fodor's view faces, it actually makes our view more reasonable and better able to explain the enormously important role of culture in shaping how concepts are acquired and used.

26.1 Against Fodor's Biological Account of Concept Acquisition

For many years, Fodor took his argument against concept learning to show that most lexical concepts are innate. We noted in Chapter 24, however, that there was a shift in his thinking about this matter starting with the publication of *Concepts* (1998). Although he continued to maintain that lexical concepts aren't learned, he became hesitant to conclude that any concepts are innate and proposed that we need to circumvent the learned/innate dichotomy. Fodor's initial idea for how to do this was to propose that concept acquisition isn't explained at the psychological level—the level of intentional states and processes—but at the neurological level. On this view, the acquisition of primitive concepts and representations is a

wholly non-cognitive affair that is to be explained directly and entirely in neurological terms. As we noted earlier, it is important to recognize that the claim here isn't simply the uncontroversial view that primitive concepts and representations are acquired via psychological processes that are realized in the brain and ultimately dependent upon the brain's activities. Rather, Fodor is making the far stronger claim that psychological processes are irrelevant to the acquisition of such concepts and representations—that their acquisition is explainable *only* in terms of non-psychological neurological processes.[1]

There is a natural worry for this type of approach to concept acquisition that Fodor himself was quick to identify. A generalized non-psychological approach to concept acquisition makes a mystery of the fact that concepts are regularly acquired through exposure to their instances. There seems to be no reason why a non-rational, non-psychological neurological process should require exposure to doorknobs to acquire the concept DOORKNOB, for example, or why such a process would lead to the production of the concept DOORKNOB on exposure to doorknobs (as opposed to GIRAFFE or some other unrelated concept). The solution can't be that doorknobs offer opportunities for confirming hypotheses about DOORKNOB or for representing the salient features of doorknobs. That would be to resort to an explanation in terms of cognitive processes and intentional states, which are explicitly forbidden on such an account. Fodor refers to this problem as the *doorknob/*DOORKNOB *problem*.

In *Concepts*, Fodor's solution to this problem took the form of a bold metaphysical theory. The reason someone acquires DOORKNOB when interacting with doorknobs is not because of their psychology, but rather because of the metaphysical nature of the property of being a *doorknob* (i.e., the referent of DOORKNOB). According to Fodor's metaphysical theory, the property of being a doorknob is partly constituted by the fact that it leads to the acquisition of DOORKNOB. Since it is in the nature of what it is to be a doorknob that, under certain conditions, we react to doorknobs by thinking DOORKNOB, then, Fodor claimed, it should no longer be mysterious that the concept DOORKNOB is normally acquired through interactions with doorknobs.

Fodor's *LOT2* theory of concept acquisition retains the core elements of this account. There is still the insistence that we should deny both learning and innateness and hold out for a generalized non-psychological theory. And *LOT2* also continues to endorse Fodor's metaphysical solution to the

[1] As we noted in Chapter 24, on our view some concepts and representations are acquired through wholly non-psychological processes as well—namely, innate concepts and representations in the acquisition base. So we have no objection to the origins of *some* concepts and representations being explained in terms of non-psychological neurological processes. The problem with Fodor's account, as we will see, is that it is meant to provide a general account that explains the origins of *all* primitive concepts and representations in terms of non-psychological neurological processes. Given our understanding of the innateness of a trait in terms of it being part of the acquisition base, such an account effectively takes all primitive concepts and representations to be innate, and so is not meaningfully different than radical concept nativism.

doorknob/DOORKNOB problem. What is distinctive about the account in *LOT2* is that it takes concept acquisition to proceed in two stages. During the first stage, a stereotype for a concept is learned. (A stereotype, for Fodor, is a representation that encodes statistical properties of features of experiences that are associated with the category picked out by a concept; for purposes of exposition in this chapter, we will follow his usage.) Importantly, the stereotype isn't identical with the concept. However, the fact that stereotypes are learned in acquiring concepts partly explains why concepts appear to be learned when, according to Fodor, they aren't. The appearance is owing to the fact that a stereotype is related to its concept, and while the concept itself isn't learned, the stereotype is. The second stage of concept acquisition on Fodor's new model occurs once the stereotype is in place. As Fodor sees it, a neurological process takes over and generates the concept from its stereotype. Crucially, the second stage is non-psychological. It is "*sub*intentional and *sub*computational…a kind of thing that our sort of brain just does" (Fodor 2008, p. 152). The second stage in Fodor's account is supposed to be where most of the action is and is why Fodor maintains that concepts aren't learned. It's because of our biological makeup that we arrive at the right concept corresponding to a stereotype; learning has nothing to do with it.

Fodor summarizes his *LOT2* account of concept acquisition as follows: "here's my story about concept acquisition: What's learned (not just acquired) are stereotypes (statistical representations of experience)" (Fodor 2008, p. 162). Once you learn a stereotype, then non-psychological, non-rational neurological processes get you to the concept: "in particular, it's a brute fact about the kind of animals we are (presumably about the kind of brains we have); and it's the bedrock on which the phenomenon of concept acquisition rests" (p. 161).

This account of concept acquisition faces a number of serious problems. Let's begin by looking at the first stage in the acquisition process—the stage where stereotypes (but not their concepts) are learned. A major difficulty with this part of his account is that it is inconsistent with Fodor's stance on learning. Our objection can be put as a dilemma. The first horn springs from Fodor's insistence that learning requires hypothesis testing. Assuming that it does, learning a concept's stereotype would necessitate putting forward hypotheses about the stereotype's individuating conditions and testing these against relevant observations. But then, following Fodor's own logic, putting forward the correct hypothesis would require having the stereotype prior to learning it, which would entail that the stereotype can't really be learned after all. Now it's true that we have rejected Fodor's circularity argument, but the present criticism concerns the internal coherence of Fodor's position. Perhaps Fodor could avoid the charge of circularity if he were to allow that not all learning reduces to hypothesis testing and were to claim that stereotypes are learned in some other unspecified way. But then he would face the second horn of the dilemma: this qualification would undermine the circularity argument against *concept* learning. Were Fodor to agree that learning doesn't necessitate hypothesis testing, there is

nothing to stop his critics from countering with the proposal that concepts are learned in this other unspecified way too.

In short, if Fodor's argument against concept learning is sound, then it undermines his biological account of concept acquisition, and, by the same token, if his biological account of concept acquisition is viable, it undermines his argument against concept learning.

But suppose we put aside Fodor's circularity argument and consider his biological account of concept acquisition simply on its own terms. Even then the account doesn't fare well. This is because of the difficulties of maintaining that concept acquisition is principally a non-psychological process. Fodor himself anticipates this challenge somewhat in the form of his doorknob/DOORKNOB problem, but he fails to appreciate the scope of the problem. For example, consider the fact that different people will often acquire different concepts on exposure to the same physical environment because they have varying interests in the same segment of the world. Dog enthusiasts have concepts corresponding to dozens of breeds, while people who are indifferent to dogs often have just a handful. Similarly, surfers have concepts corresponding to numerous types of waves and surfing conditions,[2] while most non-surfers have only a few generic concepts like CHOPPY WATER and SMALL WAVES. What's more, these examples aren't solely about the import of varying interests. They illustrate the significance of the surrounding culture. Dog enthusiasts and surfers hang out with others like them and pick up on the cultural norms of the groups they identify with. These interactions can be just as important as the interactions with the aspects of the physical world that the group cares about. But how can a purely biological process account for this fact? If mutable socially propagated norms are what matters, we need a mechanism that is calibrated to the *social* world. It's hard to see how this could be anything but a psychological mechanism, one that is chockfull of intentional states and processes.

As it happens, many concepts reflect the surrounding culture and it matters a great deal which culture it is that a learner grows up in. Medieval Europeans conceptualized health and disease in terms of humours—bodily substances that, according to the theories of the time, need to be kept in balance with one another. Few contemporary Europeans have these concepts, though they have as much exposure to instances of good and bad health as their historical counterparts. Or take the Newtonian concept GRAVITY. People who haven't been exposed to Newtonian physics aren't in a position to acquire this concept even though they experience the same sorts of causal interactions that exemplify gravitational influence (falling rocks, tides, etc.). Likewise for the logical concept VALIDITY, which is instantiated by all the valid arguments one is exposed to, but typically is acquired only in a university course in logic. Cultural forces are even more significant in cases where there is *no* physical manifestation of the items the

[2] See, e.g., Wikipedia's surfing glossary: https://en.wikipedia.org/wiki/Glossary_of_surfing

concepts are supposed to refer to. The Roman Catholic concept PURGATORY, for example, is of a place, or state of being, that no living person has actually experienced. That concept comes from cultural products—books, stories, sermons, etc.—that can only have their influence through psychological processes that extract their meaning.

One way that Fodor might try to mitigate the impact of cases like these is by claiming that the reason people in different cultures end up with different concepts is that they learn different stereotypes. This response emphasizes the fact that the second stage of concept acquisition—the crucial biological stage—can't occur until the right stereotype is in place. While this response might help with some of the cases, it faces three serious problems.

First, even when agents have access to the *same* stereotype, the surrounding culture can have a profound effect on how the world comes to be conceptualized. Colour concepts are a well-known case in point. Though people all around the world are equipped with essentially the same sensory systems, as we saw in Chapter 5, there is nonetheless a significant amount of variation in the basic colour concepts found in different cultures (Davidoff et al. 1999). The variation doesn't trace back to differences in surface reflectances or to other physical properties that are present in the environment. Rather, it's largely a matter of how different cultures have come to establish and encode the boundaries of their categories. Children inherit the local way of drawing these boundaries during the course of learning their language.

What can Fodor say about this? His biological account of concept acquisition stipulates that once the stereotype is in place, biology takes over and delivers the concept. But with colour concepts, there aren't different stereotypes in different cultures. As discussed in Chapter 5, focal instances of basic colour concepts are highly similar across cultures; it's the breadth and boundaries of the concepts that differ. So there is nothing in Fodor's account to explain why children in different cultures end up with concepts that match their own community's way of doing things. If anything, Fodor's account predicts that children across the globe should end up with exactly the same colour concepts since the same biological principles would be activated by the same stereotypes—a prediction that doesn't stand up. Much the same point can be made in light of the fact that our stereotypes for concepts are often highly impoverished and thus are unlikely to differ for related but distinct concepts. For example, for many people the stereotypes for GERBIL and HAMSTER are essentially the same. But this needn't stop them from acquiring one of these concepts but not the other (or even from acquiring both) on the basis of the very same stereotype. Fodor's view does nothing to relieve the mystery surrounding such differential concept acquisition.

The second problem with trying to use stereotype learning to explain away the appearance of psychologically mediated concept acquisition is perhaps even more damaging. Many concepts can be acquired in the absence of any stereotype at all.

In fact, Fodor himself has argued that complex concepts typically don't have stereotypes (Fodor 1981, 1998). But clearly, if a concept doesn't have a stereotype, then variable concept acquisition can't be explained by stereotypes. In particular, for all concepts that lack stereotypes, we are left with just the non-psychological biological part of Fodor's story about concept acquisition. And this part of Fodor's story has nothing to say about the evidence suggesting that concept acquisition is psychologically mediated. Consider, for example, the concept AN INVESTMENT FUND OPEN TO A LIMITED RANGE OF INVESTORS THAT UNDERTAKES A WIDER RANGE OF INVESTMENT AND TRADING ACTIVITIES THAN LONG-ONLY INVESTMENT FUNDS, AND THAT, IN GENERAL, PAYS A PERFORMANCE FEE TO ITS INVESTMENT MANAGER.[3] Prior to learning this concept (e.g., through verbal instruction), one is highly unlikely to have a stereotype associated with it. The same will be true of endlessly many complex concepts (e.g., the Scottish Country Dance concept in section 25.1). It is also likely to be true of lexical concepts with highly theoretical content, at least for many non-experts—MOLECULE, ARGON, ACETYLCHOLINE. These are not concepts that we learn by first learning a stereotype for them.[4]

This brings us to the third problem with the attempt to explain away psychologically mediated concept acquisition through stereotype learning. Many concepts are learned via the operation of psychological processes that go beyond stereotype formation. For example, some are learned alongside theories that they are embedded in (e.g., GRAVITY), and whether they are learned turns not on whether a stereotype for them is learned but on whether the learner is exposed to the relevant theory. In some cases, such as with natural kind concepts, there is arguably a default system devoted to gathering particular types of information about the kind and processing it in accordance with domain-specific inferential patterns (along the lines of the model outlined in Chapter 25). We can even predict which kinds of concepts an agent is likely to form based on considerations about the representational processes underlying concept acquisition in that domain. This makes sense if the concepts are acquired on the basis of the representational processes that support these predictions, but it is nothing short of a mystery on Fodor's biological account.

Consider once again Boyer's analysis of the origins of concepts of supernatural beings, which we discussed in Chapter 25. Boyer notes that the full range of possible supernatural concepts is far larger and more varied than what is actually found across cultures and consequently that supernatural concepts can't simply be a matter of generating new representations for strange incidents. As we saw in the previous chapter, Boyer's theory predicts, instead, that the most intuitive

[3] https://en.wikipedia.org/wiki/Hedge_fund

[4] There is room for debate about whether concepts like MOLECULE or the concept corresponding to the dance have stereotypes (Fodor 1981; Prinz 2002). For the present point, all that really matters is that these stereotypes, if they exist at all, are so anaemic that they cannot do the work that Fodor requires of them.

supernatural concepts are rooted in innate systems of inference (e.g., folk biology) and depend upon isolated counterintuitive deviations from the normal case (e.g., trees that talk), which make them memorable and apt for cultural transmission. This predication fits well with the anthropological record and with experiments that have been designed to test just these sorts of effects on memory (Boyer 2001). But Fodor's theory has no resources to explain patterns of this sort.

In sum, Fodor's positive theory of concept acquisition faces a number of serious challenges. First, there is a dilemma stemming from stage 1 of the acquisition process, which involves stereotype learning. Either his argument against concept learning undermines this aspect of his positive account or else his account of stereotype learning undermines his argument against concept learning. His account also fails to adequately explain away the robust and diverse evidence for maintaining that concept acquisition is a psychological-level phenomenon. This evidence includes the doorknob/DOORKNOB problem, but the doorknob/DOORKNOB problem is just the tip of the iceberg. More important is the fact that concept acquisition depends profoundly on one's *cultural* environment, as mediated by one's psychology. Both the number of concepts one acquires about a given domain (e.g., dogs, waves) and which specific concepts one acquires (e.g., which specific colour concepts), depend on one's cultural environment. And the dependence is often very deep, as illustrated by the many concepts which are largely cultural, with little or no grounding in one's immediate environment (e.g., PURGATORY). Moreover, the variation cannot be explained by stereotype differences given that the conceptual variation is possible without variation in stereotypes—many concepts either lack stereotypes altogether or else fail to have sufficiently robust stereotypes to discriminate between related concepts.

In stark contrast to Fodor's biological view, accounts of concept acquisition that embrace learning do not face any of these problems. The stereotype-learning dilemma disappears since concept learning accounts don't deny that either stereotypes or concepts can be learned. The doorknob/DOORKNOB problem has a straightforward and satisfying solution: concepts are often acquired through encounters with their instances because they are learned at least partly on the basis of collecting information about their instances (Laurence and Margolis 2002).[5] And the problems stemming from cultural embeddedness never arise, since concept learning accounts happily accept that many aspects of one's cultural surround are represented and feed into concept learning.

[5] In fact, calling this a solution is in some ways misleading; the problem doesn't really even arise for learning models, so it would be more accurate to say that it is avoided rather than solved. On a learning model, it's just obvious why causal interactions with doorknobs lead to the acquisition of doorknob and not, for example, giraffe—experiences with doorknobs are a great source of information about doorknobs but a terrible source of information about giraffes. This point underscores the fact that the doorknob/DOORKNOB problem really only arises in the first place because Fodor has rejected all cognitive-level stories about concept acquisition.

In retrospect, it should hardly be surprising that *LOT2*'s view of concept acquisition would face so many difficulties. Fodor's biological account in *LOT2* should have been suspect from the start since its anti-cognitivism flies in the face of the deepest motivations behind cognitive science—motivations that go back to its opposition to behaviourist accounts that likewise eschewed appeals to cognitive processes in the explanation of learning and behaviour. This becomes apparent if we consider the analogue of Fodor's theory of concept acquisition in the realm of language. A theory of language acquisition along the lines of Fodor's biological account of concept acquisition would hold that language is neither learned nor innate. The core processes involved in language acquisition, according to this sort of account, are non-rational, non-psychological neurological processes; language acquisition is simply "a brute fact about the kind of animals we are (presumably about the kind of brains we have)", and does not admit of a cognitive-level explanation. Moreover, the reason why people who are exposed to English acquire the ability to speak English (as opposed to Italian, Mandarin, or the ability to play the violin, for that matter) is because English sentences are instances of mind-dependent types; to be an English sentence just is to be the kind of thing that makes minds like ours jump to having the capacity to understand English. This account of language acquisition is clearly inadequate—precisely because it attempts to explain language development in wholly non-cognitive terms (just as *LOT2* attempts to explain conceptual development in such terms).

26.2 Is LIGHTBULB Innate?

It's time to take stock. For many, Fodor's views on the origins of concepts have been seen as providing strong grounds to reject concept nativism altogether. While we have wanted to distance ourselves from Fodor views, as we'll explain shortly, there is a way in which our own view may end up seeming almost as extreme as his and therefore just as hard to accept. Though we think that this worry is misplaced, we think that it is important to address it directly—and that doing so sheds further light on how our view fits into the larger landscape of the full space of options for accounts of the origins of concepts.

To begin, as should be clear, we take Fodor's views on the origins of concepts—both Fodor's radical concept nativism and his biological account of the origins of concepts—to be deeply problematic. On his radical concept nativism, only complex concepts have a chance of being learned, and unlearned concepts must be innate. Since he also thinks that all, or nearly all, lexical concepts are semantically primitive, this means that virtually all lexical concepts are innate. As he puts it in "The Present Status of the Innateness Controversy":

> There is "bachelor", which is supposed to mean "unmarried man"; there are the causative verbs, of which the analysis is vastly in dispute; there are jargon terms, which are explicitly and stipulatively defined…there are kinship terms, which

> are also in dispute; there is a handful of terms which belong to real, honest-to-God axiomatic systems ("triangle", "prime" as in "prime number"); and then there are the other half million or so lexical items that the OED lists. About these last apparently nothing much can be done. (Fodor 1981, p. 284)

This means that a concept like LIGHTBULB—along with at least a half million or so other lexical concepts—is innate. As we've seen in this chapter, Fodor's biological view isn't much better. LIGHTBULB might no longer be innate, but it still isn't learnable. According to this account, no one has ever learned the concept LIGHTBULB—Edison didn't learn it, and neither did you. There is just a biological fact about human beings that the concept LIGHTBULB materializes in our minds upon exposure to the lightbulb stereotype. While we have argued that there is much to learn from reflecting on Fodor's arguments for these views, the views themselves are completely implausible. So, it isn't hard to see why we have been at pains to distance ourselves from Fodor's views—for example, by calling his (1981) view *radical concept nativism* and excluding it from the class of views that we have been defending under the label *concept nativism*.

As we see it, we were right to emphasize the difference between our concept nativism and Fodor's in this way. But from a certain vantage point, one might wonder whether our own version of concept nativism is really all that different—or less extreme—than Fodor's. This is because, though we have not called attention to this fact, it turns out that *there is a rationalist account of the origins of virtually every lexical concept on our view too*. Although we don't hold that all or nearly all lexical concepts are innate as Fodor did, we do hold that there is a rationalist account of the origins of virtually every lexical concept—including LIGHTBULB. So one might wonder whether our view really is less radical than Fodor's. Have we really come all this way just to end up with a view as implausible as Fodor's?

The short answer is "no". We think that there are very clear and important differences between our view and Fodor's, and that when seen in the proper light, our view is neither extreme nor implausible. But getting clear about all of this will take a bit of unpacking.

First, why is it that, on our view, there is likely to be a rationalist account of the origins of LIGHTBULB and most other lexical concepts? In Parts II and III of the book, we argued for a rationalist account of the origins of concepts in a number of conceptual domains. But, though we didn't say so there, we in fact think that there will be some type of rationalist account for virtually every lexical concept in every conceptual domain. The reason for this is that while we only take a relatively small subset of these concepts to be innate, we think that virtually all of the rest will be innate or be acquired via rationalist learning mechanisms that trace back to some type of characteristically rationalist psychological structures in the acquisition base. This means that there will be a rationalist account of the origins of virtually every lexical concept on our view.

So how exactly is our view different from Fodor's view? How could it be neither extreme nor implausible to suppose that there is a rationalist account of the origins of virtually every lexical concept? The key is that not all rationalist accounts are equal. Rationalist accounts of the origins of concepts can take many forms and can be rationalist to very different extents despite all being rationalist.[6] The sense in which, for us, there is a rationalist account of the origins of virtually every lexical concept will be a relatively modest one in this context.

Consider, for example, what might be thought of as a *minimally rationalist account* of the origins of concepts. Such an account might involve just a few characteristically rationalist psychological structures which aren't closely aligned with the vast majority of learned concepts.[7] For example, suppose that the concept OBJECT is innate. Now imagine that OBJECT is incorporated into a domain-general learning mechanism for acquiring concepts for different types of objects in many different conceptual domains. Suppose that one of the many concepts this learning mechanism acquires is the concept ROCK (from the domain of natural objects). Then it would turn out that on this account there is a rationalist account of the origins of the concept ROCK. Why? Because even though it is acquired by a domain-general learning mechanism, that learning mechanism incorporates the innate concept OBJECT—a characteristically rationalist psychological structure—making the learning mechanism a *rationalist* domain-general learning mechanism. As we are imagining it, this same rationalist learning mechanism is capable of acquiring many other concepts in many different conceptual domains (BALL, BIRD, HAMMER, and so on). But since it traces back to a characteristically rationalist psychological structure that is part of the acquisition base, it counts as a rationalist learning mechanism all the same—even if, as rationalist learning mechanisms go, it is only rationalist to a relatively minimal extent.[8]

Other minimally rationalist accounts of the origins of concepts might turn on learning mechanisms that make use of other types of relatively minimal rationalist resources, for example, innate elementary logical concepts.[9] On a view in which

[6] As explained in Chapters 2 and 6, rationalist accounts vary along a number of independent dimensions, which determine in a coarse-grained way the extent to which an account is rationalist or empiricist (quantity, complexity, degree of articulation, diversity of content domains, abstractness, degree of domain specificity, and degree of alignment).

[7] Recall that *alignment* refers to the closeness of the relation between two content domains associated with a learning mechanism—the target domain (the domain that the learning mechanism is directed at) and the resource domain (the domain that the innate resource which the learning mechanism traces back to is directed at) (see Chapter 2).

[8] We have described this learning mechanism as incorporating OBJECT as an innate concept. But this is simply for ease of exposition; nothing turns on the supposition that this account incorporates an innate concept as opposed to some other innate domain-specific resource, such as an object-tracking system or a physical-reasoning system (see Chapter 15 for discussion of these systems).

[9] The term *minimally rationalist account* is not intended to pick out a single type of absolutely minimal account, but rather to refer to a range of accounts that are rationalist but to a substantially lesser extent than many other rationalist accounts. Moreover, since both rationalist and empiricist

elementary logical concepts are innate, any concept that is acquired via a learning mechanism that involves reasoning employing such logical concepts would count as being acquired via a rationalist learning mechanism. Much the same goes for concepts acquired in part via language (assuming a rationalist account of the origin of language), or concepts acquired in part via reasoning about mental states (assuming a rationalist account of the origins of the mental state concepts involved). In all these kinds of cases, the learning mechanism may be minimally rationalist, but it will still be rationalist. So it is not hard to see how, on an account like ours, it could turn out that there would be a rationalist story—of at least this minimal sort—for the acquisition of virtually every lexical concept. Indeed, given our arguments earlier in the book for a rationalist treatment of a broad range of concepts that play a role in the learning mechanisms that are used to acquire many other concepts, it's hard to think of examples of concepts where, on our account, there *wouldn't* be at least a minimally rationalist account of their origins.

It might be thought that this consequence (and the standard we are adopting for what makes an account rationalist) trivializes the notion of a rationalist learning mechanism—and thereby the very idea of concept nativism. But it doesn't for two reasons, and it is important to see why.

First, while minimally rationalist learning mechanisms are rationalist to only a relatively modest extent, they would be rejected by most empiricists. Consider, for example, the fairly typical empiricist view offered by Cohen et al. (2002):

> Infants are born neither with a blank slate nor a preponderance of innate core knowledge. Rather, we would argue that infants are born with a system that enables them to learn about their environment and develop a repertoire of knowledge…The system is designed to allow the young infant access to low-level information, such as orientation, sound, color, texture, and movement. (p. 1325)

Cohen and colleagues emphasize that "integration [into higher-level representations] is based upon statistical regularities or correlations in activity of those lower-level units" (p. 1326) and that "[t]his learning system applies throughout development and across domains" (p. 1327; italics removed from original). Clearly, they don't think it would be trivial to claim that the acquisition of LIGHTBULB depends on a learning mechanism which traces back to an innate object concept, a domain-specific rationalist learning mechanism for representing and

accounts are subject to various types of trade-offs among the different dimensions that determine the extent to which an account is rationalist or empiricist (see Chapter 2), there is little sense to the idea of a unique way of being minimally rationalist in any case. Even for views that are rationalist only in virtue of involving an innate domain-specific understanding of objects as such, or an innate domain-specific understanding of elementary logical concepts, there will be many different kinds of minimally rationalist accounts, which differ in terms of their degree of complexity, degree of articulation, and so on.

reasoning about mental states, or any other rationalist learning mechanism which traces back to characteristically rationalist psychological structures in the acquisition base, however modest their contribution may be. They reject all such accounts.

While an account like Cohen et al.'s is representative of empiricist accounts, it would be hard to overemphasize the difficulties that such an account faces in attempting to explain the origins of a concept of an object having a particular function or purpose or being designed with a particular intent, and ultimately to an understanding of all that is involved in possessing an artefact concept like LIGHTBULB. On such an account, the concept's acquisition (like that for all other concepts) would have to be based solely on statistical regularities regarding low-level perceptual properties (colour, texture, movement, etc.). On this basis, a learner might be able to develop a concept that *mimics* a concept like LIGHTBULB to some extent, drawing on sensorimotor representations that pick out some aspects of experiences with lightbulbs. However, such a concept would no more be the concept of a LIGHTBULB than a façade of a barn would be a barn. Pigeons can be trained to discriminate Monets from Picassos (Watanabe et al. 1995) or Stravinsky from Bach (Porter and Neuringer 1984), but no one thinks that they actually have the concepts IMPRESSIONIST PAINTING or CUBISM, since they are clearly responding to low-level sensory properties, not to an understanding of an artistic genre or a painting's place in art history. It is precisely for this reason that it is so important to control for the possibility of mere façade concepts of this sort being acquired instead of the real thing in studies with animals or with the sorts of deep learning models we discussed in Chapter 19.

The second reason why accounts of the origins of concepts that posit minimally rationalist learning mechanisms don't trivialize the notion of a rationalist learning mechanism is that this is not the only type of rationalist learning mechanism that concept nativists endorse, and certainly not the only type that we endorse. Consider LIGHTBULB again, but this time taking into account the kind of rationalist learning mechanism that we briefly sketched for artefact concepts in the previous chapter. On such an approach, the learning mechanism for LIGHTBULB wouldn't involve just the concept OBJECT but also such concepts as FUNCTION, PURPOSE, and INTENT—which, on our view, would be either innate or else themselves acquired via rationalist learning mechanisms that aren't merely minimally rationalist. Accounts like this, which involve rationalist learning mechanisms that are more complex, more articulated, and that trace back to innate resources that are more abstract and more closely aligned with the target domain, might be called *robustly rationalist accounts* of the origins of concepts.[10]

[10] Like the term *minimally rationalist account*, the term *robustly rationalist account* is not intended to pick out a single type of account, but rather to refer to a relatively broad class of accounts that vary in terms of the trade-offs they make concerning the different dimensions that determine the extent to

Robustly rationalist accounts—which were our focus throughout Parts II and III—are clearly very different from competing mainstream empiricist accounts.[11] At the same time, it should be clear that they are also miles away from Fodor's radical concept nativism. It's one thing to say that LIGHTBULB is acquired via a rationalist learning mechanism that involves concepts like OBJECT, FUNCTION, PURPOSE, and INTENT, whose own acquisition is explained in rationalist terms that are not merely minimally rationalist. It is quite another to say that LIGHTBULB and virtually every other lexical concept is simply innate (that is, a part of the acquisition base). Our rationalist account of the origin of LIGHTBULB and a myriad of culturally embedded lexical concepts couldn't be more different than both Fodor's radical concept nativist account and his biological account: Unlike both of the accounts that Fodor offers, on our account, LIGHTBULB is *learned*. But panning out from this one example and thinking about lexical concepts more generally, another difference is that, on our account, the vast majority of such concepts aren't just learned. They are learned in a way that is responsive to cultural factors. Culture plays a major role regarding both which concepts are acquired and how they are acquired.[12]

Boyd (2018) highlights the enormous importance of cultural learning with a discussion of the Burke-Wills expedition across Australia in 1860, an illustration of what he calls "lost European explorer experiments" in which healthy, educated, bands of European explorers get "stranded in an unfamiliar habitat in which an indigenous population is flourishing… [and] cannot figure out how to feed themselves, and… often die" (pp. 16–17). Burke and Wills, unable to find

which an account is rationalist or empiricist. A clear example of a robustly rationalist account would be one that postulates a learning mechanism that is closely aligned with the target domain of the innate resource it traces back to. In the case of LIGHTBULB, for example, a minimally rationalist account might say that it traces back to an innate resource that is directed at the domain of physical objects and hence one that isn't closely aligned with the domain of artefacts. By contrast, the sort of account we sketched earlier (in which particular artefact concepts are learned through a rationalist learning mechanism that traces back to an innate artefact template) is one in which the learning mechanism for LIGHTBULB and other artefact concepts is closely aligned with the domain of artefacts. The innate artefact template that the learning mechanism traces back to is directed at the domain of things with functions that are determined by a designer's intentions—a domain that is very close to, if not identical with, that of artefacts.

[11] Empiricist views vary in the extent to which they are empiricist, just as rationalist views differ in the extent to which they are rationalist. Cohen et al.'s account is a paradigmatic empiricist account. Accounts that are still empiricist but to a lesser extent (or, in other words, that are rationalist to a greater extent) might build in some further innate psychological structures, such as domain-specific biases to attend to certain types of sensory information. Accounts that are empiricist to an even lesser extent might build in a very limited amount of characteristically rationalist psychological structures, particularly when these are relatively simple, non-abstract, unarticulated, or not closely aligned with their target domains (as in Mandler's account, discussed in Chapter 21).

[12] The fact that the learning mechanism involved in acquiring a concept like LIGHTBULB traces back to an innate resource that embodies a conceptual template—one that highlights a designer's intentions—means that such an account embodies a high degree of alignment with the target domain of artefact concepts. Notably, however, this resource is not very closely aligned with any *particular* artefact concept (LIGHTBULB, GUITAR, MICROSCOPE, and so on).

sufficient food, were fortunate to encounter the Yandruwandha, a local aboriginal group, who helped them and gave them supplies—particularly some fish and a type of cake made with nardoo seeds (nardoo being a type of aquatic fern). But even with this assistance and an environment with ample sources of fish and nardoo, they didn't last long once left on their own.

As Boyd notes, "the men of the Burke-Wills expedition didn't die because they were stupid. They died because they didn't have access to the culturally transmitted knowledge that allowed the Yandruwandha to survive around Cooper's Creek" (p. 17). They didn't realize that nardoo contained toxins and didn't know the Yandruwandha's detoxification procedure. And, although there were fish in the nearby ponds, they apparently weren't able to catch any even after seeing the Yandruwandha fish with nets. As Boyd explains, this was probably because they didn't know how to create the right type of nets using the local materials. This would require knowing that cords could be made from twine from particular types of bark and roots.

By contrast, Boyd remarks,

> [f]or the Yandruwandha, Cooper's Creek was a land of plenty because they had a rich trove of culturally transmitted knowledge about how to make a living there. A Yandruwandha "Natural History Handbook" would have run to hundreds of pages with sections on the habits of game, efficient hunting techniques, how to find water, how to process toxic ferns, yams, and cycads, and so on. (p. 18)

This is only the tip of the iceberg of culturally transmitted knowledge, however:

> Australian Aborigines are famous among archaeologists for the simplicity of their technology. Nonetheless, an "Instruction Manual for Technology" would have had to cover the manufacture and proper use of nets, baskets, houses, boomerangs, fire drills, spears and spear-throwers, poisons, adhesives, shields, bark boats, ground stone tools, and much more…[moreover,] cooperation plays a crucial role in human subsistence. To become a competent Yandruwandha, you would also have needed to master "Social Policies and Procedures," "Grammar and Dictionary," and "Beliefs, Stories, and Songs," volumes of comparable length. (pp. 18–19)

The same considerations hold for any place where human beings flourish. Whenever this happens, it is partly because human learners have a psychology under which they can acquire concepts and associated information that is backed by an enormous amount of cultural knowledge.

Boyd's explanation is focused on cultural knowledge, but this knowledge is framed in terms of many culturally local *concepts* as well—concepts for particular plant and animal species, concepts for particular detoxification procedures,

concepts for particular types of raw materials, tools, weapons, rituals, norms, construction techniques, and so on. This vast array of concepts is richly informed by culturally transmitted information.

The kind of rationalist account of the origins of concepts that we endorse doesn't just acknowledge the fact that culture plays such an enormous role both in accounting for which concepts are acquired and how they are acquired; such an account is crucial to making sense of how cultural learning of this sort is even possible. Consider, for example, the important role of testimony in cultural learning. As we have discussed earlier (in Chapter 14), while our capacity for testimony supports learning across many different content domains, it is shaped in important ways that trace back to characteristically rationalist psychological structures in the acquisition base (Harris 2012; Kline 2015; Harris et al. 2018). And testimony obviously involves the use of natural language, which there are good grounds to suppose is itself acquired on the basis of a rationalist learning mechanism. Natural pedagogy, another rationalist learning mechanism that was discussed earlier (see Chapters 17 and 22), likewise supports learning across many different content domains and traces back to characteristically rationalist psychological structures in the acquisition base. The particular type of domain-general inference that it involves—inferring generalizable information about a kind—is linked to attention monitoring, ostensive communication, and referential intentions, all of which are connected with the expectations of teachers and learners and involve characteristically rationalist psychological structures (Csibra and Gergely 2009, 2011).

Other types of similar rationalist learning mechanisms involve content biases. These include mechanisms that help learners to navigate the complex problem of who to rely on for socially mediated knowledge. One example would be a mechanism that involves a *prestige bias*, which disposes learners to imitate successful models and to try to gain proximity to these models (Henrich and Gil-White 2001). Another is a *conformist bias*, which disposes learners to adopt or copy the behaviours, beliefs, norms, techniques, and practices of the majority (Henrich and Boyd 1998). Humans also have many reasoning heuristics, biases, and practices that arguably involve or are dependent on characteristically rationalist psychological structures in the acquisition base. Some of these pertain to particular content domains (e.g., values, risk), some are more general (Kahneman et al. 1982, 2000; Gilovich et al. 2002; Gigerenzer and Selten 2002; Kahneman 2011).[13]

Moreover, as we saw with LIGHTBULB, an empiricist domain-general learning account based on only low-level perceptual information is simply a non-starter for accounting for the vast array of culturally embedded concepts. Even minimally rationalist learning mechanisms are not enough. Cultural learning couldn't

[13] Some reasoning heuristics and biases may be learned, but many will not be. And many that are learned are in all likelihood learned on the basis of some type of rationalist learning mechanism.

get off the ground at all without a wide range of concepts that are either innate or else acquired via rationalist learning mechanisms that go substantially beyond minimally rationalist learning mechanisms. In previous chapters, we have engaged with local debates about the origins of many concepts in many particular domains, arguing that these concepts are either innate or else acquired via such rationalist learning mechanisms. These concepts trace back to a very rich acquisition base—one which includes a multitude of characteristically rationalist psychological structures. Since cultural learning, like all forms of learning, works with the psychological structures that the human mind has to offer, it only makes sense that cultural learning will make use of them. This means that cultural learning will draw on (among many others) concepts related to rudimentary logical and numerical content, space and time, causality, modality, agency, normativity, purpose, mental states, social groups and alliances, and the metaphysical categories that are inherent to common-sense thinking, like representations for objects, individuals, substances, events, and kinds. These concepts, which we have seen have a rationalist basis, will play central roles in structuring the input to cultural learning processes, and in guiding the learning processes themselves.

The import of such concepts is apparent when considering almost any collection of concepts informed by cultural learning—concepts such as DOLLAR (denominated in *numerical* units, backed by *social rules*, regarding what *can* be *exchanged* for goods and services), US CITIZEN (an *individual* whose *rights* are *contingent* on their being born in the country *or* naturalized), WEDDING (a *ceremonial event* conferring new *status* on the *participants* joined in marriage), PENALTY KICK (which gives a team—a type of *group*—an opportunity to gain *one* point when the opposing team violates a *rule*), GLUE (which is *manufactured* and *intended* to be an adhesive *substance*), and BREAKFAST (a meal that typically occurs in the morning—a certain *time of day*—when particular *foods* are *conventionally* eaten). These, and more or less every other concept where cultural learning plays a role, will be acquired via rationalist learning mechanisms that depend to varying degrees, and in varying ways, on a rich rationalist acquisition base of the sort that we have argued for throughout this book.

26.3 Conclusion

This chapter has taken a critical look at Fodor's biological account of the origins of concepts and has also stepped back to consider how our own view compares to Fodor's views. Given Fodor's claim that learning is impossible and his hope to avoid having to say that all concepts are innate, he has tried to find a third way, one that avoids the learned-or-innate dichotomy. In *LOT2*, this takes the form of a proposal in which concept acquisition is a biological process in which there is some non-psychological process in our brains that manages to produce a concept

upon encountering its stereotype. But as we have seen in this chapter, Fodor's biological account of concept acquisition has little to be said for it. Fodor's third option isn't viable. For all intents and purposes, concepts are either learned or innate, and any reasonable theory will hold that learning is absolutely essential to understanding the development of the human conceptual system.

How does our own account of the origins of concepts compare with Fodor's? While our account may initially seem extreme in light of the fact that virtually every lexical concept traces back to some characteristically rationalist psychological structures in the acquisition base, it is not at all radical in the end. The key to seeing why it isn't, and why our account is actually very reasonable, is the recognition that rationalist learning mechanisms are not all the same—they can vary dramatically in the number, type, complexity, degree of articulation, abstractness, and alignment of the characteristically rationalist psychological structures that they trace back to.[14] For concepts like LIGHTBULB, the characteristically rationalist psychological structures involved in their learning mechanisms are ones for which there is strong independent grounds to suppose that they are part of the acquisition base; Parts II and III provide such grounds for adopting a rationalist account of the origins of concepts like OBJECT, PURPOSE, and INTENTION. And once such concepts are in place, it only makes sense that cultural learning processes would build off of them in acquiring further concepts. Our view has emerged as more richly rationalist than Part II suggested. In addition to holding that there is a robustly rationalist account of the origins of many concepts across many conceptual domains, we also hold that there is at least a minimally rationalist account of the origin of virtually all concepts. But the very broad minimal rationalism here is only spelling out what was already implicit; it's just a consequence of the type of acquisition base we have already argued for. And as we have argued, our account—with its rich acquisition base—is not only compatible with cultural learning but should also be seen as playing a key role in explaining how culture shapes the concepts that we acquire and use.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0026

[14] Part of what may make it seem unreasonable, to first appearances, is simply the fact that the origin of a concept like LIGHTBULB is tracing back to something innate (so something in the acquisition base). But it is crucial to bear in mind that every account of the origins of concepts—rationalist or empiricist—will trace the origins of absolutely any concept at all (including LIGHTBULB) back to innate psychological structures in the acquisition base of one kind or another. If the concept is not itself innate in the sense of being psychologically primitive, it must be acquired on the basis of other psychological structures that are either innate in this sense or that trace back to further psychological structures that are innate in this sense. Seeing this helps to put into proper perspective the fact that a concept like LIGHTBULB traces back to innate rationalist psychological structures in the acquisition base.

27
Conclusion to Part IV

Can concepts be learned? One of the main claims in this book is that the answer to this question is a resounding "yes". When it comes to learning, the difference between rationalism and empiricism isn't about whether it happens, or even about how important learning is; it's about the character of these learning processes, or *how* learning works. While empiricists hold that concept acquisition is ultimately mediated by learning mechanisms that are almost wholly domain-general in character, essentially all rationalists embrace learning in accounting for the origins of concepts as well but hold that, in addition to domain-general learning mechanisms, rationalist learning mechanisms must also play a crucial role in concept acquisition. Some of the confusion surrounding this point about rationalism and learning stems from the fact that Fodor has often been seen to be *the* rationalist that empiricists must reckon with, and because Fodor has consistently rejected the very possibility that primitive concepts can be learned. But while Fodor's discussions of the origins of concepts have highlighted important theoretical issues and have rightly commanded a great deal of attention, we should not lose sight of the fact that Fodor's radical concept nativism and his rejection of the possibility of lexical concepts being learned are very extreme views that have never been widely endorsed by concept nativists. In this way, Fodor has always been an outlier even on the rationalist side of the rationalism-empiricism divide.

At the same time, much can be learned about concept acquisition by reflecting on Fodor's arguments against concept learning and the various alternatives to concept learning that he has proposed, culminating in his final discussion of these issues in *LOT2*. Part IV was essentially an extended reflection on Fodor's work on the origin of concepts, which yielded a number of important morals.

One of these is that the question of which concepts can be learned has nothing to do with whether they are complex or not. Fodor's idea that only complex concepts can be learned—and that they are learned when and only when they are appropriately assembled from a fixed stock of innate semantic primitives—has received widespread support in the cognitive science community, from rationalists and empiricists alike. Proponents of this *Acquisition by Composition model* (the ABC model) of concept acquisition broadly agree with Fodor that there is no other way for a concept to be learned, but they usually try to sidestep Fodor's conclusion that all lexical concepts are innate by also maintaining that typical lexical concepts are complex concepts. In other words, against Fodor they hold that lexical concepts can be assembled from more basic representations

(in accordance with the ABC model), but they nonetheless agree with him that concepts can be learned *only* when they are composed out of primitive concepts.

We have argued that the ABC model is right about the fact that complex concepts can be learned when composed as the ABC model says. As a result, Fodor's second thoughts in *LOT2*, questioning whether even complex concepts can be learned, are misplaced. Moreover, we have argued that not only can they be learned, but they can be learned via a process of hypothesis testing without any threat of circularity. The ABC model is right that complex concepts are learnable.

On the other hand, the ABC model is wrong that the *only* way that concepts can be learned is via composition and that only complex concepts can be learned. Primitive concepts, which fall outside of the scope of the ABC model because they lack combinatorial semantic structure, can be learned too. In Chapter 25, we initially illustrated how this is possible with reference to one particular theory of content (Fodor's asymmetric-dependence theory) and one class of concepts (natural kind concepts), but went on to show that this point generalizes to other types of concepts and other approaches to conceptual content. Thus, contrary to Fodor and to the many advocates of the ABC model, new concepts can be learned in a way that fundamentally expands the combinatorial expressive power of the conceptual system. While *LOT2* is right that the complexity of concepts has nothing to do with whether or not they are learnable, this isn't because no concepts are learnable. It's because *both* primitive and complex concepts are learnable.

In *LOT2*, Fodor didn't just reject the idea that concepts can be learned. He also proposed a generalized biological model of concept acquisition in an effort to find a third way—a general alternative to concepts being either learned or innate. In doing this, he tried to explain why it is that, though no concepts are learned on his account, they nonetheless appear to be learned. His explanation was that the appearance of learning stemmed from the fact that stereotype formation is a first step in concept acquisition (after which purely biological processes take over). In Chapter 26, we saw that there are a number of acute problems with this view. Most importantly, it fails to explain how so many concepts that lack a stereotype are acquired and why there is so much conceptual variation within and across different cultures. The obvious explanation for the breadth of conceptual variation, of course, is that concepts are learned—often on the basis of rationalist learning mechanisms.

We take this obvious explanation to be the correct one and have argued that, while it is incompatible with Fodor's radical concept nativism, it is fully compatible with our own version of concept nativism. In spelling out how our concept nativism relates to cultural learning, it emerged that there is a way in which our account of concept nativism, like Fodor's, implies that there is a rationalist account of the origins of essentially every lexical concept. But despite this commonality, our view is nothing at all like Fodor's. His rationalism about lexical concepts is the extremely implausible view that virtually all lexical concepts are

simply innate. In contrast, we have argued that most lexical concepts are learned and that this learning almost always involves some form of rationalist learning mechanism, making the overall account of these concepts a rationalist one. In some cases, this will be a minimally rationalist account. But we have argued that in many cases—as with the concept LIGHTBULB—the learning will be rationalist not merely to a minimal extent, but instead will be robustly rationalist. What's more, because cultural learning itself draws on a considerable assortment of characteristically rationalist structures in the acquisition base, concept nativism, as we understand it, isn't merely compatible with cultural learning. Concept nativism turns out to be fundamental to explaining the very possibility of cultural learning and thereby to explaining how the human mind comes to have such an enormously rich and varied conceptual system.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0027

Coda

Innate Ideas Revisited

This book began with a question from Locke. "How comes it [the mind] to be furnished? Whence comes it by that vast store, which the busy and boundless Fancy of Man has painted on it, with an almost endless variety?" (1690/1975, II.I.2, p. 104). The answer we have given has been a rationalist account of the origins of concepts, one that Locke would have recognized as a defence of innate ideas.

A crucial element in our case for concept nativism has been the understanding of what is and what isn't involved in the rationalism-empiricism debate that we developed in Part I of the book. We argued there that this debate shouldn't be viewed as a disagreement about the relative contributions of nature and nurture, or as about the cogency of the theoretical notion of innateness, or the question of whether concepts are learned or not. Many theorists have dismissed rationalist views, and even the entire rationalism-empiricism debate, by framing things in these terms. But in our view, all of these approaches are mistaken. None of these alternative ways of understanding the debate captures the reality of what is at stake among theorists seeking to understand the origins of concepts. And none can account for the fact that the rationalism-empiricism debate has been so productive in guiding research on the origins of concepts.

Our own view is that the rationalism-empiricism debate is about the contents of what we have called the *acquisition base*. In other words, it's about the ultimate psychological basis for the acquisition of psychological traits. For the debate about the origins of concepts, this means that what concept nativists and concept empiricists disagree about is what kinds of structures provide the ultimate psychological basis for the acquisition of concepts. As we have noted, any theorist at all who accepts that concepts are a part of the mind must accept that there is such an ultimate psychological basis for their acquisition. What rationalists and empiricists disagree about is what this unlearned basis consists in.

Empiricist accounts of the origins of concepts predominantly trace them back to what we have called *characteristically empiricist psychological structures* in the acquisition base. These include domain-general learning mechanisms, sensorimotor representations, and low-level attention biases. Rationalists welcome such structures as *part* of the acquisition base, but also see the acquisition base as containing numerous *characteristically rationalist psychological structures*. These include domain-specific learning mechanisms, abstract representations,

and other types of specialized psychological structures, which vary along such dimensions as complexity, diversity, articulation, abstractness, degree of domain specificity, and alignment, as discussed in Chapter 2.

As we have emphasized throughout the book, both concept nativism and concept empiricism come in many forms. It's possible to be a concept nativist while maintaining that there is a rationalist account of the origins of concepts in only a few conceptual domains. It is even possible to be a concept nativist while holding that there are *no* innate concepts—since what matters is whether the concepts the account covers are innate *or* acquired via rationalist learning mechanisms. Advocates of the core knowledge hypothesis, who are among the most prominent rationalists in cognitive science, claim that rationalist core knowledge systems are present in only a handful of conceptual domains. Our own view, by contrast, is that there is a robustly rationalist story of the origins of *many concepts* across *many conceptual domains*. Accordingly, our view stands in opposition not only to empiricist views, but also to many rationalist accounts that are less rationalist than our own.

The heart of our case for concept nativism, which we presented in Part II of the book, takes the form of a wide-ranging inference to the best explanation that incorporates the contributions of at least seven subsidiary arguments: the argument from early development, the argument from animals, the argument from universality, the argument from initial representational access, the argument from neural wiring, the argument from prepared learning, and the argument from cognitive and behavioural quirks. Though some of these arguments are widely understood to be arguments for rationalist accounts of conceptual development, others are not widely known, and even the most familiar of these arguments are frequently misunderstood or underappreciated. One of our aims has been to carefully separate out these distinct strands in the overall case for concept nativism, while clarifying the logic of each argument and addressing some of the misunderstandings associated with them. A second aim has been to illustrate each of these arguments with a variety of different case studies, in order to demonstrate both the breadth and depth of the inference to the best explanation supporting concept nativism—breadth in that these case studies argue for a rationalist account of the origins of many concepts across many content domains, and depth in that the case for a rationalist account of the origins of these many concepts will typically be supported by multiple lines of argument and a diverse body of evidence.

Our case for concept nativism would be seriously incomplete, however, if we didn't also address alternative empiricist approaches. This was the focus of Part III, where we discussed a representative sample of the most important and influential empiricist alternatives and objections to concept nativism. We argued there that none of these undermines our rationalist position. Our discussion in this part of the book also allowed us to examine a number of additional conceptual domains, and to further develop the case for concept nativism. One of the key

morals of our discussion was that many of the tools and insights associated with empiricist theorizing are perfectly compatible with concept nativism. In fact, domain-general learning approaches are not only compatible with an overall rationalist framework, they can achieve significantly greater explanatory power when they are incorporated into such a framework. This means taking conceptual development to be grounded in an acquisition base which, in addition to domain-general learning mechanisms and other characteristically empiricist psychological structures, also includes a variety of characteristically rationalist psychological structures.

While our view of the origins of concepts contrasts with both empiricist views and many rationalist views that are less rationalist than our own, it also contrasts with the most well-known rationalist account of the origins of concepts—Jerry Fodor's radical concept nativism. Part IV addressed this view and related theoretical alternatives stemming from Fodor's infamous argument against concept learning. We showed that Fodor's argument fails, but more importantly, we showed *why* it fails. Fodor's charge that hypothesis testing models are circular makes the mistake of focusing on just a single point in time—the very moment at which the correct hypothesis is confirmed—without appreciating that learning by hypothesis testing is a process that unfolds over time. Once we pan out to see this entire process, it's clear that a complex concept can be constructed and evaluated in the course of this process and so isn't something that a learner must possess before the process can even begin. We also showed that the Acquisition by Composition model of concept learning, which is not particular to Fodor but is in fact very widely accepted, does not impose the kinds of limits on concept acquisition that its advocates have assumed it does. Despite the fact that they are not composed of other representations, primitive concepts can be learned. Part IV explored some of the different ways in which such learning could proceed.

Part IV also further developed our positive account of the origins of concepts, bringing it to its full fruition. At the heart of this development was the question of how our concept nativism can accommodate the undeniable importance of cultural learning to concept acquisition. We argued that while the impact of culture on conceptual development is often thought to support empiricist accounts of the origins of concepts, cultural learning is thoroughly steeped in rationalist cognitive resources. Learning is essential to the acquisition of culturally embedded concepts, but *general-purpose* learning only goes so far. To properly understand how our species is capable of acquiring the vast array of culturally embedded concepts, a thoroughgoing rationalist account of the origins of concepts of the sort that we endorse is necessary.

In fact, we argued that meeting the challenge of culturally embedded concepts actually requires a more encompassing form of concept nativism than the view we argued for and developed in Parts I–III; as strong as the account there was, it was not strong enough. Looking back, what we effectively argued in Parts I–III

was that there is a *robustly rationalist* account of the origins of *many concepts across many conceptual domains*. But, as we argued in Part IV, the correct thing to say is actually that, in addition to there being a robustly rationalist account of the origins of many concepts across many conceptual domains, there is also at least a *minimally rationalist* account of *virtually all lexical concepts*. Crucially, however, this does not mean that learning and culture play any less of a role in our account of the origins of all of these concepts. Indeed, not only is the fact that there is at least a minimally rationalist account of the origins of virtually all lexical concepts compatible with cultural learning being vital to explaining the origins of concepts, this fact is at the heart of what makes cultural learning possible. Only a rationalist account of this sort is able to do full justice to the enormous role that culture plays in shaping human conceptual development.

Our version of concept nativism, then, takes the following form. A minimally rationalist concept nativism forms the backdrop for concept acquisition, with there being at least a minimally rationalist account of the origins of essentially all lexical concepts. And over and above this, a richer robustly rationalist form of concept nativism applies to many concepts across many conceptual domains. While recent decades have seen a virtual deluge of exciting work bearing on these issues from across a broad range of disciplines, there is still an enormous amount that we don't yet know. So there is no way that anything like a definitive list of which concepts have their origins explained in robustly rationalist terms can be given at this time.[1] But we have proposed that the following concepts and conceptual domains are among those likely to be included on this list:

> agency, animals, artefacts, belief, causation, coalitions, collections, communication, containment, cooperation, dangerous animals, death, disease, emotions, essence, events, faces, fairness, feature/property, food, formidability, function/purpose, gains and losses, geometry, giving and taking, goal, harm, hazards/danger, human body, individuals, in-group and out-group, kinship, language, life stages, logic, loyalty, meat, modality, morality, movement, music, norms, number, objects, obligation, ownership, path, plants, perception, predators and prey, preferences, prestige, sameness and difference, sex, social dominance, social groups, spatial magnitude, self, stuffs/substances, teaching/pedagogy, time, and tools.

[1] The sample of concepts and conceptual domains in the list that follows is only meant to be indicative of what a full list might include. Precisely which concepts and domains would ultimately make it onto such a list will turn on a number of factors, including, of course, whether a content domain is characterized in a course- or fine-grained manner. In some cases, we have listed individual concepts or narrow conceptual clusters in light of a prominent ongoing debate about these concepts (e.g., the debate about the origins of the concept of belief). In other cases, we have chosen to list both a superordinate category and one or more of its subordinate categories (e.g., *animals* and *dangerous animals*) because we have presented arguments or evidence bearing on there being a rationalist account for both the narrower domain(s) and the broader domain.

When Chomsky first urged researchers to revisit the debate about innate ideas, he speculated that rationalist theories of language acquisition were just the beginning and that further research would uncover rationalist systems involved in many aspects of human psychology. This book is in many ways an extended exploration of his suggestion in relation to the representations that form the basis for categorization, inference, memory, planning, decision making, and other forms of higher cognition. We have argued that, just as Chomsky speculated, a rationalist framework provides the best account of the origin of these building blocks of thought.

The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts. Stephen Laurence and Eric Margolis, Oxford University Press. © Stephen Laurence and Eric Margolis 2024. DOI: 10.1093/9780191925375.003.0028

References

Adachi, I., Kuwahata, H., & Fujita, K. (2006). Dogs recall their owner's face upon hearing the owner's voice. *Animal Cognition*, 10(1), 17–21.

Adachi, I., Chou, D. P., & Hampton, R. R. (2009). Thatcher effect in monkeys demonstrates conservation of face perception across primates. *Current Biology*, 19(15), 1270–1273.

Adams, R. M. (1975). Where do our ideas come from?—Descartes vs. Locke. In S.P. Stich (ed.), *Innate Ideas*, pp. 71–87. Berkeley, CA: University of California Press.

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., Ge, Y., Gunawardena, J., Hendrickson, R. C., Hergenrother, P. J., Huber, C. G., Ivanov, A. R., Jensen, O. N., Jewett, M. C., Kelleher, N. L., Kiessling, L. L., Krogan, N. J., Larsen, M. R., Loo, J. A., Ogorzalek Loo, R. R., Lundberg, E., MacCoss, M. J., Mallick, P., Mootha, V. K., Mrksich, M., Muir, T. W., Patrie, S. M., Pesavento, J. J., Pitteri, S. J., Rodriguez, H., Saghatelian, A., Sandoval, W., Schlüter, H., Sechi, S., Slavoff, S. A., Smith, L. M., Snyder, M. P., Thomas, P. M., Uhlén, M., Van Eyk, J. E., Vidal, M., Walt, D. R., White, F. M., Williams, E. R., Wohlschlager, T., Wysocki, V. H., Yates, N. A., Young, N. L., & Zhang, B. (2018). How many human proteoforms are there? *Nature Chemical Biology*, 14(3), 206–214.

Agrillo, C., Dadda, M., & Bisazza, A. (2007). Quantity discrimination in female mosquitofish. *Animal Cognition*, 10, 63–70.

Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2009). Use of number by fish. *PloS One*, 4(3), e4786.

Agrillo, C., Piffer, L., & Bisazza, A. (2010). Large number discrimination by mosquitofish. *PloS One*, 5(12), e15232.

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3), 299–352.

Alberts, B. (1998). The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, 92(3), 291–294.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2015). *Molecular Biology of the Cell*, 6th ed. London: Garland Science.

Alexander, L., & Moore, M. (2015). Deontological ethics. *The Stanford Encyclopedia of Philosophy* (Spring 2015 edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/>.

Altman, M. N., Khislavsky, A. L., Coverdale, M. E., & Gilger, J. W. (2016). Adaptive attention: How preference for animacy impacts change detection. *Evolution and Human Behavior*, 37(4), 303–314.

Annas, J. E. (1992). *Hellenistic Philosophy of Mind*. Berkeley, CA: University of California Press.

Antell, S., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development*, 54(3), 695–701.

Antony, L. M. (2000). Natures and norms. *Ethics*, 111(1), 8–36.

Antony, L. M. (2012). Different voices or perfect storm: Why are there so few women in philosophy? *Journal of Social Philosophy*, 43(3), 227–255.

Apperly I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 953–70.

Arcaro, M. J., & Livingstone, M. S. (2017). A hierarchical, retinotopic proto-organization of the primate visual system at birth. *eLife*, 6, e26196.

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., & Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience*, 20(10), 1–16.

Ardiel, E. L., & Rankin, C. H. (2010). An elegant mind: Learning and memory in *Caenorhabditis elegans*. *Learning & Memory*, 17(4), 191–201.
Ariew, A. S. (1996). Innateness and canalization. *Philosophy of Science*, 63, S19–S27.
Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63(7), 602–614.
Astuti, R., & Harris, P. L. (2008). Understanding mortality and the life of the ancestors in rural Madagascar. *Cognitive Science*, 32(4), 713–740.
Atran, S., & Medin, D. (2008). *The Native Mind and the Cultural Construction of Nature*. Cambridge, MA: MIT Press.
Ayer, A.J. (1946/1952). *Language, Truth, and Logic*. Mineola, NY: Dover.
Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2020). *Memory*, 3rd ed. Hove: Psychology Press.
Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, 23(5), 655.
Baillargeon, R. (2000). Reply to Bogartz, Shinskey, and Schilling; Schilling; and Cashon and Cohen. *Infancy*, 1(4), 447–462.
Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, 46, 112–124.
Baillargeon, R., & DeJong, G. F. (2017). Explanation-based learning in infancy. *Psychonomic Bulletin & Review*, 24(5), 1511–1526.
Baillargeon, R., & DeVos, J. (1991). Object permanence in young infants: Further evidence. *Child development*, 62(6), 1227–1246.
Baillargeon, R., Graber, M., DeVos, J., & Black, J. (1990). Why do young infants fail to search for hidden objects. *Cognition*, 36(3), 255–284.
Baillargeon, R., Li, J., Gertner, Y., & Wu, D. (2011). How do infants reason about physical events? In U. Goswami (ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, pp. 11–48, 2nd ed. Oxford: Wiley-Blackwell.
Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14(3), 110–118
Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K. S., Wu, D., & Bian, L., (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer, P. R. Shaver (eds.), E. Borgida, & J. A. Bargh (assoc. eds.), *APA Handbook of Personality and Social Psychology*, Vol. 1: *Attitudes and Social Cognition*, pp. 79–150. Washington, DC: American Psychological Association.
Baillargeon, R., Stavans, M., Wu, D., Gertner, Y., Setoh, P., Kittredge, A. K., & Bernard, A. (2012). Object individuation and physical reasoning in infancy: An integrative account. *Language Learning and Development*, 8(1), 4–46.
Baillargeon, R., Wu, D., Yuan, S., Li, J., & Luo, Y. (2009) Young infants' expectations about self-propelled objects. In B. M. Hood and L. R. Santos (eds.), *The Origins of Object Knowledge*, pp. 285–352. Oxford: Oxford University Press.
Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, 140(6), 1556–1581.
Bardi, L., Regolin, L., & Simion, F. (2011). Biological motion preference in humans at birth: Role of dynamic and configural properties. *Developmental Science*, 14(2), 353–359.
Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of Child Language*, 44(3), 553–590.
Barnes, M. E., Evans, E. M., Hazel, A., Brownell, S. E., & Nesse, R. M. (2017). Teleological reasoning, not acceptance of evolution, impacts students' ability to learn natural selection. *Evolution: Education and Outreach*, 10(1), 1–12.
Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind?" *Cognition*, 21(1), 37–46.

Barrett, H. C. (2009). Where there is an adaptation, there is a domain: The form-function fit in information processing. In S. M. Platek & T. K. Shackelford (eds.), Foundations in Evolutionary Cognitive Neuroscience, pp. 97–116. Cambridge University Press.

Barrett, H. C. (2015). *The Shape of Thought: How Mental Adaptations Evolve*. New York: Oxford University Press.

Barrett, H. C., & Behne, T. (2005). Children's understanding of death as the cessation of agency: A test using sleep versus death. *Cognition*, 96(2), 93–108.

Barrett, H. C., Bolyanatz, A., Broesch, T., Cohen, E., Froerer, P., Kanovsky, M., Schug, M. G., & Laurence, S. (2021). Intuitive dualism and afterlife beliefs: A cross-cultural study. *Cognitive Science*, 45(6), e12992.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A. C., Scelza, B. A., Stich, S., von Reudon, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17), 4688–4693.

Barrett, H. C., & Broesch, J. (2012). Prepared social learning about dangerous animals in children. *Evolution and Human Behavior*, 33(5), 499–508.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., & Laurence, S. (2013a). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences*, 280(1755), 20122654.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., & Laurence, S. (2013b). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences*, 280 (supplement), 1–29.

Barrett, H. C., Laurence, S., & Margolis, E. (2008). Artifacts and original intent: A cross-cultural perspective on the design stance. *Journal of Cognition and Culture*, 8, 1–22.

Barrett, H. C., Peterson, C. D., & Frankenhuis, W. E. (2016). Mapping the cultural learnability landscape of danger. *Child Development*, 87(3), 770–781.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 637–660.

Bar-Shai, N., Keasar, T., & Shmida, A. (2011). The use of numerical information by bees in foraging tasks. *Behavioral Ecolology*, 22, 317–325.

Basalla, G. (1988). *The Evolution of Technology*. Cambridge: Cambridge University Press.

Bates, E., Elman, J., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1998). Innateness and emergentism. In W. Bechtel & G. Graham (eds.), *A Companion to Cognitive Science*, pp. 590–601. Oxford: Blackwell.

Bateson, P. (2000). Taking the stink out of instinct. In H. Rose and S. Rose (eds.), *Alas, Poor Darwin*, pp. 189–207. London: Jonathan Cape.

Bateson, P. (2005). The return of the whole organism. *Journal of Biosciences*, 30(1), 31–39.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 66(2), 175–207.

Beck, J. (2013). The generality constraint and the structure of thought. *Mind*, 121(483), 563–600.

Bedny, M., Caramazza, A., Pascual-Leone, A., & Saxe, R. (2012). Typical neural representations of action verbs develop without vision. *Cerebral Cortex*, 22(2), 286–293.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27), 11312–11317.

Bellugi, U., Lichtenberger, L., Jones, W., Lai, Z., & St George, M. (2000). The neurocognitive profile of Williams syndrome: A complex pattern of strengths and weaknesses. *Journal of Cognitive Neuroscience*, 12(supplement 1), 7–29.

Berdoy, M., Webster, J. P., & Macdonald, D. W. (2000). Fatal attraction in rats infected with *Toxoplasma gondii*. *Proceedings of the Royal Society B: Biological Sciences*, 267(1452), 1591–1594.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253–3258.
Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, 11(4), 369–380.
Bergen, B. K. (2012). *Louder than Words: The New Science of How the Mind Makes Meaning*. New York: Basic Books.
Berger, S. E., Harbourne, R. T., Arman, F., & Sonsini, J. (2019). Balancing act(ion): Attentional and postural control strategies predict extent of infants' perseveration in a sitting and reaching task. *Cognitive Development*, 50, 13–21.
Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, 302(5648), 1234–1236.
Bering, J. M. (2002). The existential theory of mind. *Review of General Psychology*, 6(1), 3–24.
Berkeley, G. (1709/1975). *An Essay towards a New Theory of Vision*. In M. R. Ayers (ed.), *G. Berkeley, Philosophical Works*, pp. 1–59. Totowa, NJ: Rowman & Littlefield.
Berkeley, G. (1710/1975). *A Treatise Concerning the Principles of Human Knowledge*. In M. R. Ayers (ed.), *G. Berkeley, Philosophical Works*, pp. 61–127. Totowa, NJ: Rowman & Littlefield.
Berkeley, G. (1713/1975). *Three Dialogues between Hylas and Philonous*, 3rd ed. In M. R. Ayers (ed.), *G. Berkeley, Philosophical Works*, pp. 129–207. Totowa, NJ: Rowman & Littlefield.
Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105), 912.
Bian, L., Sloane, S., & Baillargeon, R. (2018). Infants expect ingroup support to override fairness when resources are limited. *Proceedings of the National Academy of Sciences*, 115(11), 2705–2710.
Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., & Wrangham, R. (2015). The ontogeny of fairness in seven societies. *Nature*, 528(7581), 258–261.
Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, 58, 47–73.
Block, N. (1986). Advertisement for a semantics for psychology. In P. A. French, T. Uehling Jr, & H. Wettstein (eds.), *Midwest Studies in Philosophy*, Vol. X: *Studies in the Philosophy of Mind*, pp. 615–678. Minneapolis: University of Minnesota Press.
Block, N. (1995). How heritability misleads about race. *Cognition*, 56(2), 99–128.
Bloom, P. (1996) Intention, history, and artifact concepts. *Cognition*, 60(1), 1–29.
Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1), 25–31.
Bola, Ł., Zimmermann, M., Mostowski, P., Jednoróg, K., Marchewka, A., Rutkowski, P., & Szwed, M. (2017). Task-specific reorganization of the auditory cortex in deaf humans. *Proceedings of the National Academy of Sciences*, 114(4), E600–E609.
Bonanni, R., Natoli, E., Cafazzo, S., & Valsecchi, P. (2011). Free-ranging dogs assess the quantity of opponents in intergroup conflicts. *Animal Cognition*, 14(1), 103–115.
Bor, A. (2018). Correcting for base rates in multidimensional "Who said what?" experiments. *Evolution and Human Behavior*, 39(5), 473–478.
Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 2(1), 115–129.
Boroditsky, L., & Prinz, J. (2008). What are thoughts made of? In G. R. Semin & E. R. Smith (eds.), *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, pp. 98–115. Cambridge: Cambridge University Press.
Boyd, R. (2018). *A Different Kind of Animal*. Princeton, NJ: Princeton University Press.
Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. The University of Chicago Press.
Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences*, 7(3), 119–124.
Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.
Brandom, R. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.
Brannon, E. M., & Terrace, H. (1998). Ordering of the numerosities 1 to 9 by monkeys. *Science*, 282, 746–749.
Broesch, J., Barrett, H. C., & Henrich, J. (2014). Adaptive content biases in learning about animals across the life course. *Human Nature*, 25(2), 181–199.
Brooks, R. A. (1997). From earwigs to humans. *Robotics and Autonomous Systems*, 20, 291–304.
Brooks, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.
Brown, A. A., Spetch, M. L., & Hurd, P. L. (2007). Growing in circles: Rearing environment alters spatial navigation in fish. *Psychological Science*, 18, 569–573.
Brown, D. (1991). *Human Universals*. New York: McGraw-Hill.
Brown, R. L. (2019). Infer with care: A critique of the argument from animals. *Mind & Language*, 34(1), 21–36.
Browne, K. R. (1999). *Divided Labours: An Evolutionary View of Women at Work*. New Haven, CT: Yale University Press.
Browne, K. R. (2002). *Biology at Work: Rethinking Sexual Equality*. New Brunswick, NJ: Rutgers University Press.
Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930–940.
Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), 1–19.
Bueno-Guerra, N., & Amici, F. (2018). *Field and Laboratory Methods in Animal Cognition: A Comparative Guide*. Cambridge: Cambridge University Press.
Buller, D. J. (2009). Four fallacies of pop evolutionary psychology. *Scientific American*, 300(1), 74–81.
Buller, D. J., & Hardcastle, V. (2000). Evolutionary psychology, meet developmental neurobiology: Against promiscuous modularity. *Brain and Mind*, 1(3), 307–325.
Bulnes, L. C., Marien, P., Vandekerckhove, M., & Cleeremans, A. (2019). The effects of Botulinum toxin on the detection of gradual changes in facial emotion. *Scientific Reports*, 9(1), 1–13.
Bunn, H. T. (2007). Meat made us human. In P. S. Ungar (ed.), *Evolution of the Human Diet: The Known, the Unknown and the Unknowable*, pp. 191–211. Oxford: Oxford University Press.
Burge, T. (1979) Individualism and the Mental. In P. A. French, T. E. Uehling Jr, H. K. Wettstein (eds.), *Midwest Studies in Philosophy: Studies in Metaphysics*, pp. 73–121. Minneapolis, MN: University of Minnesota Press.
Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review*, 125(3), 409–434.
Bushdid, C., Magnasco, M. O., Vosshall, L. B., & Keller, A. (2014). Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343(6177), 1370–1372.
Busigny, T., Graf, M., Mayer, E. N., & Rossion, B. (2010). Acquired prosopagnosia as a face-specific disorder: Ruling out the general visual similarity account. *Neuropsychologia*, 48(7), 2051–2067.
Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: Evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia*, 48(14), 4057–4092.
Buss, D. M. (2015). *The Handbook of Evolutionary Psychology*, 2 Vols. Hoboken, NJ: Wiley.
Buss, D. M. (2019). *Evolutionary Psychology: The New Science of the Mind*, 6th ed. New York: Pearson.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PloS One*, 12(4), e0173793.
Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112, 337–42.
Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology*, 119, 120–126.
Butterfill S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28, 606–37.
Buyukozer Dawkins, M., Sloane, S., & Baillargeon, R. (2019). Do infants in the first year of life expect equal resource allocations? *Frontiers in Psychology*, 10, 1–19.
Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5), 187–192.
Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., Tapanya, S., & Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science*, 16(5), 378–384.
Calvillo, D. P., & Hawkins, W. C. (2016). Animate objects are detected more frequently than inanimate objects in inattentional blindness tasks independently of threat. *The Journal of General Psychology*, 143(2), 101–115.
Camp, E. (2004). The generality constraint and categorial restrictions. *The Philosophical Quarterly*, 54(215), 209–231.
Cantlon, J. F., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science*, 17(5), 401–406.
Capitani, E., Laiacona, M., Mahon, B. Z., & Caramazza, A. (2003). What are the facts of category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology*, 20, 213–261.
Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10, 1–34.
Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
Carey, S. E., & Spelke, E. S. (1996). Science and core knowledge. *Philosophy of Science*, 63(4), 515–533.
Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.
Carnap, R. (1932/1959). The elimination of metaphysics through logical analysis of language. In A. J. Ayer (ed.), *Logical Positivism*, pp. 60–81. New York: The Free Press.
Carpenter, M., Pennington, B. F., & Rogers, S. J. (2001). Understanding of others' intentions in children with autism. *Journal of Autism and Developmental Disorders*, 31(6), 589–599.
Carrier, D. R., & Morgan, M. H. (2015). Protective buttressing of the hominin face. *Biological Reviews*, 90(1), 330–346.
Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, 4(14), 278–280.
Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.
Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.
Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2), 141–172.
Carruthers, P. (2020). Representing the mind as such in infancy. *Review of Philosophy and Psychology*, 11(4), 765–781.
Casasanto, D. (2014). Bodily relativity. In L. Shaprio (ed.), *The Roultledge Handbook of Embodied Cognition*, pp. 108–117. London: Routledge University Press.
Cashdan, E. (1994). A sensitive period for learning about food. *Human Nature*, 5(3), 279–291.
Cashdan, E. (1998). Adaptiveness of food learning and food aversions in children. *Social Science Information*, 37(4), 613–632.

Casini, L., & Macar, F. (1997). Effects of attention manipulation on perceived duration and intensity in the visual modality. *Memory & Cognition*, 25, 812–818.
Caves, E. M., Green, P. A., Zipple, M. N., Peters, S., Johnsen, S., & Nowicki, S. (2018). Categorical perception of colour signals in a songbird. *Nature*, 560(7718), 365–367.
Cecchini, M., Baroni, E., Di Vito, C., Piccolo, F., & Lai, C. (2011). Newborn preference for a new face vs. a previously seen communicative or motionless face. *Infant Behavior and Development*, 34(3), 424–433.
Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*, 157(1), 77–94.
Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, 11(1), 1–9.
Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, 359(6381), 1263–1266.
Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, 34, 1131–1157.
Cheeseman, J. F., Millar, C. D., Greggers, U., Lehmann, K., Pawley, M. D. M., Gallistel, C. R., Warman, G. R., & Menzel, R. (2014). Way-finding in displaced clock-shifted bees proves bees use a cognitive map. *Proceedings of the National Academy of Sciences*, 111(24), 8949–8954.
Chen, M. K., Lakshminarayanan, V. R., & Santos, L. R. (2006). How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *Journal of Political Economy*, 114(3), 517–537.
Cheney, D. L., & Seyfarth, R. M. (1999). Recognition of other individuals' social relationships by female baboons. *Animal Behaviour*, 58(1), 67–75.
Cheney, D. L., & Seyfarth, R. M. (2008). *Baboon Metaphysics: The Evolution of a Social Mind*. Chicago, IL: University of Chicago Press.
Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, 23, 149–178.
Chiandetti, C., Spelke, E. S., & Vallortigara, G. (2015). Inexperienced newborn chicks use geometry to spontaneously reorient to an artificial social partner. *Developmental Science*, 18(6), 972–978.
Chiandetti, C., & Vallortigara, G. (2010). Experience and geometry: controlled-rearing studies with chicks. *Animal Cognition*, 13.3, 463–470.
Chiandetti, C., & Vallortigara, G. (2011). Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718), 2621–2627.
Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V., Rao, P., & Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *The Lancet*, 392(10162), 2388–2396.
Chittka, L. (2022). *The Mind of a Bee*. Princeton, NJ: Princeton University Press.
Chittka, L. (2023). *The Mind of a Bee*. Princeton University Press.
Chittka, L., Faruq, S., Skorupski, P., & Werner, A. (2014). Colour constancy in insects. *Journal of Comparative Physiology A*, 200(6), 435–448.
Cho, I., Lee, Y., & Song, H. J. (2021). Six-month-olds' ability to use linguistic cues when interpreting others' pointing actions. *Infant Behavior and Development*, 64, 101621.
Choi, Y. J., Mou, Y., & Luo, Y. (2018). How do 3-month-old infants attribute preferences to a human agent? *Journal of Experimental Child Psychology*, 172, 96–106.
Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
Chomsky, N. (1959). A review of B. F. Skinner's *Verbal Behavior*. *Language*, 35(1), 26–58.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1966). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Lanham, MD: University Press of America.
Chomsky, N. (1967). Recent contributions to the theory of innate ideas. *Synthese*, 17(1), 2–11.
Chomsky, N. (1971). *Problems of Knowledge and Freedom*. New York: Pantheon Books.

Chomsky, N. (1972/2006) *Language and Mind*, 3rd ed. New York: Cambridge University Press.
Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon Books.
Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. London: Greenwood Publishing Group.
Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.
Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
Chomsky, N., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry*, 8(3), 425–504.
Chuang, M. F., Kam, Y. C., & Bee, M. A. (2017). Territorial olive frogs display lower aggression towards neighbours than strangers based on individual vocal signatures. *Animal Behaviour*, 123, 217–228.
Churchland, P. (2012). *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA: MIT Press.
Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, 3(9), 345–351.
Clearfield, M. W., & Mix, K. S. (2001). Amount versus number: Infants' use of area and contour length to discriminate small sets. *Journal of Cognition and Development*, 2(3), 243–260.
Clifford, A., Franklin, A., Davies, I. R. L., & Holmes, A. (2009). Electrophysiological markers of categorical perception of color in 7-month old infants. *Brain and Cognition*, 71, 165–172.
Cohen, L. B., Chaput, H. H., & Cashon, C. H. (2002). A constructivist model of infant cognition. *Cognitive Development*, 17, 1323–1343.
Cook, V., & Newson, M. (2007) *Chomsky's Universal Grammar: An Introduction*, 3rd ed. Oxford: Blackwell Publishing.
Cope, A. J., Vasilaki, E., Minors, D., Sabo, C., Marshall, J. A. R., & Barron, A. B. (2018). Abstract concept learning in a simple neural network inspired by the insect brain. *PLoS Computational Biology*, 14(9), e1006435.
Cordes, S., & Brannon, E. M. (2008). The difficulties of representing continuous extent in infancy: Using number is just easier. *Child development*, 79(2), 476–489.
Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 85–116. Cambridge: Cambridge University Press.
Cosmides, L., & Tooby, J. (1997). The modular nature of human intelligence. In A. B. Scheibel and J. W. Schopf (eds.), *The Origin and Evolution of Intelligence*, pp. 71–101. London: Jones and Bartlett Publishers International
Cottingham, J. (1988). *The Rationalists*. Oxford: Oxford University Press.
Coulon, M., Guellai, B., & Streri, A. (2011). Recognition of unfamiliar talking faces at birth. *International Journal of Behavioral Development*, 35(3), 282–287.
Cowell, J. M., & Decety, J. (2015). Precursors to morality in development as a complex interplay between neural, socioenvironmental, and behavioral facets. *Proceedings of the National Academy of Sciences*, 112(41), 12657–12662.
Cowie, F. (1999). *What's within? Nativism reconsidered*. New York: Oxford University Press.
Cowie, F. (2009). Why isn't Stich an eliminativist? In D. Murphy and M. Bishop (eds.), *Stich and His Critics*, pp. 74–100. Oxford: Wiley-Blackwell.
Crain, S. (2012) *The Emergence of Meaning*. Cambridge: Cambridge University Press.
Crain, S., & Khlentzos, D. (2010). The logic instinct. *Mind & Language*, 25(1), 30–65.
Crain, S., & Pietroski, P. (2001). Nature, nurture, and universal grammar. *Linguistic and Philosophy*, 24, 139–186.
Crain, S., & Thorton, R. (1998). *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.
Crane, T. (1988). The waterfall illusion. *Analysis*, 48(3), 142–147.
Crivelli, C., & Fridlund, A. J. (2019). Inside-out: From basic emotions theory to the behavioral ecology view. *Journal of Nonverbal Behavior*, 43(2), 161–194.
Crouzet, S. M., Joubert, O. R., Thorpe, S. J., & Fabre-Thorpe, M. (2012). Animal detection precedes access to scene category. *PloS One*, 7(12), e51471.
Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148–153.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567), 1149–1157.

Curtis, V., De Barra, M., & Aunger, R. (2011). Disgust as an adaptive system for disease avoidance behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1563), 389–401.

Dacke, M., & Srinivasan, M. V. (2008). Evidence for counting in insects. *Animal Cognition*, 11(4), 683–689.

Daly, C. (1994). Tropes. *Proceedings of the Aristotelian Society*, 94, 253–261. Aristotelian Society. Hoboken, NJ: Wiley.

Daly, M., & Wilson, M. (1988). *Homicide*. New York: Aldine de Gruyter.

Dancy, J. (1987). *Berkeley: An Introduction*. Oxford: Blackwell Publishers.

Darmaillacq, A.-S., Chichery, R., Shashar, N., & Dickel, L. (2006). Early familiarization overrides innate prey preference in newly hatched *Sepia officinalis* cuttlefish. *Animal Behaviour*, 71(3), 511–514.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398(6724), 203–204.

Davidson, D. (1975). Thought and talk. In his *Inquiries into Truth and Interpretation*, pp. 155–170. Oxford: Oxford University Press.

DeBellis, M. A. (1995). *Music and Conceptualizaiton*. Cambridge University Press.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford: Oxford University Press.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Viking.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3), 626.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, 311(5759), 381–384.

de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences*, 111(13), 4809–4813.

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1(4), 568–570.

De Rosa, R. (2004). Locke's *Essay, Book I*: The question-begging status of the anti-nativist arguments. *Philosophy and Phenomenological Research*, 69(1), 37–64.

Descartes, R. (1637/1985). *Optics*. In J. Cottingham, R. Stoothoff, and D. Murdoch, (trans.) *The Philosophical Writings of Descartes*, Vol. I, pp. 152–175. Cambridge: Cambridge University Press.

Descartes, R. (1641/1984). *Meditations on First Philosophy* and *Objections and Replies*. In J. Cottingham, R. Stoothoff, and D. Murdoch, (trans.) *The Philosophical Writings of Descartes*, Vol. II, pp. 1–397. Cambridge: Cambridge University Press.

Descartes, R. (1641/1991). To Mersenne, 22 July 1641. In J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny (trans.) *The Philosophical Writings of Descartes*, Vol. III: *The Correspondence*, pp. 187. Cambridge: Cambridge University Press.

Descartes, R. (1648/1985). *Comments on a Certain Broadsheet*. In J. Cottingham, R. Stoothoff, and D. Murdoch, (trans.) *The Philosophical Writings of Descartes*, Vol. I, pp. 249–311. Cambridge: Cambridge University Press.

de Villiers, J. G., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell, and K. Riggs (eds.), *Children's Reasoning and the Mind*, pp. 205–242. Hove: Psychology Press.

de Villiers, J. G., & de Villiers, P. A. (2009). Complements enable representation of the contents of false beliefs: Evolution of a theory of theory of mind. In S. Foster-Cohen (ed.), *Language Acquisition*, pp. 169–195. London: Palgrave Macmillan.

de Waal, F. (2002) *The Ape and the Sushi Master: Cultural Reflections of a Primatologist*. New York: Basic Books.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*, 21(12), 1871–1877.

Diamond, A. (1990). The development and neural bases of memory functions as indexed by the A$\overline{B}$ and delayed response tasks in human infants and infant monkeys. *Annals of the New York Academy of Sciences*, 608(1), 267–317.

Diamond, A., Cruttenden, L., & Neiderman, D. (1994). A$\overline{B}$ with multiple wells: 1. Why multiple wells are sometimes easier than two wells. 2. Memory or memory + inhibition. *Developmental Psychology*, 30, 192–205.

Diamond, J. (2013). *The World Until Yesterday: What Can We Learn from Traditional Societies?* London: Penguin.

Di Giorgio, E., Leo, I., Pascalis, O., & Simion, F. (2012). Is the face-perception system human-specific at birth? *Developmental Psychology*, 48(4), 1083–1090.

Di Giorgio, E., Lunghi, M., Rugani, R., Regolin, L., Dalla Barba, B., Vallortigara, G., & Simion, F. (2019). A mental number line in human newborns. *Developmental science*, 22(6), e12801.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., & Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 488(7414), 101–108.

Downes, S. M. (2018). Evolutionary psychology. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/entries/evolutionary-psychology/>

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. (1988) *Explaining Behavior*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, 23(5), 714–747.

Dummett, M. (1993). *Origins of Analytical Philosophy*. Cambridge, MA: Harvard University Press.

Dupré, J. (2003). Making hay with straw men. *American Scientist*, 91.1.

Dupré, J. (2001). *Human Nature and the Limits on Science*. Oxford: Oxford University Press.

Duval, A. (2019). The representation selection problem: Why we should favor the geometric-module framework of spatial reorientation over the view-matching framework. *Cognition*, 192, 103985.

Dweck, C. (2006). *Mindset: The New Psychology of Success*. London: Random House.

Dweck, C. S. (2017). *Mindset: Changing the Way You Think to Fulfil Your Potential*. Revised Edition. Hachette UK.

Dwyer, S. (2007). How not to argue that morality isn't innate: Comments on Prinz. In W. Sinnott-Armstrong (ed.), *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, Vol. 1, pp. 407–418. Cambridge, MA: MIT Press.

Eagly, A. H., & Carli, L. L. (2007). *Through the Labyrinth: The Truth about How Women Become Leaders*. Boston, MA: Harvard Business Review Press.

Eagly, A. H., & Wood, W. (2011). Feminism and the evolution of sex differences and similarities. *Sex Roles*, 64(9–10), 758–767.

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46(2), 183–201.
Edge (2005) The science of gender and science: Pinker vs. Spelke. A debate. <http://www.edge.org/3rd_culture/debate05/debate05_index.html> accessed 23 February 2014.
Ehrlich, P. R. (2000). *Human Natures: Genes, Cultures, and the Human Prospect*. Washington, DC: Island Press.
Eibl-Eibesfeldt, I. (1989). *Human Ethology*. New York: Aldine de Gruyter.
Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968), 303–306.
Ekman, P. (1992). An argument for basic emotion. *Cognition and Emotion*, 6(169–200), 1–32.
Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3), 550–553.
Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128, 203–235.
Ellis, L. (2011). Evolutionary and neuroandrogenic theory and universal gender differences in cognition and behavior. *Sex Roles*, 64, 707–722.
Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.
Elman, J. L, Bates, E. A., Johnson, M. H., Karmiliff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
Elsner, C., & Wertz, A. E. (2019). The seeds of social learning: Infants exhibit more social looking for plants than other object types. *Cognition*, *183*, 244–255.
Etienne, A. S., Berlie, J., Georgakopoulos, J., & Maurer, R. (1998). Role of dead reckoning in navigation. In S. Healy (ed.), *Spatial Representations in Animals*, pp. 54–68. Oxford: Oxford University Press.
Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.
Evans, E. M. (2002). Beyond Scopes: Why creationism is here to stay. In K. Rosengren (ed.), *Imagining the Impossible Magical, Scientific, and Religious Thinking in Children*, pp. 305–333. Cambridge: Cambridge University Press.
Evans, V. (2014). *The Language Myth: Why Language is Not an Instinct*. Cambridge: Cambridge University Press.
Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14, 219–250.
Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4), 339–357.
Farah, M. J., & Rabinowitz, C. (2003). Genetic and environmental influences on the organisation of semantic memory in the brain: Is "living things" an innate category? *Cognitive Neuropsychology*, 20(3–6), 401–408.
Farah, M. J., Rabinowitz, C., Quinn, G. E., & Liu, G. T. (2000). Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology*, 17(1–3), 117–123.
Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14), 9602–9605.
Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 594–611.
Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science*, 13(2), 150–156.
Félix, S. B., Pandeirada, J. N., & Nairne, J. S. (2019). Adaptive memory: Longevity and learning intentionality of the animacy effect. *Journal of Cognitive Psychology*, 31(3), 251–260.
Ferguson, B., & Waxman, S. R. (2016). What the [beep]? Six-month-olds link novel communicative signals to meaning. *Cognition*, 146, 185–189.
Ferkin, M. H., Pierce, A. A., Sealand, R. O., & delBarco-Trillo, J. (2005). Meadow voles, *Microtus pennsylvanicus*, can distinguish more over-marks from fewer over-marks. *Animal Cognition*, 8(3), 182–189.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, 81(2), 472–479.
Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences*, 110(38), 15231–15235.
Fessler, D. M., & Holbrook, C. (2013a). Friends shrink foes: The presence of comrades decreases the envisioned physical formidability of an opponent. *Psychological Science*, 24(5), 797–802.
Fessler, D. M., & Holbrook, C. (2013b). Bound to lose: Physical incapacitation increases the conceptualized size of an antagonist in men. *PloS One*, 8(8), e71306.
Fessler, D. M., & Holbrook, C. (2014). Marching into battle: synchronized walking diminishes the conceptualized formidability of an antagonist in men. *Biology Letters*, 10(8), 20140592.
Fessler, D. M., Holbrook, C., & Dashoff, D. (2016). Dressed to kill? Visible markers of coalitional affiliation enhance conceptualized formidability. *Aggressive Behavior*, 42(3), 299–309.
Fessler, D. M., Holbrook, C., & Snyder, J. K. (2012). Weapons make the man (larger): Formidability is represented as size and strength in humans. *PloS One*, 7(4), e32751.
Fessler, D. M., & Navarrete, C. D. (2003). Meat is good to taboo. *Journal of Cognition and Culture*, 3(1), 1–40.
Filippetti, M. L., Johnson, M. H., Lloyd-Fox, S., Dragovic, D., & Farroni, T. (2013). Body perception in newborns. *Current Biology*, 23(23), 2413–2416.
Filippetti, M. L., Orioli, G., Johnson, M. H., & Farroni, T. (2015). Newborn body perception: Sensitivity to spatial congruency. *Infancy*, 20(4), 455–465.
Filippi, R., & Karmiloff-Smith, A. (2013). What can neurodevelopmental disorders teach us about typical development? In C. R. Marshall (ed.), *Current Issues in Developmental Disorders*, pp. 193–209. London: Psychology Press.
Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (eds.), *The Handbook of Social Psychology*, Vol. 2, pp. 367–411, 4th ed. New York: McGraw-Hill.
Fiske, S. T., Lin, M., & Neuberg, S. L. (2018). The continuum model: Ten years later. In *Social Cognition*, pp. 41–75. London: Routledge.
Fitzpatrick, S. (2008). Doing away with Morgan's Canon. *Mind & Language*, 23(2), 224–246.
Flegr, J. (2007). Effects of *Toxoplasma* on human behavior. *Schizophrenia Bulletin*, 33(3), 757–760.
Fodor, J. A. (1975). *The Language of Thought*. New York: Tomas Y. Crowell.
Fodor, J. A. (1980) Reply to Putnam. In M. Piattelli-Palmarini (ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, pp. 325–334. Cambridge, MA: Harvard University Press.
Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations: Philosophical Essays on the Foundations of Cognitive Science*, pp. 257–316. Cambridge, MA: MIT Press.
Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.
Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.
Fodor, J. A. (1990a). A theory of content, I: The problem. In his *A Theory of Content and Other Essays*, pp. 51–87. Cambridge, MA: MIT Press.
Fodor, J. A. (1990b). A theory of content, II: The theory. In his *A Theory of Content and Other Essays*, pp. 89–136. Cambridge, MA: MIT Press.
Fodor, J. A. (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.
Fodor, J. A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press.
Fodor, J. A. (2001). Review: *Evolution and the Human Mind: Modularity, Language and Meta-Cognition*, Peter Carruthers and Andrew Chamberlin (eds.). *British Journal for the Philosophy of Science*, 52(3), 623–628.
Fodor, J. A. (2007). Revenge of the given. In B. P. McLaughlin and J. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, pp. 99–109. Oxford University Press.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, J. A., Garrett, M., Walker, E., & Parkes, C. (1980). Against definitions. *Cognition*, 8, 263–367.

Fodor, J. A., & Pylyshyn, Z. W. (2015). *Minds without Meanings: An Essay on the Content of Concepts*. MIT Press.

Forgács, B., Tauzin, T., Gergely, G., & Gervain, J. (2022). The newborn brain is sensitive to the communicative function of language. *Scientific Reports*, 12(1), 1–6.

Franks, N., Mallon, E. B., Bray, H. E., Hamilton, M. J., & Mischler, T. C. (2003). Strategies for choosing between alternatives with different attributes: Exemplified by house-hunting ants. *Animal Behaviour*, 65, 215–223.

Frederick, D. A., Pillsworth, E. G., Galperin, A., Gildersleeve, K. A., Filossof, Y. R., Fales, M. R., Lopez, C. R., Lurye, F. S., Phuphanich, M. E., & Snider, J. B. (2009). Analyzing evolutionary social science and its popularizations—A review of *The Caveman Mystique: Pop-Darwinism and the Debates Over Sex, Violence, and Science. Evolution and Human Behavior*, 30(4), 301–304.

Frege, G. (1892) On sense and reference. In P. Geach and M. Black (eds.), *Translations form the Philosophical Writings of Gottlob Frege*, pp. 56–78, 3rd ed. Oxford: Blackwell.

Frisby, J. P., & Stone, J. V. (2010). *Seeing: The Computational Approach to Biological Vison*, 2nd ed. Cambridge, MA: MIT Press.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, 7(11), e1002355.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Gallistel, C. R., Brown, A. L., Carey, S., Gelman, R., & Keil, F. C. (1991). Lessons from animal learning for the study of cognitive development. In S. Carey and R. Gelman (eds.), *The Epigenesis of Mind: Essays in Biology and Cognition*, pp. 3–36. Hillsdale, NJ: LEA.

Gallistel, C. R., & Gibbon, J. (2002). *The Symbolic Foundations of Conditioned Behavior*. Hillsdale, NJ: LEA.

Gallistel, C. R. (2003). The principle of adaptive specialization as it applies to learning and memory. In R. H. Kluwe, G. Lüer, and F. Rösler (eds.), *Principles of Learning and Memory*, pp. 259–280. Birkhäuser, Basel.

Galton, A. (2011). Time flies but space does not: Limits to the spatialisation of time. *Journal of Pragmatics*, 43(3), 695–703.

Garcia, J., & Koelling, R. (1966). The relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4, 123–124.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., & Duchaine, B., 2009. Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia*, 47(1), 123–131.

Gazes, R. P., Hampton, R. R., & Lourenco, S. F. (2017). Transitive inference of social dominance by human infants. *Developmental Science*, 20(2), e12367.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2019). *Cognitive Neuroscience: The Biology of the Mind*, 5th ed. W. W. Norton & Company.

Geçkin, V., Thornton, R., & Crain, S. (2017). Children's interpretation of disjunction in negative sentences: A comparison of Turkish and German. *Language Acquisition*, 25(2), 197–212.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, 14, 79–106.

Gelman, S. A. (2003) *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford: Oxford University Press.

Gennari, G., Dehaene, S., Valera, C., & Dehaene-Lambertz, G. (2023). Spontaneous supra-modal encoding of number in the infant brain. *Current Biology*, 33(10), 1906–1915.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought*, pp. 195–235. Cambridge, MA: MIT Press.

Geraci, A., & Surian, L. (2023). Preverbal infants' reactions to third-party punishments and rewards delivered toward fair and unfair agents. *Journal of Experimental Child Psychology*, 226, 105574.

Ghazanfar, A. A., Neuhoff, J. G., & Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), 15755.

Ghrear, S., Baimel, A., Haddock, T., & Birch, S. A. (2021). Are the classic false belief tasks cursed? Young children are just as likely as older children to pass a false belief task when they are not required to overcome the curse of knowledge. *PloS One*, 16(2), e0244141.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., & Conway, B. R. (2017). Color naming across languages reflects color use. *PNAS*, 114, 10785–10790.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Gilovich, T., Keltner, D., Chen, S., & Nisbett, R. E. (2015). *Social Psychology*, 4th ed. W. W. Norton & Company.

Giurfa, M. (2003). Cognitive neuroethology: dissecting non-elemental learning in a honeybee brain. *Current Opinion in Neurobiology*, 13(6), 726–735.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of "sameness" and "difference" in an insect. *Nature*, 410(6831), 930–933.

Gleitman, H., Gross, J., & Reisberg, D. (2010). *Psychology*, 8th ed. W. W. Norton & Company.

Glenberg, A. M. (2015). Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology*, 69(2), 165–171.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., & Hout, M. C. (2016). The poverty of embodied cognition. *Psychonomic Bulletin & Review*, 23(4), 959–978.

Goldman, A. (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman, A., & Beddor, B. (2015) Reliabilist epistemology. *The Stanford Encyclopedia of Philosophy* (Winter 2015 edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2015/entries/reliabilism/>.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2018). Categorization and concepts. In J. T. Wixted (ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, Vol. 3: *Language and Thought*, pp. 275–317. Hoboken, NJ: Wiley.

Gomes, N., Silva, S., Silva, C. F., & Soares, S. C. (2017). Beware the serpent: The advantage of ecologically-relevant stimuli in accessing visual awareness. *Evolution and Human Behavior*, 38(2), 227–234.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. (1967). The epistemological argument. *Synthese*, 17, 23–8.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118(1), 110–119.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, 111(1), 3–32.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56, 544–549.

Gottlieb, G. (1997). *Synthesizing Nature-Nurture: Prenatal Roots of Instinctive Behavior*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gould, J., & Gould, C., (1995). *The Honey Bee*. New York: W.H. Freeman & Co.

Gould, S. J. (1978) Sociobiology: The art of storytelling. *New Scientist*, 80, 530–533.

Gould, S. J. (1997). Evolution: The pleasures of pluralism. *New York Review of Books*, June 26, pp. 47–52.

Gould, S. J. (2000). More things in heaven and earth. In H. Rose and S. Rose (eds.), *Alas, Poor Darwin: Arguments against Evolutionary Psychology*, pp. 101–126. London: Jonathan Cape.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences*, 205(1161), 581–598.

Gouteux, S., Thinus-Blanc, C., & Vauclair, J. (2001). Rhesus monkeys use geometric and nongeometric information during a reorientation task. *Journal of Experimental Psychology: General*, 130, 509–519.

Griffiths, P. E. (2002). What is innateness? *The Monist*, 85.1, 70–85.

Griffiths, P. E., & Gray, R. D. (2004). The developmental systems perspective: Organism-environment systems as units of development and evolution. In M. Pigliucci and K. Preston (eds.), *Phenotypic Integration: Studying the Ecology and Evolution of Complex Phenotypes*, pp. 409–431. New York: Oxford University Press.

Griffiths, P. E., & Machery, E. (2008). Innateness, canalization, and "biologicizing the mind". *Philosophical Psychology*, 21, 397–414.

Gross, H., Pahl, M., Si., A., Zhu, H., Tautz, J., & Zhang, S. (2009). Number-based visual generalisation in the honeybee. *PloS One*, 4(1), e4263.

Gruber, J. S. (1965). *Studies in Lexical Relations* (Doctoral dissertation, MIT).

Hagen, T., & Laeng, B. (2016). The change detection advantage for animals: An effect of ancestral priorities or progeny of experimental design? *I-Perception*, 7(3), 2041669516651366.

Hagen, T., & Laeng, B. (2017). Animals do not induce or reduce attentional blinking, but they are reported more accurately in a rapid serial visual presentation task. *I-Perception*, 8(5), 204166951773554–25.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, 21(2), 167–179.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, 44(5), 1457–1465.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213), 665–668.

Halperin, E., Russell, A. G., Trzesniewski, K. H., Gross, J. J., & Dweck, C. S. (2011). Promoting the Middle East peace process by changing beliefs about group malleability. *Science*, 333(6050), 1767–1769.

Hamlin, J. K. (2013). Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition*, 128(3), 451–474.

Hamlin, J. K. (2014). Context-dependent social evaluation in 4.5-month-old human infants: The role of domain-general versus domain-specific processes in the development of social evaluation. *Frontiers in Psychology*, 5, 614, pp. 1–10.

Hamlin, J. K. (2015a). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in psychology*, 5, 1563, pp. 1–9.

Hamlin, J. K. (2015b). Does the infant possess a moral concept? In E. Margolis and S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: Infants prefer those who harm dissimilar others. *Psychological Science*, 24(4), 589–594.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26(1), 30–39.

Hamlin, J. K., & Wynn, K. (2012). Who knows what's good to eat? Infants fail to match the food preferences of antisocial others. *Cognitive Development*, 27(3), 227–239.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557–559, pp. 557–560.

Hamlin, J. K., Wynn, K., & Bloom, P. (2012). Reply to Scarf et al.: Nuanced social evaluation: Association doesn't compute. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22), E1427–E1427.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences*, 108(50), 19931–19936.

Harris, P. L. (2012). *Trusting What You're Told: How Children Learn from Others*. Cambridge, MA: Harvard University Press.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, 69(1), 251–273.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3), 252–264.

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, 34(3), 479–490.

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J., & Davidson, R. J. (2010). Cosmetic use of botulinum toxin-A affects processing of emotional language. *Psychological Science*, 21(7), 895–900.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 27–30 June, 770–778.

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-year-olds succeed at a verbal anticipatory-looking false belief task. *British Journal of Developmental Psychology*, 30, 14–29.

Heck, R. G. (2000). Nonconceptual content and the space of reasons. *Philosophical Review*, 109(4), 483–523.

Held, R. (1979). Development of visual resolution. *Canadian Journal of Psychology*, 33, 213–221.

Hempel, C. G. (1935/1980). The logical analysis of psychology. N. Block (ed.), *Readings in Philosophy of Psychology*, *1*, (pp. 14–23). Cambridge, MA: Harvard University Press.

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, J., & Boyd, R. (1998) The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Humall Behavior*, 19(4), 215–242.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F. J., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–814.

Henrich, J., & Gil-White, F. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3), 165–196.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.

Hepach, R., Haberl, K., Lambert, S., & Tomasello, M. (2017). Toddlers help anonymously. *Infancy*, 22(1), 130–145.

Hermer, L., & Spelke, E. S. (1996). Modularity and development: The case of spatial reorientation. *Cognition*, 61(3), 195–232.

Hernik, M., Fearon, P., & Csibra, G. (2014). Action anticipation in human infants reveals assumptions about anteroposterior body-structure and action. *Proceedings of the Royal Society B: Biological Sciences*, 281(1781), 20133205.

Hespos, S. J., & Baillargeon, R. (2001a). Reasoning about containment events in very young infants. *Cognition*, 78(3), 207–245.

Hespos, S. J., & Baillargeon, R. (2001b). Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, 12(2), 141–147.

Hespos, S. J., Ferry, A. L., Anderson, E. M., Hollenbeck, E. N., & Rips, L. J. (2016). Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science*, 27(2), 244–256.

Hespos, S., Gredebäck, G., Von Hofsten, C., & Spelke, E. S. (2009). Occlusion is hard: Comparing predictive reaching for visible and hidden objects in infants and adults. *Cognitive Science*, 33(8), 1483–1502.

Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. *Nature*, 430(6998), 453–456.

Heyes, C. (2003). Four routes of cognitive evolution. *Psychological Review*, 110, 713–727.

Heyes, C. (2012). Simple minds: A qualified defence of associative learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603), 2695–2703.

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, 17(5), 647–659.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82–97.

Holbrook, C., & Fessler, D. M. T. (2013). Sizing up the threat: The envisioned physical formidability of terrorists tracks their leaders' failures and successes. *Cognition*, 127(1), 46–56.

Holbrook, C., Piazza, J., & Fessler, D. M. T. (2014). Conceptual and empirical challenges to the "authentic" versus "hubristic" model of pride. *Emotion*, 14(1), 17–32.

Hood, B., & Willatts, P. (1986). Reaching in the dark to an object's remembered position: Evidence for object permanence in 5-month-old infants. *British Journal of Developmental Psychology*, 4(1), 57–65.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, 155, 67–76.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., Hewlett, B. S., McElreath, R., & Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences*, 110, 14586–14591.

House, P. K., Vyas, A., & Sapolsky, R. (2011). Predator cat odors activate sexual arousal pathways in brains of *Toxoplasma gondii* infected rats. *PloS One*, 6(8), e23277.

<https://en.wikipedia.org/wiki/Hedge_fund>.

<https://en.wikipedia.org/wiki/Glossary_of_surfing>.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huguet, G., Ey, E., & Bourgeron, T. (2013). The genetic landscapes of autism spectrum disorders. *Annual Review of Genomics and Human Genetics*, 14(1), 191–213.

Hume, D. (1739/1978). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, revised P. H. Nidditch. Oxford: Oxford University Press.

Hume, D. (1748/1975) *An Enquiry Concerning Human Understanding*, ed. L. A. Selby-Bigge, revised P. H. Nidditch. Oxford: Clarendon Press.

Hupbach, A., & Nadel, L. (2005). Reorientation in a rhombic environment: No evidence for an encapsulated geometric module. *Cognitive Development*, 20, 279–302.

Hurford, J. R. (2001). Languages treat 1-4 specially. *Mind & Language*, 16(1), 69–75.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, 18(3), 231–257.

Hursthouse, R. (2013) Virtue ethics. *The Stanford Encyclopedia of Philosophy* (Fall 2013 edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/ethics-virtue/>.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., Mujica, F., Coates, A., & Ng, A. Y. (2015). An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*.

Hyde, D. C., Simon, C. E., Ting, F., & Nikolaeva, J. I. (2018). Functional organization of the temporal–parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience*, 38(18), 4264–4274.

Hym, C., Dumuids, M.V., Anderson, D.I., Forma, V., Provasi, J., Briè re-Dollat, C., Granjon, L., Gervain, J., Nazzi, T., & Barbu-Roth, M. (2022). Newborns modulate their crawling in

response to their native language but not another language. *Developmental Science*, 26(1), e13248.
Ichikawa, J. J., & Steup, M. (2017). The analysis of knowledge. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/entries/knowledge-analysis/>
Inzlicht, M., & Schmader, T. (2012). *Stereotype Threat: Theory, Process, and Application*. Oxford: Oxford University Press.
Izard, V., Pica, P., Spelke, E. S., & Dehaene, S. (2011). Flexible intuitions of Euclidean geometry in an Amazonian indigene group. *Proceedings of the National Academy of Sciences*, 108(24), 9782–9787.
Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 106(25), 10382–10385.
Izard, V., & Spelke, E. S. (2009). Development of sensitivity to geometry in visual forms. *Human Evolution*, 23(3), 213–248.
Izuma, K., Matsumoto, K., Camerer, C. F., & Adolphs, R. (2011). Insensitivity to social reputation in autism. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42), 17302–17307.
Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.
Jackendoff, R. (1994). *Patterns in the Mind: Language and Human Nature*. New York: Basic Books.
Jackson, R. E. (2009). Individual differences in distance perception. *Proceedings of the Royal Society B: Biological Sciences*, 276(1662), 1665–1669.
Jackson, R. E., & Cormack, L. K. (2007). Evolved navigation theory and the descent illusion. *Attention, Perception, & Psychophysics*, 69(3), 353–362.
Jackson, R. E., & Cormack, L. K. (2008). Evolved navigation theory and the environmental vertical illusion. *Evolution and Human Behavior*, 29(5), 299–304.
Jackson, R. E., & Cormack, L. K. (2010). Reducing the presence of navigation risk eliminates strong environmental illusions. *Journal of Vision*, 10(5), 9, 1–8.
Jackson, R. E., & de Garcia, J. G. (2017). Evolved navigation illusion provides universal human perception measure. *Journal of Vision*, 17(1), 39.
Jackson, R. E., & Willey, C. R. (2011). Evolved navigation theory and horizontal visual illusions. *Cognition*, 119(2), 288–294.
Jackson, R. E., & Willey, C. R. (2013). Evolved navigation theory and the plateau illusion. *Cognition*, 128(2), 119–126.
Jacob, P. (2020). What do false-belief tests show? *Review of Philosophy and Psychology*, 11(1), 1–20.
James, W. (1890). *The Principles of Psychology*, Vol. 1. Mineola, NY: Dover Publications Inc.
Järnefelt, E., Canfield, C. F., & Kelemen, D. (2015). The divided mind of a disbeliever: Intuitive beliefs about nature as purposefully created among different groups of non-religious adults. *Cognition*, 140(C), 72–88.
Jin, K. S., & Baillargeon, R. (2017). Infants possess an abstract expectation of ingroup support. *Proceedings of the National Academy of Sciences*, 114(31), 8199–8204.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2), 201–211.
Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, 16(3), 175–179.
Johnson, M. H., Bolhuis, J. J., & Horn, G. (1985). Interaction between acquired preferences and developing predispositions during imprinting. *Animal Behaviour*, 13, 1000–1006.
Johnson, S. C. (2005). Reasoning about intentionality in preverbal infants. In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind: Structure and Contents*, pp. 139–170. Oxford: Oxford University Press.
Johnson, S. P., & Aslin, R. N. (1995). Perception of object unity in 2-month-old infants. *Developmental Psychology*, 31(5), 739–745.

Johnsson, J. I. (1997). Individual recognition affects aggression and dominance relations in rainbow trout, *Oncorhynchus Mykiss: Ethology*, 103(4), 267–282.
Jordan, H., Reiss, J. E., Hoffman, J. E., & Landau, B. (2002). Intact perception of biological motion in the face of profound spatial deficits: Williams syndrome. *Psychological Science*, 13(2), 162–167.
Joyce, R. (2013). The many moral nativisms. In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser (eds.), *Cooperation and Its Evolution*, pp. 549–572. Cambridge, MA: MIT Press.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Macmillan.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kahneman, D., Slovic, P., & Tversky, A. (2000). *Choices, Values, Frames*. Cambridge: Cambridge University Press.
Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.
Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, 116(42), 20904–20909.
Kant, I. (1781/1998). *Critique of Pure Reason*, ed. and trans. P. Guyer and A. W. Wood. Cambridge: Cambridge University Press.
Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25), 11163–11170.
Karavanich, C., & Atema, J. (1998). Individual recognition and memory in lobster dominance. *Animal Behaviour*, 56(6), 1553–1560.
Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.
Karmiloff-Smith, A. (1997). Crucial differences between developmental cognitive neuroscience and adult neuropsychology. *Developmental Neuropsychology*, 13(4), 513–524.
Karmiloff-Smith, A. (1998) Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, 2(10), 389–398.
Karmiloff-Smith, A. (2000). Why babies' brains are not Swiss army knives. In H. Rose and S. Rose (eds.), *Alas, Poor Darwin: Arguments against Evolutionary Psychology*, pp. 144–156. London: Vintage.
Karmiloff-Smith, A. (2009a). Nativism versus neuroconstructivism: Rethinking the study of developmental disorders. *Developmental Psychology*, 45(1), 56–63.
Karmiloff-Smith, A. (2009b). Preaching to the converted? From constructivism to neuroconstructivism. *Child Development Perspectives*, 3, 99–102.
Karmiloff-Smith, A., Scerif, G., & Ansari, D. (2003). Double dissociations in developmental disorders? Theoretically misconceived, empirically dubious. *Cortex*, 39(1), 161–163.
Karmiloff-Smith, A., Thomas, M., Annaz, D., Humphreys, K., Ewing, S., Brace, N., Van Duuren, M. Pike, G., Grice, S., & Campbell, R. (2004). Exploring the Williams syndrome face-processing debate: the importance of building developmental trajectories. *Journal of Child Psychology and Psychiatry*, 45(7), 1258–1274.
Karmon, D., Zoran, D., & Goldberg, Y. (2018). LaVAN: Localized and Visible Adversarial Noise. *arXiv*, 1801, 02608v2.
Karraker, K. H., Vogel, D. A., & Lake, M. A. (1995). Parents' gender-stereotyped perceptions of newborns: The eye of the beholder revisited. *Sex Roles*, 33(9–10), 687–701.
Keeley, L. H. (1996). *War before Civilization*. Oxford: Oxford University Press.
Keil, F. C. (1989) *Concepts, Kinds and Cognitive Development*. Cambridge, MA: MIT Press.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnar and M. Maratos (eds.), *The Minnesota Symposia on Child Psychology*, Vol. 25: *Modularity and Constraints in Language and Cognition*, pp. 103–137. Mahwah, NJ: Erblaum.

Keil, F. C. (1999). Nativism. In R. A. Wilson and F. C. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, pp. 583–586. Cambridge, MA: MIT Press.

Kelemen, D. (1999a). The scope of teleological thinking in preschool children. *Cognition*, 70(3), 241–272.

Kelemen, D. (1999b). Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Developmental Psychology*, 35, 1440–1453.

Kelemen, D. (2003). British and American children's preferences for teleo-functional explanations of the natural world. *Cognition*, 88(2), 201–221.

Kelemen, D. (2011). Teleological minds: How natural intuitions about agency and purpose influence learning about evolution. In K. S. Rosengren, S.K. Brem, E. M. Evans, and G. M. Sinatra (eds), *Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution*, pp. 66–92. Oxford: Oxford University Press.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology General*, 142(4), 1074–1083.

Keller, E. F. (2010). *The Mirage of a Space between Nature and Nurture*. Durham, NC: Duke University Press.

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15(4), 483–524.

Kelly, D. (2011). *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: MIT Press.

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, 43(2), 133–160.

Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What basic emotion theory really says for the twenty-first century study of emotion. *Journal of Nonverbal Behavior*, 43(2), 195–201.

Khalidi, N. (2007). Innate cognitive capacities. *Mind and Language*, 22, 92–115.

Kim, E. Y., & Song, H. J. (2015). Six-month-olds actively predict others' goal-directed actions. *Cognitive Development*, 33, 1–13.

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. In C. von Hofsten and K. Rosander (eds.), *Progress in Brain Research*, Vol. 164, pp. 257–264. London: Elsevier.

Kirsh, D. (1991) Today the earwig, tomorrow man? *Artificial Intelligence*, 47, 161–184.

Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences*, 38(2), 1–70.

Knudsen, B., & Liszkowski, U. (2012a). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science*, 15(1), 113–122.

Knudsen, B., & Liszkowski, U. (2012b). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, 17(6), 672–691.

Kobylkov, D., Rosa-Salva, O., Zanon, M., & Vallortigara, G. (2024). Innate face detectors in the nidopallium of young domestic chicks. *bioRxiv*.

Kobayashi, T., Hiraki, K., Mugitani, R., & Hasegawa, T. (2004). Baby arithmetic: One object plus one tone. *Cognition*, 91(2), B23–B34.

Kondo, N., Izawa, E.-I., & Watanabe, S. (2012). Crows cross-modally recognize group members but not nongroup members. *Proceedings of the Royal Society, B: Biological Sciences*, 279, 1937–1942.

Kornblith, H. (1993). *Inductive Inference and Its Natural Ground: An Essay on Naturalistic Epistemology*. Cambridge, MA: MIT Press.

Kornblith, H. (2002). *Knowledge and Its Place in Nature*. New York: Oxford University Press.

Kornblith, H. (2007) How to refer to artifacts. In E. Margolis and S. Laurence (eds.), *Creations of the Mind: Essays on Artifacts and Their Representation*, pp. 138–149. Oxford: Oxford University Press.

Kosakowski, H., Cohen, M., Takahashi, A., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, 32, 265–274.

Koster-Hale, J., & Saxe, R. (2013) Functional neuroimaging of theory of mind. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (eds.), *Understanding Other Minds*, pp. 132–163, 3rd ed. Oxford: Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L.L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14), 5648–5653.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830–1834.

Kripke, S. (1972/80) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012/2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. Originally published in the *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1097–1105. Lake Tahoe, NV, Dec. 2012.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308), 110–114.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, 10(4), e1343771.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, 14(5), 402–408.

Kupfer, T. R., & Fessler, D. M. (2018). Ectoparasite defence in humans: Relationships to pathogen avoidance and clinical implications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1751), 20170207.

Kurzban, R. (2002) Alas poor evolutionary psychology: Unfairly accused, unjustly condemned. *The Human Nature Review*, 2, 99–109.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, 98(26), 15387–15392.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, e253, 1–72.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

Lakusta, L., Dessalegn, B., & Landau, B. (2010). Impaired geometric reorientation caused by genetic defect. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), 2813–2817.

Laland, K. N., & Brown, G. (2002). *Sense and Nonsense: Evolutionary Perspectives on Human Behaviour*. Oxford: Oxford University Press.

Landau, B. (2009). The importance of the nativist–empiricist debate: Thinking about primitives without primitive thinking. *Child Development Perspectives*, 3(2), 88–90.

Landau, B., & Gleitman, L. R. (1985). *Language and Experience: Evidence from the Blind Child*. Harvard University Press.

Landau, B., & Hoffman, J.E. (2012). *Spatial Representation: From Gene to Mind*. New York: Oxford University Press.

Langston, R. F., Ainge, J. A., Couey, J. J., Canto, C. B., Bjerknes, T. L., Witter, M. P., Moser, E. I., & Moser, M. B. (2010). Development of the spatial representation system in the rat. *Science*, 328(5985), 1576–1580.

Larsen, C. C., Bonde Larsen, K., Bogdanovic N., Laursen, H., Græm, N., Badsberg Samuelsen, G., & Pakkenberg, B. (2006). Total number of cells in the human newborn telencephalic wall. *Neuroscience*, 13, 999–1003.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis and S. Laurence (eds.), *Concepts: Core Readings*, pp. 3–81. Cambridge, MA: MIT Press.
Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science*, 52, 217–276.
Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, 86(1), 25–55.
Laurence, S., & Margolis, E. (2005). Number and natural language. In P. Carruthers, S. Laurence, & S. Stich (eds.), *The Innate Mind: Structure and Contents*, pp. 216–235. Oxford: Oxford University Press.
Laurence, S., & Margolis, E. (2007). Linguistic determinism and the innate basis of number. In P. Carruthers, S. Laurence, & S. Stich (eds.), *The Innate Mind: Foundations and the Future*, pp. 139–170. Oxford: Oxford University Press.
Laurence, S., & Margolis, E. (2012a). The scope of the conceptual. In E. Margolis, R. Samuels, and S. Stich (eds.), *The Oxford Handbook of Philosophy of Cognitive Science*, pp. 291–317. New York: Oxford University Press.
Laurence, S., & Margolis, E. (2012b). Abstraction and the origin of general ideas. *Philosophers' Imprint*, 12(19), 1–22.
Laurence, S., & Margolis, E. (2015). Concept nativism and neural plasticity. In E. Margolis & S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts*, pp. 117–147. Cambridge, MA: MIT Press.
Lea, S. E., Slater, A. M., & Ryan, C. M. (1996). Perception of object unity in chicks: A comparison with the human infant. *Infant Behavior and Development*, 19(4), 501–504.
Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105(2), 395–438.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
Leding, J. K. (2019). Adaptive memory: Animacy, threat, and attention in free recall. *Memory & Cognition*, 47(3), 383–394.
Lee, S. A., Sovrano, V. A., & Spelke, E. S. (2012). Navigation as a source of geometric knowledge: Young children's use of length, angle, distance, and direction in a reorientation task. *Cognition*, 123, 144–161.
Lee, S. A., & Spelke, E. S. (2008). Children's use of geometry for reorientation. *Developmental Science*, 11(5), 743–749.
Lee, S. A., & Spelke, E. S. (2011). Young children reorient by computing layout geometry, not by matching images of the environment. *Psychonomic Bulletin & Review*, 18(1), 192–198.
Lee, S. A., & Vallortigara, G. (2015). Bumblebees spontaneously map location of conspecific using geometry and features. *Learning and Motivation*, 50, 32–38.
Lehrman, D. S. (1953). A critique of Konrad Lorenz's theory of instinctive behavior. *Quarterly Review of Biology*, 28(4), 337–363.
Lehrman, D. S. (1970). Semantic and conceptual issues in the nature–nurture problem. In L. R. Aronson, D. S. Lehrman, E. Tobach, & J. S. Rosenblatt (eds.), *Development and Evolution of Behavior*, pp. 17–52. San Francisco: W. H. Freeman.
Leibniz, G. W. (1705/1996) *New Essays on Human Understanding*, ed. and trans. P. Remnant & J. Bennett. Cambridge: Cambridge University Press.
Leibniz, G. W. (1714/1965). *Monadology*. In P. Schrecker & A. M. Schrecker (trans. and eds.) *Monadology, and Other Philosophical Essays*, pp. 215–277. New York: Bobbs-Merrill Co.
Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, 40, 1–62.
Lerner, R. M. (2015) Preface. In W. F. Overton, P. C. M. Molenaar, & R. M. Lerner (eds.), *Handbook of Child Psychology and Developmental Science: Theory and Method*, pp. xv–xxi, 7th ed. Hoboken, NJ: John Wiley & Sons, Inc.
Leo, I., Angeli, V., Lunghi, M., Dalla Barba, B., & Simion, F. (2018). Newborns' face recognition: The role of facial movement. *Infancy*, 23(1), 45–60.
Leo, I., & Simion, F. (2009). Face processing at birth: A Thatcher illusion study. *Developmental Science*, 12(3), 492–498.
Leslie, A. M. (1982). The perception of causality in infants. *Perception*, 11(2), 173–186.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review*, 94(4), 412–426.
Leslie, A. M. (1988). The necessity of illusion: Perception and thought in infancy. In L. Weiskrantz (ed.), *Thought without Language*, pp. 185–210. Oxford: Oxford Science Publications.
Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld, & S. A. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 119–148. New York: Cambridge University Press.
Leslie, A. M. (2004), Who's for learning? *Developmental Science*, 7(4), 417–419.
Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265–288.
Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, 43(3), 225–251.
Leslie, A. M., Xu, F., Tremoulet, P., & Scholl, B. J. (1998). Indexing and the object concept: Developing what and where systems indexing. Trends in Cognitive Sciences, 2(1), 10–18.
Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47, 1940–1951.
Levin, B., & Pinker, S. (1991) Introduction. In B. Levin & S. Pinker (eds.), *Lexical and Conceptual Semantics*, pp. 1–7. Oxford: Blackwell.
Levinson, S. (2003) Language and mind: Let's get the issues straight. In D. Gentner and S. Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought*, pp. 25–46. Cambridge, MA: MIT Press.
Lewis, J. D., & Elman, J. L. (2001) A connectionist investigation of linguistic arguments from poverty of the stimulus: Learning the unlearnable. In J. Moore & K. Stenning (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp. 552–557. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Lewkowicz, D. J. (2011). The biological implausibility of the nature-nurture dichotomy and what it means for the study of infancy. *Infancy*, 16(4), 331–367.
Li, Z., Liu, J., Zheng, M., & Xu, X. S. (2014). Encoding of both analog- and digital-like behavioral outputs by one C. elegans interneuron. *Cell*, 159(4), 751–765.
Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). Preverbal infants infer third-party social relationships based on language. *Cognitive science*, 41(S3), 622–634.
Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental science*, 13(6), 900–906.
Libertus, M. E., Starr, A., & Brannon, E. M. (2014). Number trumps area for 7-month-old infants. *Developmental Psychology*, 50, 108–112.
Lieberman, D., Oum, R., & Kurzban, R. (2008). The family of fundamental social categories includes kinship: Evidence from the memory confusion paradigm. *European Journal of Social Psychology*, 38(6), 998–1012.
Lima, S. L. (1984). Downy woodpecker foraging behavior: Efficient sampling in simple stochastic environments. *Ecology*, 65, 166–174.
Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. In O. Houdé & G. Borst (eds.), *Cambridge Handbook of Cognitive Development*, pp. 168–194. Cambridge: Cambridge University Press.
Lindová, J., Novotná, M., Havlicek, J., Jozífková, E., Skallová, A., Kolbeková, P., Hodný, Z., Kodym, P., & Flegr, J. (2006). Gender differences in behavioural changes induced by latent toxoplasmosis. *International Journal for Parasitology*, 36(14), 1485–1492.
Lindsey, D. T., & Brown, A. M. (2021). Lexical color categories. *Annual Review of Vision Science*, (7), 605–631.
Lingnau, A., Strnad, L., He, C., Fabbri, S., Han, Z., Bi, Y., & Caramazza, A. (2014). Cross-modal plasticity preserves functional specialization in posterior parietal cortex. *Cerebral Cortex*, 24, 541–549.
Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological science*, 14(5), 396–401.

Lloyd, E. A. (1999). Evolutionary psychology: The burdens of proof. *Biology and Philosophy*, 14, 211–233.
Locke, J. (1690/1975). *An Essay Concerning Human Understanding*, ed. P. H. Nidditch. Oxford: Oxford University Press.
Loomis, J. M., Klatzky, R. L., Golledge, R. G., & Philbeck, J. W. (1999). Human navigation by path integration. In R. Golledge (ed.), *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*, pp. 125–151. Baltimore: Johns Hopkins University Press.
Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent. *Developmental Science*, 14(2), 453–460.
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science*, 16(8), 601–608.
Luo, Y., Baillargeon, R., Brueckner, L., & Munakata, Y. (2003). Reasoning about a hidden object after a delay: Evidence for robust representations in 5-month-old infants. *Cognition*, 88(3), 23–32.
Luo, Y., & Choi, Y. J. (2012). Infants attribute to agents goals and dispositions. *Developmental Science*, 15(5), 727–728.
Lyon, B. E. (2003). Egg recognition and counting reduce costs of avian conspecific brood parasitism. *Nature*, 422(6931), 495–499.
Machery, E. (2011). Developmental disorders and cognitive architecture. In P. R. Adriaens, & A. De Block (eds.), *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory*, pp. 91–116. Oxford: Oxford University Press.
Mackie, J. L. (1976). *Problems from Locke*. Oxford: Oxford University Press.
Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, 63(3), 397–405.
Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology*, 60(1), 27–51.
Mahon, B. Z., Schwarzbach, J., & Caramazza, A. (2010). The representation of tools in left parietal cortex is independent of visual experience. *Psychological Science*, 21(6), 764–771.
Mallon, R. (2016). *The Construction of Human Kinds*. Oxford: Oxford University Press.
Mallon, R., & Weinberg, J. (2006). Innateness as closed process invariance. *Philosophy of Science*, 73, 323–344.
Mameli, M. (2008). On innateness: The clutter hypothesis and the cluster hypothesis. *Journal of Philosophy*, CV, 719–736.
Mameli, M., & Bateson, P. (2006). Innateness and the sciences. *Biology and Philosophy*, 21, 155–188.
Mandler, J. M. (1988). How to build a baby: On the development of an accessible representational system. *Cognitive Development*, 3(2), 113–136.
Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 99(4), 587.
Mandler, J. M. (2004). *The Foundations of Mind: Origins of Conceptual Thought*. Oxford: Oxford University Press.
Mandler, J. M. (2008a). On the birth and growth of concepts. *Philosophical Psychology*, 21(2), 207–230.
Mandler, J. M. (2008b). Infant concepts revisited. *Philosophical Psychology*, 21(2), 269–280.
Mandler, J. M. (2012). On the spatial foundations of the conceptual system and its enrichment. *Cognitive Science*, 36(3), 421–451.
Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, 23(2), 263–298.
Mandler, J. M., & Cánovas, C. P. (2014). On defining image schemas. *Language and Cognition*, 6(4), 510–532.
Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, 8(3), 291–318.
Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G. F. (2004). *The Birth of the Mind: How a Tiny Number of Genes Creates the Complexities of Human Thought*. New York: Basic Books.
Marcus, G. F. (2009). Misrepresentational innateness. *Child Development Perspectives*, 3(2), 94–95.
Marcus, G. F. (2018). Innateness, AlphaZero, and artificial intelligence. ArXiv Preprint. ArXiv:1801.05667.
Marcus, G. F., & Keil, F. C. (2008). Concepts, correlations, and some challenges for connectionist cognition. *Behavioral and Brain Sciences*, 31(6), 722–723.
Margolis, E. (1998). How to acquire a concept. *Mind & Language*, 13(3), 347–369.
Margolis, E. (2017). Infants, animals, and the origins of number. *Behavioral and Brain Sciences*, 40, 31.
Margolis, E. (2020). The small number system. *Philosophy of Science*, 87, 1–22.
Margolis, E., & Laurence, S. (1999). *Concepts: Core Readings*. Cambridge, MA: MIT Press.
Margolis, E., & Laurence, S. (2007a). The ontology of concepts—abstract objects or mental representations? *Noûs*, 41(4), 561–593.
Margolis, E., & Laurence, S. (2007b). *Creations of the Mind: Theories of Artifacts and Their Representation*. Oxford: Oxford University Press.
Margolis, E., & Laurence, S. (2008). How to learn the natural numbers: Inductive inference and the acquisition of number concepts. *Cognition*, 106(2), 924–939.
Margolis, E., & Laurence, S. (2011). Learning matters: The role of learning in concept acquisition. *Mind & Language*, 26(5), 507–639.
Margolis, E., & Laurence, S. (2013). In defense of nativism. *Philosophical Studies*, 165, 693–718.
Margolis, E., & Laurence, S. (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press.
Margolis, E., & Laurence, S. (2019). Concepts. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/entries/concepts/>.
Margolis, E., & Laurence, S. (2023). Making sense of domain specificity. *Cognition*, 240, 105583.
Margolis, J., & Fisher, A. (2002) *Unlocking the Clubhouse: Women in Computing*. Cambridge, MA: MIT Press.
Margoni, F., Baillargeon, R., & Surian, L. (2018). Infants distinguish between leaders and bullies. *Proceedings of the National Academy of Sciences of the United States of America*, 115(38), E8835–E8843.
Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology*, 54(8), 1445–1455.
Margoni, F., Surian, L., & Baillargeon, R. (2022). The violation-of-expectation paradigm: A conceptual overview. <https://doi.org/10.31234/osf.io/5fsxj>.
Marno, H., Farroni, T., Dos Santos, Y. V., Ekramnia, M., Nespor, M., & Mehler, J. (2015). Can you see what I am talking about? Human speech triggers referential expectation in four-month-old infants. *Scientific Reports*, 5(1), 13594.
Martinho, A., & Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different". *Science*, 353(6296), 286–288.
Martins, Y., Pelchat, M. L., & Pliner, P. (1997). "Try it; it's good and it's good for you": Effects of taste and nutrition information on willingness to try novel foods. *Appetite*, 28(2), 89–102.
Mascalzoni, E., Regolin, L., & Vallortigara, G. (2009). Mom's shadow: Structure-from-motion in newly hatched chicks as revealed by an imprinting procedure. *Animal Cognition*, 12(2), 389–400.
Mascalzoni, E., Regolin, L., & Vallortigara, G. (2010). Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences of the United States of America*, 107(9), 4483–4485.
Mascalzoni, E., Regolin, L., Vallortigara, G., & Simion, F. (2013). The cradle of causal reasoning: Newborns' preference for physical causality. *Developmental Science*, 16(3), 327–335.
Mascaro, O., & Csibra, G. (2012). Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences*, 109(18), 6862–6867.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition*, 78(1), 1–26.

Mattes, R.D. (1991). Learned food aversions: A family study. *Physiology & Behavior* 50(3), 499–504.

Mayer, U., Rosa-Salva, O., & Vallortigara, G. (2017). First exposure to an alive conspecific activates septal and amygdaloid nuclei in visually-naïve domestic chicks (Gallus gallus). *Behavioural Brain Research*, 317, 71–81.

Mayer, A., & Träuble, B. E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International Journal of Behavioral Development*, 37(1), 21–28.

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development*, 82(4), 1224–1237.

McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science*, 268(5210), 569–573.

McBrearty, S., & Brooks, A. S. (2000). The revolution that wasn't: A new interpretation of the origin of modern human behavior. *Journal of Human Evolution*, 39(5), 453–563.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2: *Psychological and Biological Models*. Cambridge, MA: MIT Press.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, 15(11), 776–781.

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology*, 39(1), 83–90.

Mechner, F., & Guevrekian, L. (1962). Effects of deprivation upon counting and timing in rats. *Journal of the Experimental Analysis of Behavior*, 5, 463–466.

Medin, D. L., & Atran, S. (1999). *Folkbiology*. MIT Press.

Medin, D. L., & Rips, L. J. (2005). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*, pp. 37–81. Cambridge: Cambridge University Press.

Mehr, S. A., Song, L. A., & Spelke, E. S. (2016). For 5-month-old infants, melodies are social. *Psychological science*, 27(4), 486–501.

Meinhardt, M. J., Bell, R., Buchner, A., & Röer, J. P. (2019). Adaptive memory: Is the animacy effect on memory due to richness of encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3), 416–426.

Meristo, M., Strid, K., & Surian, L. (2016). Preverbal infants' ability to encode the outcome of distributive actions. *Infancy*, 21(3), 353–372.

Meristo, M., & Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition*, 129(1), 102–113.

Michotte, A. (1946/1963). *The Perception of Causality*, trans. T. R. Miles & E. Miles. New York: Basic Books.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4), 143–152.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.

Mill, J. S. (1882). *A System of Logic, Ratiocinative and Inductive*, 8th ed. London: Longmans, Green, and Co.

Millikan, R. (1984) *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Nazi, A., Pak, J., Tong, A., Srinivasa, K., Hang, W., Tuncer, E., Le, Q. V., Laudon, J., Ho, R., Carpenter, R., & Dean, J. (2021). A graph placement methodology for fast chip design. *Nature*, 594(7862), 207–212.

Mix, K. S., Levine, S. C., & Newcombe, N. S. (2016). Development of quantitative thinking across correlated dimensions. In A. Henik (ed.), *Continuous Issues in Numerical Cognition: How Many or How Much*, pp. 3–35. Cambridge, MA: Academic Press.

Mondschein, E. R., Adolph, K. E., & Tamis-LeMonda, C. S. (2000). Gender bias in mothers' expectations about infant crawling. *Journal of Experimental Child Psychology*, 77(4), 304–316.

Moon, C., Cooper, R., & Fifer, W. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, 16, 495–500.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Paediatrica*, 102(2), 156–160.

Moore, D. S. (2001). *The Dependent Gene: The Fallacy of "Nature vs. Nurture"*. New York: Henry Holt and Company.

Moore, D. S. (2009). Probing predispositions: The pragmatism of a process perspective. *Child Development Perspectives*, 3, 91–93.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–26 July 2017, Honolulu, HI, pp. 1765–1773.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), 2688–2692.

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., Ibáñez, A., García, A., Gianelli, C., Jones, J. L., Madden, J., Alifano, F., Bergen, B., Bloxsom, N. G., Bub, D. N., Cai, Z. G., Chartier, C. R., Chatterjee, A., Conwell, E., Cook, S. W., Davis, J. D., Evers, E. R. K., Girard, S., Harter, D., Hartung, F., Herrera, E., Huettig, F., Humphries, S., Juanchich, M., Kühne, K., Lu, S., Lynes, T., Masson, M. E. J., Ostarek, M., Pessers, S., Reglin, R., Steegen, S., Thiessen, E., D., Thomas, L. E., Trott, S., Vandekerckhove, J., Vanpaemel, W., Vlachou, M., Williams, K., & Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*, 29(2), 613–626.

Morgan, M. H., & Carrier, D. R. (2013). Protective buttressing of the human fist and the evolution of hominin hands. *Journal of Experimental Biology*, 216(2), 236–244.

Morris, K. V., & Mattick, J. S. (2014). The rise of regulatory RNA. *Nature Reviews: Genetics*, 15(6), 423–437.

Müller M., & Wehner, R. (2010). Path integration provides a scaffold for landmark learning in desert ants. *Current Biology*, 20, 1368–1371.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, 104(4), 686.

Murphy, G. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murty, N. A. R., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences*, 117(37), 23011–23020.

Nairne, J. S. (2022). Adaptive education: Learning and remembering with a stone-age brain. *Educational Psychology Review*, 34, 2275–2296.

Nairne, J. S., & Pandeirada, J. N. (2008). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science*, 17(4), 239–243.

Neander, K. (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Neary, K. R., & Friedman, O. (2014). Young children give priority to ownership when judging who should use an object. *Child Development*, 85(1), 326–337.

Neuhoff, J. G. (1998). Perceptual bias for rising tones. *Nature*, 395(6698), 123–123.
Neuhoff, J. G., Long, K. L., & Worthington, R. C. (2012). Strength and physical fitness predict the perception of looming sounds. *Evolution and Human Behavior*, 33(4), 318–322.
Neumeyer, C. (1998). Comparative aspects of colour constancy. In V. Walsh & J. Kulikowski (eds.), *Perceptual Constancy*, pp. 323–351. Cambridge: Cambridge University Press.
New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America*, 104(42), 16598–16603.
New, J. J., & German, T. C. (2015). Spiders at the cocktail party: an ancestral threat that surmounts inattentional blindness. *Evolution and Human Behavior*, 36(3), 165–173.
Newcombe, N. (2002). The nativist-empiricist controversy in the context of recent research on spatial and quantitative development. *Psychological Science*, 13, 395–401.
Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151–166.
Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.
Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.
Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
Nichols, S. (2021). *Rational Rules: Towards a Theory of Moral Learning*. Oxford: Oxford University Press.
Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H.-Y. (2016). Rational learners and moral rules. *Mind & Language*, 315, 530–554.
Nichols, S., & Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.
Niedenthal, P. M. (2008). Emotion concepts. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (eds.), *Handbook of Emotions*, Vol. 3, pp. 587–600. New York: The Guildford Press.
Niedenthal, P. M., Winkielman, P., Mondillon, L., & Vermeulen, N. (2009). Embodiment of emotion concepts. *Journal of Personality and Social Psychology*, 96(6), 1120.
Niedenthal, P. M., Wood, A., & Rychlowska, M. (2014). Embodied emotion concepts. In L. Shaprio (ed.), *The Roultledge Handbook of Embodied Cognition*, pp. 240–249. New York: Routledge University Press.
Nisbett, R. E. (2009). *Intelligence and How to Get It: Why Schools and Cultures Count*. WW Norton & Company.
Novack, M. A., Brentari, D., Goldin-Meadow, S., & Waxman, S. (2021). Sign language, like spoken language, promotes object categorization in young hearing infants. *Cognition*, 215, 104845.
Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H. Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sovi, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., & Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44–53.

Oakes, L. M., & Luck, S. J. (2014). Short-term memory in infancy. *The Wiley Handbook on the Development of Children's Memory*, Vols. I/II, 151–180. Hoboken, NJ: Wiley.
Oaten, M., Stevenson, R. J., & Case, T. I. (2009). Disgust as a disease-avoidance mechanism. *Psychological Bulletin*, 135, 303–321.
O'Hearn, K., Roth, J. K., Courtney, S. M., Luna, B., Street, W. Terwillinger, R., & Landau, B. (2011). Object recognition in Williams syndrome: Uneven ventral stream activation. *Developmental Science*, 14(3), 549–565.
Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology General*, 130(3), 466.
Olmstead, M. C., & Kuhlmeier, V. A. (2015). *Comparative Cognition*. Cambridge: Cambridge University Press.
Olsson, P., Wilby, D., & Kelber, A. (2016). Quantitative studies of animal colour constancy: Using the chicken as model. *Proceedings of the Royal Society B: Biological Sciences*, 283(1830), 20160411.
O'Neill, E. (2015). Relativizing innateness: Innateness as the insensitivity of the appearance of a trait with respect to specified environmental variation. *Biology & Philosophy*, 30(2), 211–225.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255–258.
O'Reilly, R. C., & McClelland, J. L. (1992). The self-organization of spatially invariant representations. Tech. Rep. No. PDP.CNS.92.5, Carnegie Mellon University.
Orioli, G., Bremner, A. J., & Farroni, T. (2018). Multisensory perception of looming and receding objects in human newborns. *Current Biology*, 28(22), R1294–R1295.
Overton, W. F. (2006). Developmental psychology: Philosophy, concepts, methodology. In R. M. Lerner (ed.), *Handbook of Child Psychology: Theoretical Models of Human Development*, Vol. 1, pp. 18–88, 6th ed. Hoboken, NJ: John Wiley & Sons Inc.
Oyama, S. (2000). *The Ontogeny of Information*. Durham, NC: Duke University Press.
Pagliarini, E., Crain, S., & Guasti, M. T. (2018). The compositionality of logical connectives in child Italian. *Journal of Psycholinguistic Research*, 47(1), 1–35.
Pagliarini, E., Lungu, O., van Hout, A., Pintér, L., Surányi, B., Crain, S., & Guasti, M.T. (2022). How adults and children interpret disjunction under negation in Dutch, French, Hungarian and Italian: A cross-linguistic comparison, *Language Learning and Development*, 18(1), 97–122.
Pagliarini, E., Reyes, M. A., Guasti, M. T., Crain, S., & Gavarró, A. (2021). Negative sentences with disjunction in child Catalan. *Language Acquisition*, 28(2), 153–165.
Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, 51(4), 550–572.
Papeo, L., Hochmann, J. R., & Battelli, L. (2016). The default computation of negated meanings. *Journal of Cognitive Neuroscience*, 28(12), 1980–1986.
Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research*, 134, 427–445.
Pauen, S. (2002). The global-to-basic level shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioral Development*, 26(6), 492–499.
Paulus, M., & Sabbagh, M. A. (2018). Special Issue on: Understanding theory of mind in infancy and toddlerhood. *Cognitive Development*, 46, 1–124
Peacocke, C. (1992). *A Study of Concepts*. MIT Press.
Peacocke, C. (2001). Does perception have a nonconceptual content? *Journal of Philosophy*, 98(5), 239–264.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Peelen, M. V., Bracci, S., Lu, X., He, C., Caramazza, A., & Bi, Y. (2013). Tool selectivity in left occipitotemporal cortex develops without vision. *Journal of Cognitive Neuroscience*, 25(8), 1225–1234.
Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120(3), 302–321.
Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214–216.

Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition*, 153, 175–181.
Peterson, C. C., Peterson, J. L., & Webb, J. (2000). Factors influencing the development of a theory of mind in blind children. *British Journal of Developmental Psychology*, 18(3), 431–447.
Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O-Toole, A. J. (2018). Face recognition accuracy of forensic examiners, super-recognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24), 6171–6176.
Piaget, J. (1930/2013). *The Child's Conception of Physical Causality*, trans. M. Gabain. London: Routledge.
Piaget, J. (1952). *The Origins of Intelligence in Children*. Madison, CT: International University Press.
Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.
Pietraszewski, D. (2018). A reanalysis of crossed-dimension "who said what?" paradigm studies, using a better error base-rate correction. *Evolution and Human Behavior*, 39(5), 479–489.
Pietraszewski, D. (2021). The correct way to test the hypothesis that racial categorization is a byproduct of an evolved alliance-tracking capacity. *Scientific Reports*, 11(1), 3404.
Pietraszewski, D. (2022). A (failed) attempt to falsify the alliance hypothesis of racial categorization: Racial categorization is not reduced when crossed with a nonalliance category. *Journal of Experimental Psychology: General*, 151(9), 2195–2203.
Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PloS One*, 9(2), e88534.
Pietraszewski, D., Curry, O. S., Petersen, M. B., Cosmides, L., & Tooby, J. (2015). Constituents of political cognition: Race, party politics, and the alliance detection system. *Cognition*, 140(C), 24–39.
Pinker, S. (1997). *How the Mind Works*. New York: Norton.
Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature*. London: Penguin.
Pinker, S. (2004). Why nature & nurture won't go away. *Daedalus*, 133(4), 5–17.
Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. New York: Allen Lane.
Plato (*c*.360 BCE / 1961) *Phaedo*. In E. Hamilton & H. Cairns (eds.), *Plato: The Collected Dialogues*, pp. 40–98. Princeton, NJ: Princeton University Press.
Plato (*c*.360 BCE / 1992). *The Republic*, trans. G. M. A. Grube. Indianapolis: Hackett.
Plato (*c*.380 BCE / 1961). *Meno*. In E. Hamilton & H. Cairns (eds.), *Plato: The Collected Dialogues*, pp. 353–384. Princeton, NJ: Princeton University Press.
Plebe, A., & Mazzone, M. (2016). Neural plasticity and concepts ontogeny. *Synthese*, 193(12), 3889–3929.
Pointer, M. R., & Attridge, G. G. (1998). The number of discernible colours. *Color Research & Application*, 23(1), 52–54.
Porter, D., & Neuringer, A. (1984). Music discrimination by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 10(2), 138–148.
Povinelli, D. J., & Eddy, T. J., (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, 61(3), 1–152.
Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, 110(41), E3965–E3972.
Powell, L. J., & Spelke, E. S. (2018). Human infants' understanding of social imitation: Inferences of affiliation from third party observations. *Cognition*, 170, 31–48.
Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
Prinz, J. (2005). The return of concept empiricism. In H. Cohen & C. Leferbvre (eds.), *Handbook of Categorization in Cognitive Science*, pp. 679–695. London: Elsevier Science Ltd.
Prinz, J. (2007). Is morality innate? In W. Sinnott-Armstrong (ed.), *Moral Psychology: The Evolution of Morality. Adaptations and Innateness*, Vol. 1. Cambridge, MA: MIT Press.

Prinz, J. (2012). *Beyond Human Nature: How Culture and Experience Shapes Our Lives*. New York: Allen Lane.
Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistics Review*, 19, 9–50.
Pun, A., Birch, S. A., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences*, 113(9), 2376–2381.
Pun, A., Birch, S. A. J., & Baron, A. S. (2021). The power of allies: Infants' expectations of social obligations during intergroup conflict. *Cognition*, 211, 104630.
Pun, A., Birch, S. A., & Baron, A. S. (2022). Infants infer third-party social dominance relationships based on visual access to intergroup conflict. *Scientific Reports*, 12(1), 18250.
Putnam, H. (1967). The "innateness hypothesis" and explanatory models in linguistics. *Synthese*, 17, 12–22.
Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, 70(19), 699–711.
Quartz, S. R. (2003). Innateness and the brain. *Biology and Philosophy*, 18(1), 13–40.
Quine, W. V. O. (1936/1976) Truth by convention. In his *The Ways of Paradox and Other Essays*, pp. 77–106, revised and enlarged ed. Cambridge, MA: Harvard University Press.
Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. O. (1969a). Linguistics and philosophy. In S. Hook (ed.), *Language and Philosophy: A Symposium*, pp. 95–98. New York: NYU Press.
Quine, W. V. O. (1969b). Natural kinds. In his *Ontological Relativity & Other Essays*, pp. 114–138. New York: Columbia University Press.
Rakoczy, H., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, 7(1), 17–21.
Ratcliffe, J. M, Fenton, M. B., & Galef Jr, B. G. (2003). An exception to the rule: Common vampire bats do not learn taste aversions. *Animal Behaviour*, 65, 385–389.
Ray, E., & Heyes, C. (2011). Imitation in infancy: the wealth of the stimulus. *Developmental Science*, 14(1), 92–105
Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, 13(10), 439–446.
Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, 102(23), 8386–8391.
Regolin, L., Rugani, R., Stancher, G., & Vallortigara, G. (2011). Spontaneous discrimination of possible and impossible objects by newly hatched chicks. *Biology Letters*, 7(5), 654–657.
Regolin, L., & Vallortigara, G. (1995). Perception of partly occluded objects by young chicks. *Perception and Psychophysics*, 57(7), 971–976.
Reid, T. (1785/2002). *Essays on the Intellectual Powers of Man*, ed. D. Brookes & K. Haakonssen. Edinburgh: Edinburgh University Press.
Reid, T. (1788/2011). *Essays on the Active Powers of Man*, reissue ed. Cambridge: Cambridge University Press.
Reiss, J. E., Hoffman, J. E., & Landau, B. (2005). Motion processing specialization in Williams syndrome. *Vision Research*, 45(27), 3379–3390.
Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2010). Preserved functional specialization for spatial processing in the middle occipital gyrus of the early blind. *Neuron*, 68(1), 138–148.
Rescher, N. (1966) A new look at the problem of innate ideas. *British Journal for the Philosophy of Science*, 17(3), 205–218.
Revusky, S., & Garcia, J. (1970). Learned associations over long delays. In G. H. Bower & J. T. Spence (eds.), *The Psychology of Learning and Motivation*, pp. 1–84. New York: Academic.
Rey, G. (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Oxford: Blackwell Publishers.
Rey, G. (2001). Physicalism and psychology: A plea for a substantive philosophy of mind. In C. Gillett & B. M. Loewer (eds.), *Physicalism and its Discontents*, pp. 99–128. Cambridge: Cambridge University Press.

Rey, G. (2014). Innate and learned: Carey, mad dog nativism, and the poverty of stimuli and analogies (yet again). *Mind & Language*, 29(2), 109–132.
Rezlescu, C., Barton, J. J., Pitcher, D., & Duchaine, B. (2014). Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia. *Proceedings of the National Academy of Sciences*, 111(14), 5123–5128.
Rhodes, M., Hetherington, C., Brink, K., & Wellman, H. M. (2015). Infants' use of social partnerships to predict behavior. *Developmental Science*, 18(6), 909–916.
Richardson, R. C. (2007). *Evolutionary Psychology as Maladapted Psychology*. Cambridge, MA: MIT Press.
Richerson, P., & Boyd, R. (2005) *Not by Genes Alone*. Chicago University Press.
Rips, L. J. (2017). Core Cognition and Its Aftermath. *Philosophical Topics*, 1(45), 157–179.
Rilling, M., & McDiarmid, C. (1965). Signal detection in fixed-ratio schedules. *Science*, *148*(3669), 526–527.
Robbins, J., & Rumsey, A. (2008). Introduction: Cultural and linguistic anthropology and the opacity of other minds. *Anthropological Quarterly*, 81(2), 407–420.
Rogers, T. T., Hocking, J., Mechelli, A., Patterson, K., & Price, C. (2005). Fusiform activation to animals is driven by the process, not the stimulus. *Journal of Cognitive Neuroscience*, 17(3), 434–445.
Rogers, T. T., & McClelland, J. L (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.
Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, 31(6), 689–714.
Rose, H. (2000). Colonising the social sciences? In H. Rose & S. Rose (eds.), *Alas, Poor Darwin: Arguments against Evolutionary Psychology*, pp. 106–128. London: Jonathan Cape.
Rose, H., & Rose, S. (2000). Introduction. In H. Rose & S. Rose (eds.), *Alas, Poor Darwin: Arguments against Evolutionary Psychology*, pp. 1–13. London: Jonathan Cape.
Rose, S. (2000). Escaping evolutionary psychology. In H. Rose & S. Rose (eds.), *Alas, Poor Darwin: Arguments against Evolutionary Psychology*, pp. 299–320. London: Jonathan Cape.
Roussel, E., Padie, S., & Giurfa, M. (2012). Aversive learning overcomes appetitive innate responding in honeybees. *Animal Cognition*, 15(1), 135–141.
Rozin, P. (1990). Development in the food domain. *Developmental Psychology*, 26(4), 555.
Rozin P., & Kalat, J., (1971). Specific hungers and poison avoidance as adaptive specializations of learning. *Psychological Review*, 78, 459–486.
Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, 50(4), 703.
Rubin, J. Z., Provenzano, F. J., & Luria, Z. (1974). The eye of the beholder: Parents' views on sex of newborns. *American Journal of Orthopsychiatry*, 44(4), 512.
Rugani, R., Regolin, L., & Vallortigara, G. (2010). Imprinted numbers: Newborn chicks' sensitivity to number vs. continuous extent of objects they have been reared with. *Developmental Science*, 13(5), 790–797.
Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, 347(6221), 534–536.
Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (eds.), *An Introduction to Neural and Electronic Networks*, pp. 405–420. Cambridge, MA: Academic Press.
Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart & J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: *Foundations*, pp. 110–146. Cambridge, MA: MIT Press.
Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: *Foundations*. Cambridge, MA: MIT Press.
Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (eds), *Attention and performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, pp. 3–30. MIT Press.

Russell, B. (1912). *The Problems of Philosophy*. Oxford: Oxford University Press.
Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105, 669–680.
Sameroff, A. (2005). The science of infancy: Academic, social, and political agendas. *Infancy*, 7(3), 219–242.
Samet, J. (2008). The historical controversies surrounding innateness. *The Stanford Encyclopedia of Philosophy* (Fall 2008 edition), E. N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2008/entries/innateness-history/.
Samet, J., & Flanagan, O. (1989) Innate representations. In S. Silvers (ed.), *Rerepresentation*, pp. 189–210. New York: Kluwer Academic Publishers.
Sampson, G. (2005). *The "Language Instinct" Debate*, revised ed. New York: Continuum.
Samuels, R. (2002). Nativism in cognitive science. *Mind and Language*, 17, 233–265.
Sann, C., & Streri, A. (2007). Perception of object shape and texture in human newborns: Evidence from cross-modal transfer tasks. *Developmental Science*, 10(3), 399–410.
Santos, L. R., & Rosati, A. G. (2015). The evolutionary roots of human decision making. *Annual Review of Psychology*, 66(1), 321–347.
Sarpal, D., Buchsbaum, B. R., Kohn, P. D., Kippenhan, J. S., Mervis, C. B., Morris, C. A., Meyer-Lindenberg, A., & Berman, K. F. (2008). A genetic model for understanding higher order visual processing: functional interactions of the ventral visual stream in Williams syndrome. *Cerebral Cortex*, 18(10), 2402–2409.
Saul, J. (2013). Implicit bias, stereotype threat and women in philosophy. In F. Jenkins & K. Hutchison (eds.), *Women in Philosophy: What Needs to Change?*, pp. 39–60. Oxford: Oxford University Press.
Sauter, D. A., Crasborn, O., Engels, T. F. S., Kamiloglu, R. G., Sun, R., Eisner, F., & Haun, D. B. M. (2019). Human emotional vocalisations can develop in the absence of auditory learning. *Emotion*, 20(8), 1435–1445.
Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6), 2408–2412.
Sauter, D. A., LeGuen, O., & Haun, D. (2011). Categorical perception of emotional facial expressions does not require lexical categories. *Emotion*, 11(6), 1479.
Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? Simple associations may explain moral reasoning in infants. *PloS One*, 7(8), e42698.
Schachner, A., Zhu, L., Li, J., & Kelemen, D. (2017). Is the bias for function-based explanations culturally universal? Children from China endorse teleological explanations of natural phenomena. *Journal of Experimental Child Psychology*, 157, 1–20.
Schacter, D. L., Gilbert, D. T., Wegner, D. M., & Nock, M. K. (2017). *Psychology*, 4th ed. Duffield: Worth Publishers.
Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, 20(2), 99–103.
Schilling, T. (2000). Infants' looking at possible and impossible screen rotations: The role of familiarization. *Infancy*, 1(4), 389–402.
Schlingloff, L., Csibra, G., & Tatone, D. (2020). Do 15-month-old infants prefer helpers? A replication of Hamlin et al. (2007). Royal Society Open Science, 7(4), 191795.
Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, 124(3), 325–333.
Schmidt, M. F. H., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PloS One*, 6, e23223.
Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, 23, 842–847.
Schneirla, T. C. (1957). The concept of development in comparative psychology. In D. B. Harris (ed.), *The Concept of Development*, pp. 78–108. Minneapolis: University of Minnesota Press.
Schubert, C. (2009). The genomic basis of the Williams–Beuren syndrome. *Cellular and Molecular Life Sciences*, 66(7), 1178–1197.

Scola, C., Holvoet, C., Arciszewski, T., & Picard, D. (2015). Further evidence for infants' preference for prosocial over antisocial behaviors. *Infancy*, 20(6), 684–692.
Scott, R. M. (2014). Post hoc versus predictive accounts of children's theory of mind: A reply to Ruffman. *Developmental Review*, 34(3), 300–304.
Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80(4), 1172–1196.
Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, 21(4), 237–249.
Scott, R. M., Baillargeon, R., Song, H.-J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, 61(4), 366–395.
Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, 15(2), 181–193.
Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, 82, 32–56.
Scrivner, C., Holbrook, C., Fessler, D. M., & Maestripieri, D. (2020). Gruesomeness conveys formidability: Perpetrators of gratuitously grisly acts are conceptualized as larger, stronger, and more likely to win. *Aggressive Behavior*, 46(5), 400–411.
Searle, J. R. (1992). *The Rediscovery of the Mind*. MIT press.
Seligman, M. (1970). On the generality of the laws of learning. *Psychological Review*, 77(5), 406–418.
Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., Von Rueden C, Krauss A, & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B: Biological Sciences*, 277(1699), 3509–3518.
Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., & Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B: Biological Sciences*, 276(1656), 575–584.
Sell, A., Hone, L. S., & Pound, N. (2012). The importance of physical strength to human males. *Human Nature*, 23(1), 30–44.
Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9), 668–671.
Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., Osanai, H., & Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, 22(2), 353.
Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger Syndrome. *Science*, 325(5942), 883–885.
Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, 5(1), 399–426.
Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands, *Proceedings of the National Academy of Sciences*, 113(47), 13360–13365.
Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, 110(40), 15937–15942.
Seyfarth, R. M., & Cheney, D. L. (2015). The evolution of concepts about agents: Or, what do animals recognize when they recognize an individual? In E. Margolis & S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press.
Shapiro, L. (2019). *Embodied Cognition*. London: Routledge.
Sharma, J., Angelucci, A., & Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780), 841–847.
Shea, N. (2012). Genetic representation explains the cluster of innateness-related properties. *Mind & Language*, 27(4), 466–493.
Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Sheehan, M. J., & Tibbetts, E. A. (2011). Specialized face learning is associated with individual recognition in paper wasps. *Science*, 334(6060), 1272–1275.
Shultz, S., & Vouloumanos, A. (2010). Three-month-olds prefer speech to other naturally occurring signals. *Language Learning and Development*, 6(4), 241–257.
Shweder, R. (1991). *Thinking through Cultures: Expeditions in Cultural Psychology*. Cambridge, MA: Harvard University Press.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T. Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–59.
Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2), 809–813.
Simon, T. J. (1997). Reconceptualizing the origins of number knowledge: A "non-numerical" account. *Cognitive Development*, 12(3), 349–372.
Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5(4), 644–649.
Sinnott-Armstrong, W. (2015) Consequentialism. *The Stanford Encyclopedia of Philosophy* (Winter 2015 edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2015/entries/consequentialism/>.
Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, 114(21), 5545–5550.
Skinner, B. F. (1957). *Verbal Behavior*. New York: Appleton-Century-Crofts.
Slater, A., & Bremner, G. (2017). *An Introduction to Developmental Psychology*, 3rd ed. Hoboken, NJ: Wiley.
Slater, A., Morison, V., Somers, M., & Mattock, A. (1990). Newborn and older infants' perception of partly occluded objects. *Infant Behavior and Development*, 13, 33–49.
Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, 23(2), 196–204.
Sloman, S. (2005). *Causal Models: How People Think about the World and its Alternatives*. Oxford: Oxford University Press.
Smith, A. D., McKeith, L., & Howard, C. J. (2013). The development of path integration: Combining estimations of distance and heading. *Experimental Brain Research*, 231, 445–455.
Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *PloS One*, 8(3), e59510.
Snedeker, J. (2008). Reading *Semantic Cognition* as a theory of concepts. *Behavioral and Brain Sciences*, 31(6), 727–728.
Sober, E. (1988). Aportioning causal responsibility. *Journal of Philosophy*, 85(6), 303–318.
Sober, E. (1998). Innate knowledge. In E. Craig (ed.), *The Routledge Encyclopedia of Philosophy*, Vol. 4, pp. 794–797. London: Routledge.
Song, H.-J., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*, 109(3), 295–315.
Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13(6), 907–912.
Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, 130(1), 1–10.
Sovrano, V. A., Bisazza, A., & Vallortigara, G. (2003). Modularity as a fish (*Xenotoca eiseni*) views it: Conjoining geometric and nongeometric information for spatial reorientation. *Journal of Experimental Psychology: Animal Behavior Processes*, 29(3), 199.

Sovrano, V. A., Potrich, D., & Vallortigara, G. (2013). Learning of geometry and features in bumblebees (*Bombus terrestris*). *Journal of Comparative Psychology*, 127(3), 312.

Sovrano, V. A., Rigosi, E., & Vallortigara, G. (2012). Spatial reorientation by geometry in bumblebees. *PLoS One*, 7(5), e37449.

Spelke, E. S. (1991). Physical knowledge in infancy: Reflections on Piaget's theory. In S. Carey & R. Gelman (eds.), *The Epigenesis of Mind: Essays on Biology and Cognition*, pp. 133–169. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, 21(2), 181–200.

Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In D. Gentner & S. Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought*, pp. 277–311. Cambridge, MA: MIT Press.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, 60(9), 950–958.

Spelke, E. S. (2022). *What Babies Know: Core Knowledge and Composition*, Vol. 1. Oxford: Oxford University Press.

Spelke, E. S., Breinlinger, K., Jacobson, K., & Phillips, A. (1993). Gestalt relations and object perception: A developmental study. *Perception*, 22, 143–1501.

Spelke, E. S., & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603), 2784–2793.

Spelke, E., Lee, S. A., & Izard, V. (2010). Beyond core knowledge: Natural geometry. *Cognitive Science*, 34(5), 863–884.

Spelke, E. S., & Newport, E. L. (1998). Nativism, empiricism, and the development of knowledge. In R. M. Lerner (ed.), *Handbook of Child Psychology*, Vol. 1: *Theoretical Models of Human Development*, pp. 275–340, 5th ed. New York: Wiley.

Spelke, E. S., & Tsivkin, S. (2001). Initial knowledge and conceptual change: Space and number. In M. Bowerman & S. C. Levinson (eds.), *Language Acquisition and Conceptual Development*, pp. 70–100. Cambridge: Cambridge University Press.

Spencer, J., Blumberg, M., McMurray, B., Robinson, S., Saumelson, L., & Tomblin, J. (2009). Short arms and talking eggs: Why we should no longer abide the nativist–empiricist debate. *Child Development Perspectives*, 3(2), 79–87.

Spencer, J., Samuelson, L., Blumberg, M., McMurray, B., Robinson, S., & Tomblin, J. (2009). Seeing the world through a third eye: Developmental systems theory looks beyond the nativist–empiricist debate. *Child Development Perspectives*, 3(2), 103–105.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, 67, 415–37.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4–28.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, pp. 39–67. Cambridge: Cambridge University Press.

Sperber, D., & Hirschfeld, L. (2006), Culture and modularity. In P. Carruthers, S. Laurence, & S. Stich (eds.), *Innate Mind: Culture and Cognition*, pp. 149–164. New York: Oxford University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press.

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (eds.), *Innate Mind: Culture and Cognition*, pp. 280–301. New York: Oxford University Press.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230), 91–94.

Starr, A., & Brannon, E. M. (2015). Evidence against continuous variables driving numerical discrimination in infancy. *Frontiers in Psychology*, 6, 1–6.

Stavans, M., & Baillargeon, R. (2018). Four-month-old infants individuate and track simple tools following functional demonstrations. *Developmental Science*, 21(1), e12500.
Stavans, M., Lin, Y., Wu, D., & Baillargeon, R. (2019). Catastrophic individuation failures in infancy: A new model and predictions. *Psychological Review*, 126(2), 196.
Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7), 509–528.
Sterelny, K. (1989) Fodor's nativism. *Philosophical Studies*, 55, 119–141.
Sterelny, K. (2003). *Thought in a Hostile World*. Oxford: Blackwell.
Sterelny, K. (2010). Moral nativism: A sceptical response. *Mind & Language*, 25(3), 279–297.
Stich, S. P. (1975). Introduction: The idea of innateness. In S. P. Stich (ed.), *Innate Ideas*, pp. 1–22. Berkeley, CA: University of California Press.
Stiles, J. (2009). On genes, brains, and behavior: Why should developmental psychologists care about brain development? *Child Development Perspectives*, 3(3), 196–202.
Strawson, P. F. (1966). *Bounds of Sense: Essay on Kant's "Critique of Pure Reason"*. London: Methuen.
Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, 144(3), 570.
Striem-Amit, E., Dakwar, O., Reich, L., & Amedi, A. (2012). The large-Scale Organization of "Visual" Streams Emerges without Visual Experience. *Cerebral Cortex*, 22(7), 1698–1709.
Striem-Amit, E., Vannuscorps, G., & Caramazza, A. (2017). Sensorimotor-independent development of hands and tools selectivity in the visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 114(18), 4787–4792.
Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences*, 105(1), 394–398.
Suzuki, K., & Kobayashi, T. (2000). Numerical competence in rats (*Rattus norvegicus*): Davis and Bradford (1986) extended. *Journal of Comparative Psychology*, 114(1), 73–85.
Symons, D. (1992) On the use and misuse of Darwinism in the study of behavior. In J. Barkow, L. Cosmides, & J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 137–159. New York: Oxford University Press.
Symposium on Solutions to Fodor's Puzzle of Concept Acquisition, Annual Cognitive Science Society (2005), Stressa, Italy. <https://www.harvardlds.org/wp-content/uploads/2017/01/Niyogi_Snedeker_2005-1.pdf>.
Sznycer, D., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., Smet, D. D., Ermer, E., Kim, S., Kim, S., Li, N. P., Seal, M. F. L., McClung, J., O, J., Ohtsubo, Y., Quillien, T., Schaub, M., Sell, A., van Leeuwen, F., Cosmides, L., and Tooby, J. (2017). Cross-cultural regularities in the cognitive architecture of pride. *Proceedings of the National Academy of Sciences*, 114(8), 1874–1879.
Sznycer, D., Xygalatas, D., Alami, S., An, X.-F., Ananyeva, K. I., Fukushima, S.,… Tooby, J. (2018). Invariances in the architecture of pride across small-scale societies. *Proceedings of the National Academy of Sciences*, 10, 1–6.
Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, 29(3), 288–299.
Tager-Flusberg, H., Plesa-Skwerer, D., Faja, S., & Joseph, R. M. (2003). People with Williams syndrome process faces holistically. *Cognition*, 89(1), 11–24.
Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12, 49–100.
Talmy, L. (2000). *Toward a Cognitive Semantics*, Vol. 1: *Concept Structuring Systems*. Cambridge, MA: MIT Press.
Tan, E., & Hamlin, J. K. (2022). Mechanisms of social evaluation in infancy: A preregistered exploration of infants' eye-movement and pupillary responses to prosocial and antisocial events. *Infancy*, 27(2), 255–276.
Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, 137, 47–62.

Tatone, D., Hernik, M., & Csibra, G. (2019). Minimal cues of possession transfer compel infants to ascribe the goal of giving. *Open Mind*, 3, 31–40.
Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159.
Téglás, E., Ibanez-Lillo, A., Costa, A., & Bonatti, L. L. (2014). Numerical representations and intuitions of probabilities at 12 months. *Developmental Science*, 18(2), 183–193.
Téglás, E., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332(6033), 1054–1059.
Tenter, A. M., Heckeroth, A. R., & Weiss, L. M. (2000). *Toxoplasma gondii*: From animals to humans. *International Journal for Parasitology*, 30, 1217–1258.
The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*, 282, 2012–2018.
The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74.
Thelen, E. (2000). Grounded in the world: Developmental origins of the embodied mind. *Infancy*, 1(1), 3–28.
Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, 24(1), 1–34.
Thelen, E., & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Perception and Action*. Cambridge, MA: MIT Press.
Thomas, A. J., Thomsen, L., Lukowski, A. F., Abramyan, M., & Sarnecka, B. W. (2018). Toddlers prefer those who win but not when they win by force. *Nature Human Behaviour*, 2(9), 662.
Thomasson, A. (2007). Artifacts and human concepts. In E. Margolis & S. Laurence (eds.), *Creations of the Mind: Theories of Artifacts and Their Representation*, pp. 52–73. Oxford: Oxford University Press.
Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, 9, 483–484.
Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, 331(6016), 477–480.
Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology*, 92(1), 144–151.
Timp, W., & Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. *Science Advances*, 6(2), eaax8978.
Ting, F., He, Z., & Baillargeon, R. (2019). Toddlers and infants expect individuals to refrain from helping an ingroup victim's aggressor. *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), 6025–6034.
Tomasello, M., & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
Tooby, J., & Cosmides, L. (1989) The innate versus the manifest: How universal does universal have to be? *Behavioral and Brain Sciences*, 12(1), 36–37.
Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 19–136. New York: Oxford University Press.
Tooby, J., & Cosmides, L. (2005). Conceptual foundations of evolutionary psychology. In D. M. Buss (ed.), *The Handbook of Evolutionary Psychology*, pp. 5–67. Hoboken, NJ: Wiley & Sons.
Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, 321(5897), 1831–1834.
Tracy, J., & Matsumoto, D. (2008). The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences of the United States of America*, 105(33), 11655.
Tracy, J. L., & Robins, R. W. (2008a). The automaticity of emotion recognition. *Emotion*, 8(1), 81–95.

Tracy, J. L., & Robins, R. (2008b). The nonverbal expression of pride: Evidence for cross-cultural recognition. *Journal of Personality and Social Psychology*, 94(3), 516–530.
Tracy, J. L., Shariff, A. F., & Cheng, J. T. (2010). A naturalist's view of pride. *Emotion Review*, 2(2), 163–177.
Tracy, J. L., Shariff, A. F., Zhao, W., & Henrich, J. (2013). Cross-cultural evidence that the nonverbal expression of pride is an automatic status signal. *Journal of Experimental Psychology General*, 142(1), 163–180.
Turati, C., Bulf, H., & Simion, F. (2008). Newborns' face recognition over changes in viewpoint. *Cognition*, 106(3), 1300–1321.
Turati, C., Gava, L., Valenza, E., & Ghirardi, V. (2013). Number versus extent in newborns' spontaneous preference for collections of dots. *Cognitive Development*, 28(1), 10–20.
Turati, C., Macchi Cassia, V., Simion, F., & Leo, I. (2006). Newborns' face recognition: Role of inner and outer facial features. *Child Development*, 77(2), 297–311.
Twyman, R. M. (2014) *Principles of Proteomics*, 2nd ed. London: Garland Science.
Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.
Tybur, J. M., Laakasuo, M., Ruff, J., & Klauke, F. (2016). How pathogen cues shape impressions of foods: the omnivore's dilemma and functionally specialized conditioning. *Evolution and Human Behavior*, 37(5), 376–386.
Uller, C., Carey, S., Huntley-Fenner, G., & Klatt, L. (1999). What representations might underlie infant numerical knowledge? *Cognitive Development*, 14(1), 1–36.
Valenza, E., & Bulf, H. (2011). Early development of object unity: Evidence for perceptual completion in newborns. *Developmental Science*, 14(4), 799–808.
Vallortigara, G., & Regolin, L. (2006). Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology*, 16(8), R279–R280.
Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLoS Biology*, 3(7), e208.
Vallortigara, G., Zanforlin, M., & Pasti, G. (1990). Geometric modules in animals' spatial representations: A test with chicks (*Gallus gallus domesticus*). *Journal of Comparative Psychology*, 104(3), 248.
van Buren, B., & Scholl, B. J. (2017). Minds in motion in memory: Enhanced spatial memory driven by the perceived animacy of simple shapes. *Cognition*, 163, 87–92.
Van Essen, D. (2004). Organization of visual areas in macaque and human cerebral cortex. In L. M. Chalupa & J. S. Werner (eds.), *The Visual Neurosciences*, Vol. 1, pp. 507–521. Cambridge, MA: MIT Press.
Vannuscorps, G., Wurm, M. F., Striem-Amit, E., & Caramazza, A. (2019). Large-scale organization of the hand action observation network in individuals born without hands. *Cerebral Cortex*, 29(8), 3434–3444.
Van Strien, J. W., & Isbell, L. A. (2017). Snake scales, partial exposure, and the Snake Detection Theory: A human event-related potentials study. *Scientific Reports*, 7(1), 1–9.
Verhage, M., Maia, A. S., Plomp, J. J., Brussaard, A. B., Heeroma, J. H., Vermeer, H., Toonen, R. F., Hammer, R. E., van den Berg, T. K., Missler, M., Geuze, H. J., & Südhof, T. C. (2000). Synaptic assembly of the brain in the absence of neurotransmitter secretion. *Science*, 287(5454), 864–869.
Versace, E., Martinho-Truswell, A., Kacelnik, A., & Vallortigara, G. (2018). Priors in animal and artificial intelligence: where does learning begin? *Trends in Cognitive Sciences*, 22(11), 963–965.
Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The tuning of human neonates' preference for speech. *Child Development*, 81(2), 517–527.
Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate? *Developmental Science*, 17(6), 872–879.
Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E. M., Bulnes, L. C. Caldwell, T. L.,

Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S.,Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Ozdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., & Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7(11), 483–488.

Wang, J. J., & Feigenson, L. (2019). Is empiricism innate? Preference for nurture over nature in people's beliefs about the origins of human knowledge, *Open Mind*, Early Access Version, 1–12.

Wang, J. J., Libertus, M. E., & Feigenson, L. (2018). Hysteresis-induced changes in preverbal infants' approximate number precision. *Cognitive Development*, 47, 107–116.

Wang, L., & Leslie, A. M. (2016). Is implicit theory of mind the "real deal"? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language*, 31(2), 147–176.

Wang, S. H. (2019). Regularity detection and explanation-based learning jointly support learning about physical events in early infancy. *Cognitive psychology*, 113, 101219.

Wang, S.-H., & Baillargeon, R. (2008). Can infants be "taught" to attend to a new physical variable in an event category? The case of height in covering events. *Cognitive Psychology*, 56(4), 284–326.

Wang, S. H., Baillargeon, R., & Paterson, S. (2005). Detecting continuity violations in infancy: A new account and new evidence from covering and tube events. *Cognition*, 95(2), 129–173.

Wang, S. H., & Kohne, L. (2007). Visual experience enhances infants' use of task-relevant information in an action task. *Developmental Psychology*, 43(6), 1513.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765), 1301–1303.

Warneken, F., & Tomasello, M. (2013). Parental presence and encouragement do not influence helping in young children. *Infancy*, 18(3), 345–368.

Warneken, F., & Tomasello, M. (2014). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. Developmental psychology, 44(6), 1785–1788.

Warrington, E. K., & McCarthy, R. A. (1983). Category specific access dysphasia. *Brain*, 106, 859–878.

Watanabe, S., Sakamoto, J., & Wakita, M. (1995). Pigeons' discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior*, 63(2), 165–174.

Wehner, R., & Srinivasan, M. (1981). Searching behavior of desert ants, genus *Cataglyphis* (Formicidae, Hymenoptera). *Journal of Comparative Physiology*, 142, 315–338.

Weigelt, S., Koldewyn, K., Dilks, D. D., Balas, B., McKone, E., & Kanwisher, N. (2014). Domain-specific development of face memory but not face perception. *Developmental Science*, 17(1), 47–58.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20(3), 470–477.

Weiskopf, D. 2008: The origins of concepts. *Philosophical Studies*, 140, 359–384.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New Directions. *Annual Review of Psychology*, 66(1), 173–196.

Werker J. F., & Tees R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Developement*, 7, 49–63.

Wertz, A. E. (2019). How plants shape the mind. *Trends in Cognitive Sciences*, 23(7), 528–531.

Wertz, A. E., & Wynn, K. (2014a). Thyme to touch: Infants possess strategies that protect them from dangers posed by plants. *Cognition*, 130(1), 44–49.

Wertz, A. E., & Wynn, K. (2014b). Selective social learning of plant edibility in 6- and 18-month-old infants. *Psychological Science*, 25(4), 874–882.

White, D. J., Ho, L., & Freed-Brown, G. (2009). Counting chicks before they hatch: Female cowbirds can time readiness of a host nest for parasitism. *Psychological Science*, 20(9), 1140–1145.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, 116(3), 580–601.

White, P. A. (2012). The experience of force: The role of haptic experience of forces in visual perception of object motion and interactions, mental simulation, and motion-related judgments. *Psychological Bulletin*, 138(4), 589–615.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Wierzbicka, A. (2001). Leibnizian linguistics. In I. Kenesei & R. M. Harnish (Eds), *Perspectives on Semantics, Pragmatics, and Discourse: A Festschrift for Ferenc Kiefer* (Vol. 90), pp. 229–253. John Benjamins Publishing.

Wierzbicka, A. (2015). Innate conceptual primitives manifested in the languages of the world and in infant cognition. In E. Margolis & S. Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts*, pp. 379–412. Cambridge, MA: MIT Press.

Wilcox, T., Woods, R., Tuggy, L., & Napoli, R. (2006). Shake, rattle, and…one or two objects? Young infants' use of auditory information to individuate objects. *Infancy*, 9(1), 97–123.

Wilcoxon, H. C., Dragoin, W. B., & Kral, P. A. (1971). Illness-induced aversions in rat and quail: Relative salience of visual and gustatory cues. *Science*, 171, 826–828.

Williamson, M. (2012). *How Proteins Work*. London: Garland Science.

Wills, T. J., Cacucci, F., Burgess, N., & O'Keefe, J. (2010). Development of the hippocampal cognitive map in preweanling rats. *Science*, 328(5985), 1573–1576.

Wilson, M. L., Hauser, M. D., & Wrangham, R. W. (2001). Does participation in intergroup conflict depend on numerical assessment, range location, or rank for wild chimpanzees? *Animal Behaviour*, 61(6), 1203–1216.

Wilson, R. A. (2008). The drink you have when you're not having a drink. *Mind & Language*, 23, 273–283.

Witherington, D. C., Overton, W. F., Lickliter, R., Marshall, P. J., & Narvaez, D. (2018). Metatheory and the primacy of conceptual analysis in developmental science. *Human Development*, 61(3), 181–198.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Wittlinger, M., Wehner, R., & Wolf, H. (2006). The ant odometer: Stepping on stilts and stumps. *Science*, 312, 1965–1967.

Włodarczyk, A., Elsner, C., Schmitterer, A., & Wertz, A. E. (2018). Every rose has its thorn: Infants' responses to pointed shapes in naturalistic contexts. *Evolution and Human Behavior*, 39(6), 583–593.

Włodarczyk, A., Rioux, C., & Wertz, A. E. (2020). Social information reduces infants' avoidance of plants. *Cognitive Development*, 54, 100867.

Wood, J. N., & Wood, S. M. W. (2020). One-shot learning of view-invariant object representations in newborn chicks. *Cognition*, 199, 104192.

Wood, J. N., Ullman, T. D., Wood, B. W., Spelke, E. S., & Wood, S. M. W. (2024). Object permanence in newborn chicks is robust against opposing evidence. arXiv.

Woo, B. M., & Spelke, E. S. (2023). Toddlers' social evaluations of agents who act on false beliefs. *Developmental Science*, *26*(2), e13314.

Wolbers, T., Wiener, J. M., Mallot, H. A., & Büchel, C. (2007). Differential recruitment of the hippocampus, medial prefrontal cortex, and the human motion complex during path integration in humans. *The Journal of Neuroscience*, 27(35), 9408–9416.

Wolbers, T., Zahorik, P., & Giudice, N. A. (2011). Decoding the direction of auditory motion in blind humans. *NeuroImage*, 56(2), 681–687.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology General*, 136(1), 82–111.

Wolff, P. (2017). Force dynamics. In M. R. Waldmann (ed.), *The Oxford Handbook of Causal Reasoning*. New York: Oxford University Press.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. In B. H. Ross (ed.), *The Psychology of Learning and Motivation*, Vol. 58, pp. 167–202. Cambridge, MA: Academic Press.

Woo, B. M., Liu, S., & Spelke, E. (2021). Open-minded, not naïve: Three-month-old infants encode objects as the goals of other people's reaches. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43), 514–520.

Woo, B. M., & Spelke, E. (2020). Toddlers' social evaluations of agents who act on false beliefs. *Developmental Science*, 26(2), e13314.

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, 168, 154–163.

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences*, 110(34), 14000–14005.

Wood, J. N. (2016). A smoothness constraint on the development of object recognition. *Cognition*, 153, 140–145.

Wood, J. N., & Spelke, E. S. (2005). Infants' enumeration of actions: Numerical discrimination and its signature limits. *Developmental science*, 8(2), 173–181.

Wood, J. N., & Wood, S. M. (2016). The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society, B*, 283(1829), 20160166.

Woodruff Carr, K., Perszyk, D. R., & Waxman, S. R. (2021). Birdsong fails to support object categorization in human infants. *PloS One*, 16(3), e0247430.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1–34.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, 22(2), 145–160.

Wright, L. (1973). Functions. *Philosophical Review*, 82, 139–68.

Wu, C. T., Crouzet, S. M., Thorpe, S. J., & Fabre-Thorpe, M. (2015). At 120 msec you can spot the animal but you don't yet know it's a dog. *Journal of Cognitive Neuroscience*, 27(1), 141–149.

Wynn, K. (1992). Issues concerning a nativist theory of numerical knowledge. *Mind & Language*, 7(4), 367–381.

Wynn, K., Bloom, P., & Chiang, W. C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, 83(3), B55–B62.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, 89(1), B15–B25.

Xu, F., & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology*, 25(1), 103–108.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74(1), B1–B11.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, 8(1), 88–101.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

Yang, C. (2006). *The Infinite Gift: How Children Learn and Unlearn the Languages of the World*. New York: Scribner.

Yang, J., Kanazawa, S., Yamaguchi, M. K., & Kuriki, I. (2016). Cortical response to categorical color perception in infants investigated by near-infrared spectroscopy. *Proceedings of the National Academy of Sciences*, 113(9), 2370–2375.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, 47(10), 2065–2072.

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui,

P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., & Yang, H. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 296, 79–92.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., de Sousa, B. L., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W. J., Santos, M. R., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Hansen, R. S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, R. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., & The Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515, 355–364.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11, 207–219.

Zipple, M. N., Caves, E. M., Green, P. A., Peters, S., Johnsen, S., & Nowicki, S. (2019). Categorical colour perception occurs in both signalling and non-signalling colour ranges in a songbird. *Proceedings of the Royal Society B*, 286(1903), 20190524.

Ziv, T., & Sommerville, J. A. (2017). Developmental differences in infants' fairness expectations from 6 to 15 months of age. *Child Development*, *88*(6), 1930–1951.

Index

Note: Boxes and figures are indicated by an italic "*b*", "*f*", and notes are indicated by "n." following the page number

Since the index has been created to work across multiple formats, indexed terms for which a page range is given (e.g., 52–53, 66–70, etc.) may occasionally appear only on some, but not all of the pages within the range.