Embodied Decisions and the Predictive Brain

Christopher D. Burr

University of BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Arts

November 2016

Word count: 81766

Abstract

Decision-making has traditionally been modelled as a serial process, consisting of a number of distinct stages. The traditional account assumes that an agent first acquires the necessary perceptual evidence, by constructing a detailed inner representation of the environment, in order to deliberate over a set of possible options. Next, the agent considers her goals and beliefs, and subsequently commits to the best possible course of action. This process then repeats once the agent has learned from the consequences of her actions and subsequently updated her beliefs. Under this interpretation, the agent's body is considered merely as a means to report the decision, or to acquire the relevant goods. However, embodied cognition argues that an agent's body should be understood as a proper part of the decision-making process. Accepting this principle challenges a number of commonly held beliefs in the cognitive sciences, but may lead to a more unified account of decision-making.

This thesis explores an embodied account of decision-making using a recent framework known as predictive processing. This framework has been proposed by some as a functional description of neural activity. However, if it is approached from an embodied perspective, it can also offer a novel account of decision-making that extends the scope of our explanatory considerations out beyond the brain and the body. We explore work in the cognitive sciences that supports this view, and argue that decision theory can benefit from adopting an embodied and predictive perspective.

Dedication and Acknowledgements

It is not uncommon to hear researchers state that their work can be a lonely task. I have not felt this to be the case, and I am confident that this is thanks to the support of my family and close friends. I should start by acknowledging the unending support of my wife, who has been by my side since my first years as an undergraduate. Without your support, encouragement, and love, I do not believe I would have been able to achieve as much as I have done. The same goes for my family, who help ground me and remind me not to take life too seriously! Special thanks go to my parents for all of their many forms of support over the years.

I have had the pleasure of conducting my doctoral research at the University of Bristol, and during my time I have met many fantastic people. The Department of Philosophy is a friendly and supporting place, and I will be very sad to leave it behind. My thanks to all of the staff and students that I have interacted with, no matter how brief it may have been, and especially to the following (in no particular order): Finn Spicer, Samir Okasha, Havi Carel, Anya Farrenikova, James Ladyman, Tudor Baetu, Karim Thebault, Bengt Autzen, Alexander Bird, Anthony Everett, Kit Patrick, Vincenzo Politi, Seiriol Morgan and Andrew Pyle. In addition, I have really enjoyed being a part of the postgraduate community, and hope that I have helped make it enjoyable for others. My thanks to the following postgraduates: Niall Paterson, Nick Cosstick, Ben Springett, Richard Bowles, Prakhar Manas, Aaron Guthrie, Bon-Hyuk Koo, Sam Roberts, Alejandra Casas Munoz, Aadil Kurji, Jessica

Laimann, Lilit Movsisyan, Alice Monypenny and Mo Abolkheir.

I must also single out two postgraduates especially: Oliver Lean and Max Jones. You have both provided an enormous amount of support, and have been kind enough to help me with my research in so many ways. I am deeply indebted to the both of you, and wish you the very best in your own academic careers. You will always remain true rivals, in only the most positive of manners!

As a member of the Epistemic Utility Theory: Foundations and Applications project—generously supported by the European Research Council—I have had the fortune of meeting many fantastic researchers who have visited the project over the last three years. In addition to these visits, I have had the pleasure of working alongside Jason Konek, Ben Levinstein, Pavel Janda, Patricia Rich and Greg Gandenberger—my thanks to all of you. I wish you all the best of luck, and I hope that our paths continue to cross in the future.

Thank you to the members of the Embodied Cognition and Bayesian Brain reading groups. I am very glad that I organised these groups, as it allowed me to meet so many interesting people from outside of the Philosophy department. My thanks to you all for the many stimulating discussions, which helped me produce the thesis that you are currently reading.

My thanks to Wanja Wiese and Thomas Metzinger for organising a superb conference on the Philosophy of Predictive Processing. Your invitation to participate encouraged me to formulate many of the ideas that are contained within this thesis, and connected me with many superb researchers.

A special thanks to Andy Clark for writing the paper that started me on a path that led to the production of this thesis, and for your encouragement along the way. Thank you also for agreeing to be my external examiner, along with Samir Okasha, and for what I imagine will be a challenging but rewarding conversation during my viva.

I would also like to thank the following (in no particular order) for a variety of interesting discussions that have helped me with some of the ideas in this thesis: Zoe Drayson, Nathan Lepora, Giovanna Colombetti, Mark Sprevak, Dave Ward, Alistair Isaac, Michael Miller, Jakob Hohwy, Michael Anderson, Regina Fabry, Martin Butz, Lisa Quadt, Joe Dewhurst, Krys Dolega, Jonny Lee, Daniel Burnston, Roberto Fumagalli, William Ramsey, Kevin O'Regan, William Bechtel, Robert Rupert, David Spurrett, Harry Farmer, Alexandra Georgescu, Anna Ciaunica and Catarina Dutilh-Novaes.

Finally, my thanks to Richard Pettigrew. You truly embody the academic principles and standards that I hope to attain myself someday. You have been a superb supervisor and friend, and I am forever grateful for the opportunity that you have given me. I hope that I am able to offer the same encouragement, support and inspiration, which you have shown me, to a student of my own in the not too distant future.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE:..........................

Publications

Some of the material in this thesis is reprinted from the following publications:

1. Burr, Christopher and Max Jones (2016) “The body as laboratory: Prediction-error minimization, embodiment, and representation.” In: *Philosophical Psychology* 29.4, pp. 586-600

2. Burr, C (Forthcoming) “Embodied Decisions and the Predictive Brain” In: *OpenMIND: The Philosophy of Predictive Processing.* Ed. by Thomas Metzinger and Wanja Wiese.

In line with the statement made in the Author’s Declaration, where this is representative of collaborative work references have been provided, and only those sections of this previous research that were written by myself have been replicated in full.

Contents

Introduction

Who's Making The Decisions Around Here?

Consider the following situation: you are busy writing a doctoral thesis, which has a deadline that is fast approaching. You have already been writing for a couple of hours without a break, and begin to notice that your productivity is decreasing. What should you do?

It is almost lunchtime, and you consider the fact that your tiredness is a product of your hunger. However, there is a chance that if you take too long of a break you will be unable to pick up where you left off. Perhaps you should continue working for a bit longer and have lunch once the word count has been reached, or maybe it is better to simply take a short break now to make a coffee, and hope that the caffeine will allow you to persist with the writing for a bit longer. Should you stop here? Does this set of options collectively characterise the decision problem? What about doing something else entirely? Far from being a simple matter of choosing between a few well-delineated options, it appears that you must also figure out what options are available to you. Everyday decisions, such as these, do not come fully-formed, neatly representing a set of exhaustive and mutually exclusive options to choose between. It seems that agents like ourselves are first required to determine a set of options prior to deliberating about, and then committing to one of them. However, is this the right way to characterise what is actually happening when we make decisions?

A complete account of how we make decisions, among other things, should be able to explain the full breadth of these processes. Unfortunately, the apparent simplicity of the current example belies an enormous complexity that is found at the level of the mechanisms involved in making these sorts of everyday decisions. For example, how is the initial process of specifying the possible options achieved? It will be argued that answering this question, as well as others, requires an appreciation of multiple levels of scientific explanation, spanning multiple disciplines and physical scales. This will involve challenging the idea that the brain is the seat of decision-making, a picture that casts some central executive region as computing the relevant decision variables, before instructing the body of the agent to carry out the necessary behaviour. It will also involve looking beyond the physiological boundaries of the agent, to understand the inseparability of ecological considerations from the process of decision-making. To do this, over the course of this thesis we will develop and defend an embodied version of a framework known as predictive processing. We will then use this perspective to explore an account of decision-making. Such a view, we argue, is preferable to alternative accounts of decision-making.

Before we can answer the question, "Who's making the decisions around here?" (as specified in the title of this section) we need to first determine *what it is we are looking for*, and also *where we should look*. Of course these are just vague ways of stating the following questions: which system or phenomenon are we trying to explain, and which theoretical framework (and corresponding methodology) is relevant for our purpose? As we are interested in the study of decision-making, an answer to the first question should provide details that help us demarcate which physical processes can usefully be described as constituting, and contributing to, decision-making. The second question should specify the theoretical framework that attempts to provide an account of the mechanisms that are involved in the phenomenon under investigation.

Framed as it is currently, the first question assumes that a satisfactory definition of decision-making has already been provided. To rectify this, we shall take decision-making to be the process of selecting an action from a set of *alternative options*. We will allow, as in the example above, for an incomplete set of possible options to be specified prior to selection. We will also allow for the act to be both *epistemic* and *pragmatic* in nature, and state that the set of alternative options be restricted to only those that are viable for the system in question. This definition will suffice for the present discussion, and will be made more precise in the relevant chapters.

What about the second question? Should we turn to decision theory to provide the necessary answers? Of course the answer is an obvious 'yes', but we need to be a bit more precise, as decision theory is an interdisciplinary project to which philosophers, economists, psychologists and statisticians, among others, contribute. There is also the fact that decision theory is separated into *descriptive* and *normative* approaches, where the first is viewed as an empirical approach that aims to provide an account of how decisions are made, and the second is understood as providing prescriptions for what decision-makers are rationally required to do (Peterson, 2009).

The separation of the two approaches is sensible given the unsurprisingly large number of questions we could ask about a variety of decision-making systems, which may also differ widely in their capacities. To understand why two systems differ comparatively, it may be necessary to point to a difference in the mechanisms that the two systems utilise, as well as providing a general (and justifiable) standard for the evaluation of the systems under investigation. The former is considered the domain of descriptive decision theory, whereas the latter is considered the domain of normative decision theory. However, it is also sensible to ask whether the evaluative standard we adopt is appropriate given the systems we are interested in understanding. Kant's famous doctrine of 'ought implies can' (Kant, 1781), though by no means uncontroversial (cf. Stern, 2004), is widely accepted as a constraint on normative requirements. In short, if a system is incapable of acting in the manner prescribed by a norm of decision theory, the application of the norm is inappropriate. Despite a personal interest in the relationship between descriptive and normative decision theory we restrict discussion in this thesis to the former.

Given the simple definition provided of decision-making as the process of selecting an action from a set of alternative options, we can begin to narrow our search by asking which systems can usefully be described as making decisions. Should a flower that appears to track the movement of the sun through the process known as heliotropism be described as making a decision to do so? What about the insect that exhibits similar phototaxic behaviour, but which ends up moving towards a fatal source of artificial light? Was this ostensibly "suicidal" behaviour the product of a decision-making process, or can an alternative, and more appropriate explanation be given for its behaviour? If asked to draw a line that demarcates systems that can be usefully described as possessing the capacity for decision-making from those that do not, it is likely that most would draw a line that subsequently delineates a subset of living systems—probably based on some perceived degree of organisational complexity. Though there will invariably be disagreement about whether it is inappropriate to ascribe decision-making capacities to the plant and not the insect, most would agree that a rock at least makes no decisions at all. It is beyond the scope of this thesis to explore these sorts of debates in any real depth, and therefore we initially restrict our discussion to human cognition, appealing to studies using non-human animals insofar as they have explanatory interest to humans. However, there is also a theme of evolutionary continuity that runs throughout this thesis, and in the final chapters we will return to some of these topics. The following section provides an outline of this thesis, including a brief description of the questions that will be explored.

Thesis Outline

The first two chapters are summative in nature, providing the reader with some necessary context and terminological clarification.

In *chapter 1* we begin by providing a short history of the cognitive sciences. This history will introduce some of the concepts and debates that help to clarify and situate our understanding of the two questions posed in the previous section. It starts from the time that a unified discipline emerged in the 1960s, and ends with a discussion concerning embodied cognition. Embodied cognition is sometimes seen as a post-cognitivist paradigm, and to demonstrate why this is the case, this chapter contrasts the two approaches. However, embodied cognition is also composed of a number of separate approaches, and we review some of the themes that have been defended by those who work in this area. It should be noted that we do not commit ourselves to any particular theme throughout the course of this thesis. Instead, given that embodied cognition is best understood (at present) as a research program, we try to emphasise the need for explanatory pluralism where possible. This does not mean ignoring conflicting explanations, but neither does it mean we need to become involved with every micro-debate that arises in the course of discussion (e.g. representationalism versus anti-representationalism). The position we wish to defend over the course of this thesis is that decision-making (in the descriptive sense) is best understood from an embodied perspective, and that predictive processing offers a promising framework to develop our understanding of this process. As long as the core of embodied cognition (i.e. a rejection of cognitivism and an inseparable explanatory role for the body) is maintained, it should not matter whether all of the themes are vindicated. We can accept this, while at the same time acknowledging that greater focus on the conceptual foundations of the embodied cognition research program is a worthwhile pursuit.

In *chapter 2*, we provide an introduction to the contemporary framework known as predictive processing. This framework overturns the idea that the brain is a passive system that receives inputs from, and constructs a detailed representation of, the world. Instead, predictive processing argues that the brain evolved to help coordinate adaptive action selection by anticipating salient future states of the world on the basis of top-down predictions of sensory information. These predictions emerge from a hierarchically-organised generative model, which is encoded by the brain, and is constantly updated on the basis of incoming information from the world. We will argue that, contrary to some interpretations, the predictive processing framework need not be construed as supporting an internalist conception of the mind. Instead, we will argue the framework is best understood from an embodied perspective.

Before we explore how predictive processing accounts for decision-making, we first explain what it means to say that decision-making is embodied. This is done in *chapter 3* where we look at a number of approaches to understanding decision-making. First, we explore the traditional cognitivist picture of decision-making and problem-solving. Second, we briefly introduce the contemporary field of neuroeconomics, alongside an introduction to expected utility theory. Lastly, we argue that recent neurophysiological evidence challenges these conceptions, and instead points to a need to reconsider decision-making from an embodied perspective.

In *chapter 4*, we start to explore how the idea that decision-making is an embodied behaviour can be accommodated by an embodied account of predictive processing. This requires blurring the boundaries between perception, cognition, action and emotion, and also reconsidering the fundamental role of the brain, body and world in shaping cognition. As a consequence of this, the traditional cognitivist picture—a picture that views perception, cognition and action as separate, encapsulated processes—is severely undermined. This also means rejecting the cognitivist conception of decision-making as a serial process of deliberation and commitment, which is also independent of sensorimotor processes.

In *chapter 5*, having undermined the cognitivist account of decision-making, we begin to further develop our account of decision-making. Turning our attention first to the brain, we explore more fully the notion of *precision-weighting* that is introduced in chapter 2. This notion is an important component of predictive processing, and we explore how work in cognitive neuroscience presents a compelling reason to reconsider the brain in a more interactive manner. To explain how the interactive, predictive brain is able to support effective decision-making, we are required to first give up on the traditional ontological commitments of cognitive psychology, and begin to look outwards beyond the brain for a new taxonomy.

To complete our discussion of embodied decisions and the predictive brain, we turn in *chapter 6* to build a novel proposal for how the body and the world provide important constraints on decision-making. Most of the empirical evidence that will have been explored in earlier chapters pertains to *habitual* decision-making, but this is only one form of decision-making. Therefore, we begin by discussing the differences between habitual and deliberative decisions, and why it is important that predictive processing is able to account for both forms. Initially it appears as though embodied decision-making is only able to account for the former. To respond to this worry, we turn to consider some of the ways that additional processes constrain our choice behaviour, and whether this affects our understanding of embodied decision-making. This chapter identifies a novel approach to decision-making, but acknowledges that a full analysis requires further development.

We end with some further remarks regarding the interpretation of decision theory, and also bring the discussion back to some of the questions raised in this introductory chapter by exploring some ideas in comparative psychology.

Chapter 1

Cognitive Systems

> "One might say that cognitive science has a very long past but a relatively short history." (Gardner, 1985)

As discussed in the introduction, the focus of this thesis is decision-making, and we will be exploring this process from the perspective of cognitive science. This chapter serves as both a motivation and a foundation for the subsequent chapters.

Gardner's characterisation of the dual history of cognitive science recognises the long interest we have had in questions pertaining to the mind and behaviour, but also acknowledges the more recent emergence of a recognised scientific framework in which to study these phenomena. The disciplines that formed the interdisciplinary framework were initially unified in terms of their rejection of some aspects of the preceding paradigms (e.g. introspectionism and behaviorism).[1] However, on its own,

[1] Given the interdisciplinary nature of cognitive science, it is tempting to opt for the pluralistic label of the 'cognitive sciences', rather than the singular and arguably more monolithic term 'cognitive science'. Though the latter is more frequently used in the literature, increased usage of the former is advisable, in order to recognise the pluralistic nature of the scientific practice and study of mind and behaviour. Throughout this thesis, we will opt for the following: when historical accuracy is called for, we will use the singular term, but otherwise the plural term will be adopted.

the foil of a previous paradigm cannot stand as an appropriate foundation for the unification of a diverse set of distinct disciplines, even if it helps a nascent framework develop. So what provided unification?

1.1 The Birth of Cognitive Science

In their wonderful book, *How the Body Shapes the Way We Think*, Rolf Pfeifer and Josh Bongaard (2007) discuss the inception of the field of artificial intelligence, which the authors claim is best viewed as commencing with the Dartmouth conference in 1956 where the "fathers of AI" such as, Marvin Minsky, John McCarthy, Allen Newell, Herbert Simon, and Claude Shannon, discussed "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." (Dartmouth Artificial Intelligence Project Proposal, McCarthy et al., Aug. 31, 1955, cited in Pfeifer and Bongaard, 2007, p. 28). However, 1956 was not just a pivotal year for artificial intelligence. Research in AI had a number of early successes with information processing and abstract symbol manipulation, which many of those attending the conference had been involved in formalising. Due to this success the idea that intelligent behaviour could be thought of as rule-governed symbol manipulation was quickly picked up by other fields. George Miller—an influential psychologist during these formative years—singles out 11th September 1956 as "the moment of conception of cognitive science" (Miller, 2003, p. 142). It was the day of a symposium organised by the 'Special Interest Group in Information Theory' at the Massachusetts Institute of Technology. Instead of being restricted to artificial intelligence, this symposium gathered researchers from psychology, linguistics, artificial intelligence and neuroscience, and Miller states,

> "I left the symposium with a conviction, more intuitive than rational,

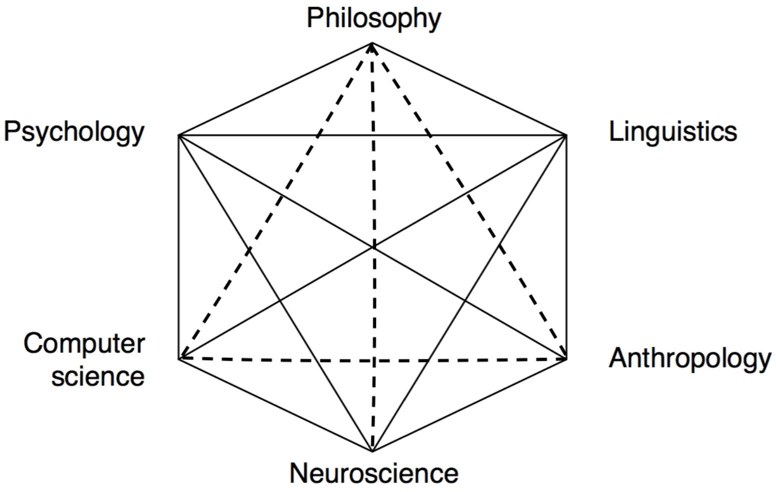


Figure 1.1: Solid lines represent relations that Miller claims were established on the basis of a respectable research programme in the late 1970s. Dotted lines represent connections that have been made since (e.g. relationship between philosophy and the neurosciences). Adapted from (Miller, 2003).

> that experimental psychology, theoretical linguistics, and the computer simulation of cognitive processes were all pieces from a larger whole and that the future would see a progressive elaboration and coordination of their shared concerns." (ibid., p. 143)

His conviction was well-founded, and following this pivotal year, the interdisciplinary project of cognitive science took off. The interdisciplinary ties at the time, were characterised in a report by Miller, which claimed the relations between the various disciplines were those depicted in Figure 1.1. These relations were characterised by the methodological tools shared by the respective disciplines, which in turn led to novel interdisciplinary frameworks (e.g. neuroscience and computer science could be connected by approaches such as cybernetics). Miller acknowledges that nowadays all 15 possible links could be argued to exist, on the basis of respectable research programmes. But is there a non-trivial assumption that is shared by the various disciplines?

1.2 Cognitivism

Although the foil of behaviourism, as previously discussed, is not an appropriate unifying thread on its own, it does point us towards a reason for the emergence of cognitive science as an interdisciplinary project in 1956.

In discussing a conversation that took place with Noam Chomsky, Miller claims that "defining psychology as the science of behavior was like defining physics as the science of meter reading. If scientific psychology was to succeed, mentalistic concepts would have to integrate and explain the behavioral data" (ibid., p. 142, emphasis added). Psychological behaviourism is often seen as eschewing the use of mental states as theoretical posits.[2] As the previous quotation highlights, in the 1950s this methodological limitation was becoming a contentious issue for those who desired a truly explanatory and integrative framework. The response to this limitation, has since become known as one of the defining features of *cognitivism*: the classical view of cognitive science that emerged in the late 1950s.

The proponent of *cognitivism* views cognition as a system of internal, brain-based processes that are formed on the basis of sensory input, and stored as abstract symbols, which can then be transformed in a deterministic manner for the control of motor behaviour (Neisser, 1967). Cognition is, therefore, a form of information-processing, and in line with the historical considerations above, the dominant metaphor of this perspective views the brain as some sort of computational system. The computations are taken to operate over a set of stored symbols, which stand-in or

[2] Though see (Barrett, 2012) for a less dogmatic reading.

represent some object or event. These symbols function both as perceptual and conceptual stand-ins for the world, and when appropriately transformed, as instructions for motor behaviour. Therefore, they are understood as the internal *causes* of behaviour, and are not identical with behaviour itself. This leads to a rejection of the behaviourist's guiding methodological and philosophical perspective. As Barrett (2012, p. 18) highlights, "[a]lthough internal rules and representations are not available for direct inspection, they can, however, be inferred, via observation and experiment, from the behaviour they cause." As such, the cognitivist framework allowed researchers and experimenters to posit explanatory representational states in order to overcome the limitations of the behaviourist paradigm and to focus on an abstract problem-solving characterisation of cognitive tasks. With a sufficient level of abstraction, these tasks could in turn be implemented in the so-called "expert systems" of artificial intelligence (Pfeifer and Bongard, 2007, p. 27). Such systems are indicative of the classical approach to AI, or in Haugeland's (1985) terminology, "Good Old-Fashioned Artificial Intelligence".

This classical perspective places a particular conception of human intelligence centre-stage, and focuses on readily formalisable processes such as natural language, knowledge acquisition and representation, formal reasoning, and playing games such as chess—all areas that were amenable to the methods being discussed at the aforementioned Dartmouth conference. The expert systems of AI were positioned to replace human experts in these circumscribed domains, and when coupled with the philosophical theory of functionalism, this classical cognitivist approach flourished. As Thompson (2007, emphasis added, p.5) states:

> "Cognitivism goes hand in hand with functionalism in the philosophy of mind, which in its extreme computational form holds that the *embodiment* of the organism is essentially *irrelevant* to the nature of the mind. It is the software, not the hardware, that matters most for mentality."

As with any good philosophical debate, opposition quickly emerged, taking issue with the thoroughly brain-bound position of mind and cognition that was emerging. Understanding the correct target of this opposition, however, first requires us to make a number of conceptual distinctions and subsequent clarifications about terminological usage. The following sections will briefly explore these foundational issues, but will fall short of providing a complete critical analysis.

1.2.1 The Computational Brain

To begin, we can distinguish the notion of *computation* from *computationalism*, where the former refers to a formalisable concept that is well-defined by the theory of computation, and the latter, by contrast, refers to the view that the mathematical apparatus of the theory of computation can be applied to model various cognitive phenomena. For example, *neurocomputationalism* would be the more exclusive view that the activity of neurons can be modelled (and subsequently explained) as computing in the sense defined by the theory of computation.

This distinction may seem straightforward, but is complicated by a number of conceptual worries. Firstly, though the notion of *computation* is well-defined mathematically, it is an abstract definition that is independent from any particular physical system. Therefore, its usage often varies depending on the context of discussion, and can be viewed as highly restrictive (e.g. language of thought hypothesis (Fodor, 1975)) or fully general (e.g. pancomputationalism (Putnam, 1991)). This is important, as many disagreements turn on whether a particular theory in the cognitive sciences is computational or not, but if the notion is being employed without a fixed or precise meaning, then it will be hard to settle the debates. For example, it is common to hear that classical cognitive architectures are differentiated from later connectionist architectures due to their computational commitments:

> "Connectionism gives solace [...] to philosophers who think that relying on the pseudoscientific intentional or semantic notions of folk psychology (like goals and beliefs) mislead psychologists into taking the computational approach [...]" (Fodor and Pylyshyn, 1993, p. 4)[3]

Therefore, we need to provide a way of understanding the notion of computation, such that we can determine whether several theories are truly in conflict with one another. So, what is computation?

The Theory of Computation

In his (1936/37), Turing set out the now-famous notion of a Turing machine, with the aim of providing a method for answering the Entscheidungsproblem: the *decision problem* of whether an *effective method* could be found for determining the provability of statements expressed in first-order logics. Along with Alonzo Church, Turing established that no effective method could in fact be determined, but the notion of an *effective method* (or algorithm), for determining the solution to a function (i.e. the output value of a function for a particular input), led to the formation of the Turing machine: an abstract mathematical device or system that functions in an algorithmic manner. In short, any function that could be solved by an algorithmic process (i.e. a set of rules that necessarily lead to a solution) would be computable by a Turing machine, or using the equivalent notion of *recursion* from Church's (1936) lambda calculus. Turing noticed that the two formal definitions provided by himself and Church were coextensive with the informal notion of computation as effective

[3] It is important to note, that Fodor and Pylyshyn (1993) did not believe that connectionism was non-computational, but differed from classical cognitive architectures due to its rejection of the latter's commitment to discrete symbolic processes, which were the atomic digits that the computations were defined over.

method, leading to the formation of what is now known as the Church-Turing thesis. We will here focus on the notion of a Turing machine for simplicity.

A Turing machine has two components: a (potentially) infinite tape, and a finite-state machine with a read/write head. The tape is divided into finite cells, on which there is written a symbol (taken from a finite alphabet), or simply a blank space. This tape serves as both the vehicle for input into the system, and as a working memory store for the computational process. The finite-state machine positions its read/write head over each cell, and then depending on a) the current state of the machine, b) the symbol on the cell being read, and c) the rule specified by the machine table (or programme in the case of a universal Turing machine), the finite-state machine can perform a number of operations (i.e. move left or right, or write a new symbol on the tape). An explanation of how this abstract machine is able to compute functions need not concern us.[4] More important is the fact that this description is necessarily abstract, and thus independent of any physical implementation.

Any system that is able to realise the functional roles specified by the above components, can be considered computational in a generic sense. This leads to an often overlooked aspect of Turing's work, which both Wells (2005) and Anderson (2014) have emphasised as important, and is evident in the following quotes:

> "Computing is normally done by *writing certain symbols on paper*." (Turing, 1936/37, emphasis added, p.249)

A 'computer' in Turing's time was a human, employed to perform laborious calculations by following algorithmic processes by writing on a piece of paper. The abstraction away from these details, into the suggestive notion of a Turing *machine*, was supposed to bring greater generality to the notion of computation, while retaining the emphasis on an algorithmic process. The generality was important, as the

[4] See (Clark, 2013a) for a simple introduction.

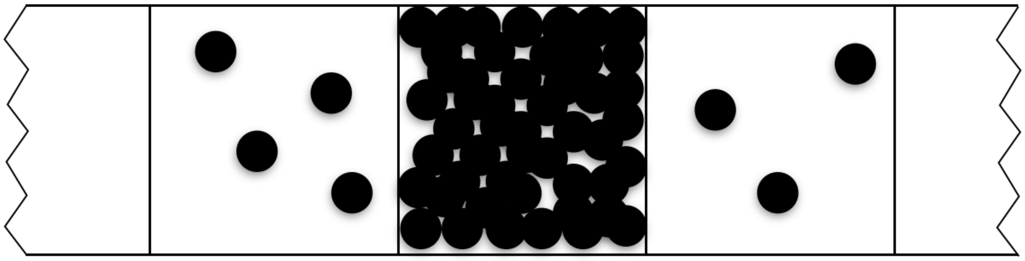


Figure 1.2: A portion of a tape in a Turing machine.

continuation of the above quote highlights:

> "We may suppose that there is a bound B to the number of symbols of squares which the computer can observe at one moment. If he wishes to observe more, he must use successive observations." (ibid., emphasis added, p.250)

Unlike an abstract Turing machine, a human observer realising the functional role of one is limited by their perceptual apparatus, not merely in the temporal sense that Turing highlights, but also in the level of discernible details that they are receptive to. Imagine you are instructed to compute in the sense defined by a Turing machine. The cells on the tape can contain a number of dots, which are to be processed according to a finite set of instructions laid out in the machine table. You proceed as usual, until you come to the cell depicted in Figure 1.2. Such a cell would obviously be a terrible design choice for a human computer as it is too difficult to determine the exact number of dots.

Far from simply being a limitation for human computers, however, the above example highlights a technical challenge that all engineers working in information and communication technologies will be acutely aware of—the problem of signal noise. That is, how to ensure that the range of symbols a machine must compute is not so large, or too fine-grained, that the receiver (or read/write head in Turing's terminology) is unable to distinguish between two distinct states. When dealing with analog currents of electricity as the vehicle for communication, one solution is to map ranges of the analog signal onto a restricted set of discrete (or quantized) states. As far as the receiving system is concerned, this discreteness need only exist at the level of symbols being computed, the underlying vehicle may in reality be a continuous (or analog) input. This means that understanding the commitments of a computational theory requires us to understand the claims made about the properties of the symbolic representations, as well as whether they are neurally-plausible in terms of their implementation. Why is this important?

Piccinini and Scarantino (2010) argue that the notion of computation outlined by a Turing machine is an example of what they term *digital computation*. They state:

> "[...] the computation of a Turing-computable function is a digital computation because Turing-computable functions are by definition functions of a denumerable domain—a domain whose elements may be counted—and the arguments and values of such functions are, or may be represented by, strings of digits."

This can be contrasted with what they term *generic* computation, and *analog* computation. They define the former as the processing of vehicles according to rules that are sensitive to certain properties of the vehicles and, specifically, to differences between portions of the vehicles.[5] This is general enough to include both digital

[5] Given the fact that generic and analogue notions of computation need not take strings of discrete symbols as inputs, the more encompassing notion of 'vehicle' is adopted to account for this greater level of generality. A vehicle therefore refers to any input that is computed by one of the notions of computation.

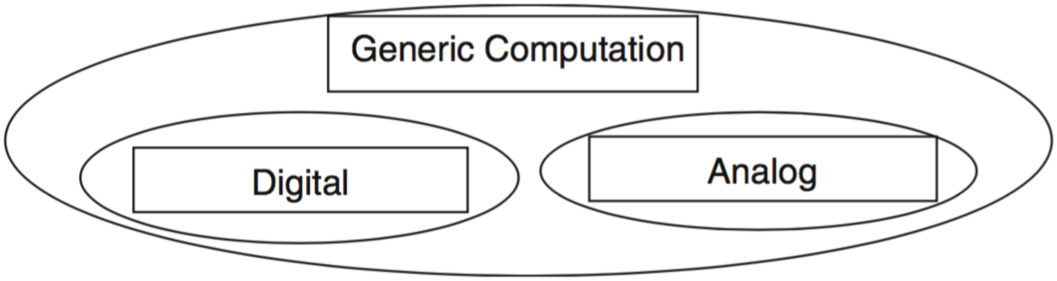


Figure 1.3: Types of computation. Reprinted from (Piccinini and Scarantino, 2010).

computation and analog computation, the latter of which they describe as the manipulation of continuous variables, which can vary continuously over time and take any real values within certain intervals—it is therefore uncountable and differentiated from digital computation. The relationship between the different notions is depicted in Figure 1.3.

These distinctions are important to note, as when proponents of computationalism make claims about cognition being a form of computation, depending on the type of computation being discussed, we should expect corresponding differences in the properties of the vehicles being processed.[6]

Computationalism

With the necessary distinctions in place, we can now turn to the notion of *computationalism*, which we previously defined as the view that the formal notion of computation can be applied to model various cognitive and neural phenomena. One of the first attempts to do this was put forward by McCulloch and Pitts (1943), who

[6] More in-depth details of each subset of computation is explored further in (Piccinini and Scarantino, 2010). However, as this thesis is not directly concerned with whether any of these particular notions is applicable to the brain and cognition, further discussion is unnecessary.

noted that neural activity, due to its "all-or-none" character, could be described as a type of digital computation. By "all-or-none", they were referring to the observation that an action potential (or spike) produced in the soma of a neuron, was independent of the strength of incoming signals from neighbouring neurons. In short, if the threshold for the production of an action potential is sufficiently met, the strength of the stimulus is irrelevant—either it occurs or it does not.

They argued that this property supports an argument that what the brain does is best understood as digital computation, as the spikes can be considered the basic atomic components of the strings of symbols that form the inputs to the computational process. However, not long after the publication of their work, others argued that this picture overlooked the important role that neurotransmitters (and the endocrine system) have on neural activity. For example, (Gerard, 1951, p. 12) stated:

> "[...] chemical factors (metabolic, hormonal, and related) which influence the functioning of the brain are analogical, not digital. What is perhaps not fully recognized is the tremendously important role that these play not only in the abnormal but also in the perfectly normal functioning of the nervous system."

Here we see a dispute that may potentially be resolved by means of accumulating empirical data (e.g. what processes are in fact responsible for regulating neural activity). If, as McCulloch and Pitts (1943) argue, the action potential is the primary vehicle of computation, and is importantly digital rather than analogue, then perhaps the brain is a digital computer. However, if the vehicles of computation extend to include analog processes, then perhaps (Gerard, 1951) is correct. Piccinini and Scarantino (2010, p. 12) are ambivalent with respect to this debate, but nevertheless maintain that:

> "[...] current evidence suggests that the vehicles of neural processes are

> neuronal spikes and that the functionally relevant aspects of neural processes are medium-independent aspects of the spikes—primarily, spike rates. [...] spike trains appear to be another case of medium-independent vehicles, in which case they qualify as proper vehicles for generic computations. Assuming that brains process spike trains and that spikes are medium-independent vehicles, it follows by definition that brains perform computations in the generic sense."

This position could be challenged, as there are a number of ways in which it is possible to encode a pattern of spikes (e.g. a rate code (average production of spikes), a timing code (specific timings of spikes), a population code (population-wide groupings of neural spikes), and a synchrony code (synchronicity across neurons) (Eliasmith and Anderson, 2003, p. 7)). This is likely the reason why Piccinini and Scarantino (2010) adopt the weakest notion of *generic computation* in order to remain ambivalent. However, is this move entirely appropriate?

Note that what is being defended is a notion of *neurocomputationalism*, which we defined earlier as the view that the activity of *neurons* can be modelled as computing in some sense (whether digital, analog or merely generic). However, it is not immediately evident what would be wrong with accepting this position, but choosing to remain agnostic with regards to whether the *mind* should therefore be understood in computational terms as well. Perhaps one could appeal to something like the *personal/sub-personal* distinction (Dennett, 1969), or the notion of sub-doxastic states (Stich, 1978) to justify an explanatory pluralism. If so, then we could retain the explanatory power of computational modelling as it applies to the neural level, which has been enormously productive in the field of computational neuroscience, without being limited by the constraints of the computer metaphor at other levels of explanation. Although we will not directly argue in favour of either position, it is important to understand this debate in order to subsequently understand the lim-

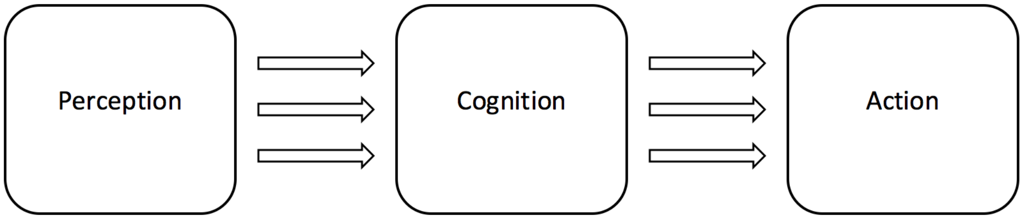


Figure 1.4: The Classical Sandwich Model.

itations of adopting a computationalist account of the mind. As a bridge into this discussion, it is helpful to turn to the work of Susan Hurley (1998), and her critique of the "classical sandwich model" of the mind.

1.2.2 The Classical Sandwich

Hurley's purpose in describing the 'classical sandwich model' was to explore the mainstream (at her time of writing) view of the mind (cognitivism), which she claimed was committed to a couple of things.

Firstly, it took perception and action to be distinct processes, separate from one another and peripheral to the inner symbolic processing associated with cognition (see Figure 1.4). Perception, as described by this view, is an incremental process by which the detachable symbols used in computation are produced in a serial manner, first detecting simple properties such as lines or edges and progressively building up to richer, more complex representations of the environment. Action is the end product of the transformative, algorithmic process, whereby the representational symbols are combined with various goal representations such as desires, in order to guide behaviour.

Secondly, this implies that cognition, as a process of symbolic manipulation, was "virtually central, even if the mere implementation of cognitive processes is distributed"[7] (ibid., p. 401). Furthermore, perception and action are not simply separate from each other, but also separate and encapsulated[8] from cognition, and therefore if the mind is a product of these encapsulated processes, "[t]he mind is a kind of sandwich, and cognition is the filling" (1998, p. 401).

Hurley took the 'classical sandwich model' to highlight three important questions. First, is cognition 'classical' in the sense of being a computational process operating over symbolic representations? Second, is cognitive processing central and distinct from perception and action? Third, are perception and action separate from each other? We have already touched upon some of the difficulties of settling the first question. However, it may arguably be easier to answer it when couched in terms of classical computation, rather than computation *tout court*, due to the additional constraints placed upon the types of vehicles that the former postulates (e.g. strings of symbols with sentence-like properties) (Fodor, 1975). We will return to this question shortly.

Answering the other two questions satisfactorily will require a more prolonged discussion, which will require the remainder of this thesis to achieve, even in the restricted case of decision-making. However, it is possible to make a few tentative remarks at present. The argument to be defended in this thesis is that current theoretical and empirical evidence supports a negative answer to the latter two

[7] Important to note, is that although often seen as critiquing classical cognitivism in the sense defended by (Fodor and Pylyshyn, 1993; Fodor, 1975), Hurley's addition of the possible implementation of cognitive processes being distributed, means that some forms of feed-forward *connectionist* architectures would also captured by the sandwich model.

[8] The term encapsulated may be unfamiliar to some readers. It is commonly used in discussions of modularity and cognitive penetrability (Fodor, 1983), where the mind is thought of as realised by distinct neural structures or modules, which have specific functions. These modules are considered encapsulated if their functions are cognitively impenetrable or unaffected by other cognitive domains.

questions—cognition is not central and distinct from perception and action, and neither are perception and action separate from one another. However, adopting this view is not grounds for dismissing the idea that some processes in brain, at the right level of description, are computational. Many have seen the rejection of cognitivism to be a rejection of its computational underpinnings. However, this move would be too rash, and is likely a result of too monolithic an understanding of cognitivism. It is important to bear in mind that even if we can reject the claim that cognition is not classical computation, it may still be possible to argue for some form of *neurocomputationalism*, or a different notion of cognitive computation altogether. As we will see in the next section, what aspects of cognitivism are rejected by embodied cognition is not always shared equally among the various positions.

1.3 Embodied Cognition

Even the most cursory glance at the literature on embodied cognition is sufficient to instill the idea that the label 'embodied cognition' is employed in a number of ways, and to sometimes incompatible ideas or methodologies. To make matters more confusing, there has been a recent tendency to move away from the label of 'embodied cognition' to the more encompassing label '4e approaches to cognition' (Barrett, 2015; Hohwy, 2016; Menary, 2010). Here, '4e' highlights the conjunction of four views known as: embodied, embedded, extended and enactive (sometimes another term 'affective' is included, despite its unwillingness to conform to the pattern).

In order to bring cogency to the discussion, some have attempted to specify the themes or commitments of embodied cognition. Wilson (2002), for example, claims that it is composed of six views (here paraphrased):

1. Cognition is situated in the context of a real-world environment.

2. Cognition is time-pressured, and should be understood in terms of how it functions under these constraints.

3. Cognition is offloaded on to the environment as a result of information processing constraints of the organism.

4. Cognition is constituted in part by the environment, and the idea of a brain-bound mind is therefore an inappropriate object of investigation.

5. Cognition is carried out primarily for action and the guidance of an organism in its environment.

6. Off-line cognition is body-based and grounded in mechanisms that primarily evolved for the guidance of action.

The explanatory strength of these views is varied, as is the amount of support that each receives. For the time being, it is worth highlighting that, unlike cognitivism, embodied cognition is still in its *relative* infancy, and a guiding set of principles on which practitioners and advocates agree is yet to form.[9] An understanding of the reasons behind this lack of conceptual and methodological structure must certainly come from recognising the interdisciplinary nature of the cognitive sciences, and the multitude of explanatory interests that are to be found within the many disciplines. However, can we be more specific?

[9] It would not be inappropriate to say that *embodied cognition* represents a revolutionary turn away from the normal practice of cognitive science qua cognitivism, in the sense expressed by Thomas Kuhn's (1962) famous account of scientific practice. It is also fair to state that it has not yet acquired the status of 'normal science'.

1.3.1 Post-Cognitivism: The Symbol Grounding Problem

In a similar vein to the cognitivist's response to the limitations of behaviourism, embodied cognition also emerged as a result of dissatisfaction with some of the conceptual challenges raised by its predecessor (cognitivism). An oft-cited problem in the context of this conceptual shift is the *symbol grounding problem*—in short, the problem of determining the meaning of abstract symbols.

To emphasise this problem, Harnad (1990) explores some of the commitments of cognitivism in relation to explaining and understanding 'physical cognitive systems' (he adopts Newell's (1980) term "symbol systems" in lieu of cognitive systems):

> "A symbol system is:
> (1) a set of arbitrary physical tokens (scratches on paper, holes on a tape, events in a digital computer, etc.) that are
> (2) manipulated on the basis of explicit rules that are
> (3) likewise physical tokens and strings of tokens. The rule-governed symbol-token manipulation is based
> (4) purely on the shape of the symbol tokens (not their "meaning"), i.e. it is purely syntactic, and consists of
> (5) rulefully combining and recombining symbol tokens. There are
> (6) primitive atomic symbol tokens and
> (7) composite symbol-token strings. The entire system and all its parts—the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules—are all
> (8) semantically interpretable: The syntax can be systematically assigned a meaning (e.g. as standing for objects, as describing states of affairs)."
> (ibid., p. 336)

Harnad states that for cognitivists such as (Fodor, 1975; Pylyshyn, 1984) these commitments also apply to mental phenomena such as beliefs and desires, and not merely neural processes. The problem with these commitments is that they lead to a problem of 'content determination', whereby it is hard to see how the symbols can be said to refer intrinsically to any states in the world, or to paraphrase, how the symbolic representations can mean anything without the presence of an external observer viewing them. This is because by (1) and (4), the symbols are necessarily arbitrary, and manipulated according to just their syntactic properties. Furthermore, as a result of (2), (3), (5), (6) and (7), all parts and processes of the system are accounted for in this manner, such that semantic meaning cannot creep in artificially at some higher-level. Therefore, and in accordance with (8) it is only the system, taken as a whole, which can be *interpreted* semantically, and this necessarily requires an observer. However, though this may be satisfactory for cognitive systems such as computers, it seems to leave open the well-known question of how systems such as ourselves can be said to have mental representations that carry non-derived content.

To illustrate this problem, Harnad describes the problem of an agent with no knowledge of the Chinese language, attempting to determine the meaning of Chinese symbols using only a Chinese-Chinese dictionary. In a manner that is reflective of Searle's (1980) famous Chinese room thought experiment, Harnad (1990, p. 340) argues that it is difficult to see how one could "ever get off the symbol/symbol merry-go-round", and how symbol meaning is "grounded in something other than just more meaningless symbols?" This is the worry that has come to be known as the symbol grounding problem:

> "How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?" (Harnad, 1990, p. 335)

In other words, how can thoughts acquire meaning if they are simply arbitrary strings of symbols, related to other arbitrary symbols. Meaning cannot arise in this manner; it must be grounded in something else. Defenders of the cognitivist framework have attempted to provide various responses to this challenge, such as positing some sort of casual-dependancy relation between the symbol and the referent in the world (Fodor, 1992), or through an information-theoretic account of reliable correlation (Dretske, 1981). However, one of the motivations for turning to an embodied framework is that this issue doesn't arise for an embodied account (Robbins and Aydede, 2009).[10]

Proponents of embodied cognition claim that the way around the symbol grounding problem is to claim that meaning is necessarily grounded in facts about the agent's embodiment. The manner in which this is explained can vary. Some begin by arguing for a pragmatic solution, such that if an agent is capable of exploiting the incoming sensory information in an appropriate way (e.g. fulfilling its purpose of acquiring food) then this is all that needs to be said of meaning, as the agent has "understood" the meaning of the sensory stimulation (Pfeifer and Bongard, 2007). However, this only pushes the problem back a step, as we then need to account for the notion of purpose, in order to determine when a representation is *appropriate*,

[10] It should be noted that many varied attempts in the philosophy of mind literature have tried to solve the problem of meaning, or the problem of representation more generally. In fact, an incredible amount of time and research has been dedicated to the topic of mental representations since the decline of psychological behaviourism (Anderson and Rosenberg, 2008; Brooks, 1991; Clark and Toribio, 1994; Cummins, 1989; Dretske, 1981; Field, 1978; Fodor, 1992; Grush, 2004; Millikan, 1995; Ramsey, 2015). This makes a summary of the highly disparate viewpoints so challenging as to seem like a fool's errand unless fundamental to the task at hand. Simply put, there is not enough space to rehearse them here, and the aim of this thesis is not to decide, which of them is most likely correct. Instead, I have chosen to use the symbol grounding problem as a bridge between cognitivism and embodied cognition to highlight one (of several) contrastive areas between the two frameworks.

or when it is inappropriate, perhaps due to mis-representing the object in some way. Others emphasise the situated (or embedded) nature of cognition[11], leading to the claim that cognition is dynamic, and always unfolding in the environment, such that the meaning of any mental state is determined naturally during agent-environment interaction (we will explore some examples of this view shortly).

One of the most influential accounts comes from the work of psychologist Lawrence Barsalou (1999) and his work on perceptual symbols (also see Barsalou, 2008; Prinz and Barsalou, 2000). The notion of perceptual symbols draws a contrast between the arbitrary, *amodal* symbols postulated by classical cognitivism, and the *modal* symbols of Barsalou's embodied account. When an agent perceives the world, they do so through different *modes* of sensation (e.g. vision, audition, proprioception), though the perception of token objects need not be presented in all of the possible modes (e.g. I can hear a bird singing without being able to see the bird). The cognitivist would require that these perceptual signals from the world are transduced, at the point of contact with the various sensory receptors, into some neutral or amodal code (akin to something like the binary code of a computer), in order to be transformed, combined and possibly decoded and recoded into some action plan—Barsalou claims this is unnecessary. Instead, he argues, "cognition is inherently perceptual, sharing systems with perception at both the cognitive and the neural levels." (Barsalou, 1999, p. 577) During perceptual experience, he claims, the task of the brain is to capture bottom-up patterns of neural activation in sensorimotor regions, so that these patterns can be stored as modal symbols, and later reactivated (or simulated) for a wide-range of cognitive tasks (e.g. conceptualisation, reasoning, decision-making etc.). Importantly this simulation redeploys the same sensorimotor regions (or parts of them) that were initially activated during perception. For example, thinking of

[11] The distinction between situated and embedded cognition is sometimes collapsed, though see (Robbins and Aydede, 2009) for an account of their differences.

kicking a ball, or even the word ‘kick’ would re-activate the same sensorimotor regions that would have been involved in past experiences of kicking a ball (Hauk, Johnsrude, and Pulvermüller, 2004; Pulvermüller, 1999).

Three points are noteworthy. Firstly, the emphasis that Barsalou places on the use of *symbols* should give us pause to consider what exactly is being rejected in the cognitivist picture. Barsalou explicitly states that “traditional approaches (i.e. cognitivism) are correct in postulating the importance of symbolic operations for interpreting experience” (Barsalou, 2008, p. 622). Although this may allow for the retention of computational processes, the arbitrariness of the amodal symbols cannot be retained—a new account of representation is therefore required. Secondly, when Barsalou claims that cognition is inherently perceptual, and involves the reactivation of stored perceptual symbols, he is also denying the centrality of cognition. This means that, contra the ‘classical sandwich model’, cognition is not an encapsulated process, separate from perceptual or motor systems in the brain, but is instead intertwined with, or *constituted* by, sensorimotor activity. Finally, by arguing that these symbols are necessarily perceptual, their meaning is grounded in facts about the agent’s embodiment. For example, the type of perceptual capacities that the agent has will necessarily shape the thoughts it can have, as well as the content they carry—a congenitally blind person will not have the same set of thoughts as a fully sighted individual; a chameleon will not have the same neural representations as a fly. This provides important explanatory constraints on the sorts of theories that the cognitive sciences should pursue.

Irrespective of how the symbol grounding problem is approached[12], the empha-

[12] For example, Chemero (2011) adopts a Gibsonian notion of affordances, and argues for a more direct approach to understanding perception, cognition and action. This approach eschews the notion of representations in favour of dynamical modelling, whereas Hutto and Myin (2013) take an even more radical perspective and try to dismiss the problem of content altogether by eliminating

sis is on the agent's embodiment and dynamic interaction with the environment. This requires understanding cognition and behaviour as arising not from within an encapsulated central system that merely crunches symbols, but as a dynamic and interactive process that is inseparable from the embodied behaviour of the agent. Furthermore, it requires understanding the representational (and indeed conceptual) capacities of the agent as being fundamentally shaped by the two aforementioned processes. Here we reach a potential source of disagreement among embodied theorists: are all of these themes necessary conditions?

1.3.2 Three Themes of Embodiment

Shapiro (2011) aims to capture the idea of embodiment by outlining three themes:

Conceptualisation: an organism's acquisition and use of conceptual knowledge, on which it relies to understand its world, is determined in some sense by the (dynamic) properties of its body and sensory organs.

Replacement: mental representations, which were ubiquitous in cognitivist explanations of behaviour, are inappropriate and in some cases misleading tools for understanding an organism's skilful interaction with the world. They should be replaced by alternative tools, such as dynamical systems theory, which emphasise the coupled, reciprocal nature of extended systems, rather than brain-bound processes.

Constitution: the constituents of the mind might comprise objects and properties apart from those found in the head, which were traditionally understood as mere causes of cognition.

it from their framework entirely.

As with Wilson's six views, these three themes are also accepted and dismissed to varying degrees throughout the community of self-described embodied cognition researchers, with some ongoing disagreements arising about whether they are all compatible with one another. Let's explore each in turn.

Conceptualisation

The term 'embodied' for the proponent of the conceptualisation thesis means two things. First, cognition and thought depend upon the kinds of experience that arise from the possession of a body with various sensorimotor capacities—a type of embodied action. Therefore, understanding cognition requires understanding the capacities of the body. For example, the properties of the visual system or the range of actions afforded by the agent's body (e.g. gripping or flying). Second, these capacities are to be understood as being embedded in a wider environmental context, which is taken to include a biological and socio-cultural context. This means exploring not only how an agent interacts with its local, physical environment, but also its evolutionary lineage and any relevant socio-cultural norms. Shapiro cites Varela, Thompson, and Rosch (1991) (VTR) as being a prototypical case of the conceptualisation thesis.

For VTR, understanding cognition requires understanding embodied action. They claim that "sensory and motor processes, perception and action, are fundamentally inseparable in *lived cognition*" (ibid., p. 173). The emphasis on lived cognition, is motivated by Varela's earlier work with Humberto Maturana (Maturana and Varela, 1980), in which they outlined the *theory of autopoiesis*. In short, the theory was concerned with the dynamic, *self-producing* processes that sustain life.[13] The term 'autopoiesis' was coined to stand-in as a label for the processes of circular-organisation, which they argued constitute the basis of life, and within which they situated their

[13] We will have more to say of the theory of autopoiesis in chapter 4.

understanding of cognition as a fundamentally lived experience. Within this framework, perception and action not only enable an agent to successfully interact with their environment, but also change the agent's experience. For example, as I move throughout the world, I open up new possibilities for perceptual experience, and simultaneously close off others. Action determines new perceptions, which in turn disclose possible future actions, which in turn determine further perceptions, which in turn... you get the idea! However, this intertwined nature of embodied action is further shaped by the various properties of the agent's body. One needs only think of a non-human animal with a radically different perceptual (or motor) system to our own (e.g. monocular versus binocular vision) to appreciate the truth of this statement.

This idea was explored by O'Regan and Noe (2001), who introduced the notion of *sensorimotor contingencies*, and were also influenced by the work of James Gibson (1979) and the *ecological approach* to psychology. Central to their view was the acknowledgement that there is more information available in the environment to an organism that is capable of interaction than is available to a purely passive perceiver. Active perceivers can detect invariant features in the dynamics of sensory input, relative to their interactions with the environment. For instance, as an organism moves directly towards an object, it will increasingly fill a larger portion of the visual field relative to the speed and movement of the agent. This occurs when the object is stationary with respect to the active perceiver, and thus provides additional information about the world. An active perceiver can exploit these reliable properties of sensorimotor interactions in order to learn about the features of the world, which would otherwise require inferential processes (Gregory, 1980). These predictable relationships between action and perceptual input are what they term *sensorimotor contingencies* (O'Regan and Noe, 2001).

A further feature of the ecological approach to perception is to highlight the action-oriented nature of perceptual processes. On a more traditional theory of perception, information only becomes available for the guidance of action once a perceptual representation has been formed and passed on to cognitive systems. However, this need not be the case for active perceivers. This is because the kinds of complex invariant features that can be detected by an active perceiver have immediate relevance for action. Rather than first perceptually representing external objects and then inferring the consequences for action, active perceivers are able to directly perceive *affordances*, which are best understood as opportunities for action (Gibson, 1979).[14]

As described, the first aspect of the conceptualisation thesis (outlined at the start of this section) may seem relatively uncontroversial, but difficulties arise with the second aspect. The issue arises when we try to understand the nature of the embedding relation. Why is it important to understand the capacities as *embedded* in a wider context? Shapiro points to a section of VTR's work, where they discuss two positions:

Chicken Position: The world out there has pregiven properties. These exist prior to the image that is cast on the cognitive system, whose task is to recover them appropriately (whether through symbols or global subsymbolic states).

Egg Position: The cognitive system projects its own world, and the apparent reality of this world is merely a reaction of internal laws of the system. (Varela,

[14] There are a number of metaphysical complications that arise on closer inspection of the notion of affordances. Rather, than dealing with this huge literature, we will instead use the term *action opportunities* throughout the thesis to distance our view from that of Gibson's. We acknowledge that plenty of conceptual challenges remain, but regretfully do not have the space to deal with them directly in this thesis. Chemero (2011) provides a useful summary of some of the challenges with Gibson's theory of affordances, and discusses a significant portion of the literature that has subsequently been published in response.

> Thompson, and Rosch, 1991, p. 171, cited in Shapiro, 2011)

These positions can be seen as endorsing realism and idealism about properties of the world respectively. But which comes first in the case of perceptual experience for VTR?

The chicken position is untenable, as perception of the world depends on the sensorimotor capacities of the organism. However, the egg position is also untenable if we acknowledge the evolutionary factors that shaped our sensorimotor capacities over time—denying a mind-independent reality, as idealism would have us do, is inconsistent with this. Therefore, VTR wish to collapse the distinction between these two positions, and allow that biological and socio-cultural factors determine and shape our experience (denying idealism), while also maintaining that all perceptual experience of the world is necessarily organism-relative (denying realism). To do this, they argue, requires understanding the capacities of an agent as necessarily embedded in its environment—always and everywhere inseparable from it, unless we want to risk collapsing back into one of the aforementioned positions.

These topics have been explored in depth by more recent work in enactivism (one of the 4E approaches), and are certainly worthy of continued investigation (Noe, 2004; Thompson, 2007). In addition, the work of Lakoff and Johnson (1980) has been of considerable interest to researchers in fields such as anthropology and linguistics in exploring the way our bodies and their dynamics alter and shape our conceptual and linguistic practices (Boroditsky and Gaby, 2010; Casasanto, 2009).

Replacement

The second theme of embodied cognition is primarily concerned with the methodological tools used to model and explain cognition. Most notable in this regard is the use of dynamical systems theory (a mathematical theory that describes how rule-governed systems change over time) as a *replacement* for the symbolic representations of cognitivism. As such, cognition is modelled as a dynamic process that is closely coupled with its environment, and the behaviour of cognitive systems should be understood in a similarly dynamic manner. This involves the use of differential equations that describe the continuous changes in the state of a system—the complete map of these changes is accordingly known as the the *state space*. Three points need to be made explicit regarding the notion of replacement and dynamical systems: the emergence of *self-organisation* as a property of dynamical systems, the idea of *coupling*, and the commitment to *antirepresentationalism*.

A number of researchers defend what Shapiro calls the replacement theme. Of note are developmental psychologists Esther Thelen and Linda Smith (1994), roboticists Randy Beer (2000) and Rodney Brooks (1991), and philosopher Tony Chemero (2011). Rather than looking at specific cases, it will be more instructive to see what is common to their approaches. For this we can turn to the prototypical example of dynamical systems presented in (Kelso, 1995): Rayleigh-Bénard convection.

As with the aforementioned dynamicists, Kelso believes dynamical systems theory to be a superior alternative to cognitivism. He states:

> "This is an entirely different image from the brain as a computer with stored contents or subroutines to be called up by a programme. In nature's pattern-forming systems, contents aren't contained anywhere but are revealed only by the dynamics. Form and content are thus inextricably connected and can't ever be separated." (ibid., p. 1)

As an example of these "pattern-forming" systems, he describes the simple system of oil being heated from below and cooled from above.[15] When the oil is heated weakly from below there is no large-scale motion. The oil is in a rest state as the

[15] He is also careful to highlight that the simplicity of this example is no guide to the complexities

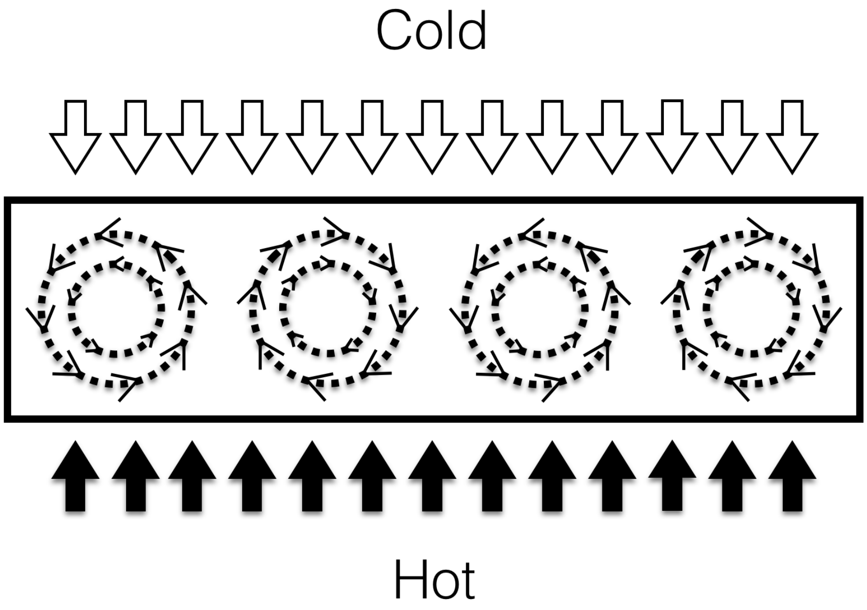


Figure 1.5: Rayleigh Rayleigh-Bénard convection - at a critical value of the temperature gradient the molecules in the oil display an emergent collective behaviour and begin to roll.

heat is able to quickly dissipate, and the molecules continue to move in a random motion with no overall discernible pattern. However, as the temperature increases, the random motion of the molecules begins to organise into a coordinated whole, following an orderly rolling motion (depicted in Figure 1.5). The reason for this is due to the rising of the less dense oil at the bottom, met with the falling of the cooler oil at the top.

In this example, the temperature gradient is known as a *control parameter*, because it controls or affects the state of the oil molecules, and the amplitude of the convection rolls that emerge are known as the *collective variable* (or *order parameter*). Of interest for proponents of the *replacement* thesis is that the collective variable is *emergent* and *self-organising.* As Kelso (ibid., p. 7) states:

> "[T]he control parameter does not prescribe or contain the code for the emerging pattern. It simply leads the system through the variety of possible patterns or states."

Once the pattern has emerged, the behaviour of the individual molecules is in turn governed by the convection rolls; the emergent collective behaviour of the system influences the behaviour of the lower-order constituents. This *circular causality*, as Kelso describes it, is one of the main conceptual differences between dynamical systems and the serial computational processing in cognitivism.

This leads to the idea of coupling. Components of a system are *coupled* when the mathematical description of the behaviour of one component includes as a term the behaviour of the other (as is the case in the equations which describe the above system, see (ibid.)). In coupled systems it is not possible to isolate the behaviour of

of dynamic pattern formation in the brain. He explores dynamic modelling and coordination of neural states in (Kelso, 2012).

one of the components from the others. Therefore, understanding the behaviour of the system and its components requires a broader perspective.

Some argue that when a complete mathematical description of a system's behaviour can be determined, thus providing a predictive, counterfactual supporting model, the task of explaining the system's behaviour has been provided (cf. Chemero, 2011). In addition, supporters of the replacement theme go further in arguing that the parameters and variables that make up the mathematical descriptions of dynamical systems are not representational in the manner defended by the cognitivist. This latter point is a source of contention, and continues to be debated (see (Bechtel, 2009) and (Stepp, Chemero, and Turvey, 2011) as an example of this debate).

Chemero (2011) has provided a comprehensive reason as to why this debate has been sustained for so long, and argues that we should restrict the debate over representationalism and anti-representationalism to an *epistemic* question. We should ask whether the best explanation of cognitive systems involve representations, rather than the *metaphysical* question regarding the nature of cognitive systems (i.e. do they contain representations). Anderson and Rosenberg (2008, p. 56) also draw attention to the explanatory role of representations—highlighting the distinctively epistemic nature of question. They claim that the debate should ask "not what a representation is, but what it does for the representing agent". Both authors acknowledge that whether cognition can be explained without positing representations is an empirical matter that has not been settled.

Shapiro (2011, p. 156) argues that the types of behaviour that are amenable to dynamical modelling represent "too thin a slice of the full cognitive spectrum" to argue in favour of replacement wholesale. Instead it may be preferable to seek a rapprochement between dynamical modelling for minimally-cognitive behaviours (e.g. perceptually-guided action), and those which Clark and Toribio (1994) call "representation-hungry" behaviours (e.g. long-term planning or reasoning involving a distal (decoupled) object, which requires representation). Following Chemero (2011), any further discussion of representations in this thesis, will be restricted to the question of whether they are suitable explanatory posits.

Constitution

Finally, we come to the theme of constitution, which can be described as the view that cognition is comprised of objects, events and their properties that are not necessarily found solely within the brain. This means that cognitivism, which claims that cognition is simply the processing of symbolic representations in the brain, is incomplete. However, constitution does not have to be interpreted as entirely anti-computational, nor a wholesale rejection of cognitivism.

As an example of this work, Shapiro (2011) cites Andy Clark (e.g. 1997a, 2008). Of interest are the following two themes[16]:

Nontrivial Causal Spread: behaviours that the cognitivist claims result from the product of inner symbol-crunching (or from an otherwise well-demarcated system in more general cases) are in fact best explained by appealing to external mechanisms spread across the body and the world (Clark, 2008; Wheeler and Clark, 1999). The motion of so-called "passive walkers" is an example of nontrivial causal spread, as their ability to perform the function of walking depends (nontrivially) on "far-flung" environmental factors, i.e. gravity, friction, incline of a slope (Clark, 2008).[17]

Principle of Ecological Assembly: problem-solving depends on the environmen-

[16] Shapiro (2011) explores further themes, which due to space limitations have not been considered directly here, but will appear in later discussion

[17] For those unfamiliar with them, many videos displaying the behaviour of passive walkers are available to watch on YouTube.

tal resources an organism has available to it, where the environment can be considered to include the body. How much the organism contributes to the task, and how much is exploited from the environment, will be determined largely by what is most efficient. As an example, Clark (ibid.) gives the case of a tile-assembly task studied by Ballard, Hayhoe, and Pook (1997), which we discuss in detail in chapter 3. He claims that the principle of ecological assembly can be described as the view that "the canny cognizer tends to recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort."

Note that *Nontrivial Causal Spread* is not making the uncontroversial, and somewhat trivial claim that worldly objects have causal effects on the body and the mind. Rather, as with the earlier distinction made by Chemero (2011), this is a claim about how best to understand and explain the causes of a cognitive system's behaviour (i.e. what theoretical posits are required for a complete account of behaviour, and are they found entirely within the symbol-processing brain). Developing on this thread, the *principle of ecological assembly* adopts an evolutionary perspective, and argues that adaptive behaviours are likely to result from the exploitation of any resources that contribute to efficient and effective problem-solving, irrespective of where they may fall on some brain-environment boundary (e.g. orienting ingredients in a particular order to simplify the task of following a recipe). This goes beyond the trivial causal claim mentioned above, and leads to the following theme, which bears resemblance to aspects of both the conceptualisation and replacement theses:

Open Channel Perception: rather than positing inner symbols, which mediate between the world and action by constructing rich representations, open channel perception can often exploit the invariants in the optic array, which correlate reliably with certain features of the world. This idea is often discussed in ecological psychology (Gibson, 1979), but is also a key aspect of embodied robotics (Brooks, 1991; Steels, 2003). One of Clark's favourite examples to highlight this point is the idea of optical acceleration cancellation in the trajectory of a fly ball in baseball (Clark, 1997a, 2008, 2015). Rather than requiring complex computation of a trajectory, in order to guide motor behaviour towards some spot in the outfield, a baseball player can simply keep perception of the ball fixed in their line of sight by running at the appropriate speed and in the appropriate direction. By doing this, the outfielder will naturally arrive in the right spot so as to catch the ball, simply by exploiting the close coupling between their perceptual and motor systems, and without any complex computational resources. We will explore this example more thoroughly in chapter 3.

It is important to note that Clark is by no means an anti-representationalist, and acknowledges the importance of representations in accounting for some aspects of cognition (Clark and Toribio, 1994). Furthermore, he has argued in defence of a conciliatory role for the computational explanations that posit representations, and the more dynamic accounts outlined above (Clark, 1997a). The following two strictures he outlines demonstrate this more inclusive attitude:

> 1. Beware of putting too much into the head. Adaptive behavior emerges from a complex balancing act that incorporates neural, bodily, and environmental influences.
>
> 2. Beware of narrow visions of the form and content of putative internal representational systems. Such systems may involve indexical-functional (action-oriented) contents and may not require expression in the form of compositional codes and classical programmes. (Clark, 1997b, p. 475)

Whereas theorists such as Chemero (2011) and Kelso (1995) may wish to abolish the use of computationalism and representationalism from the cognitive sciences, Clark believes them to be ineliminable.

We can interpret this complementarity of computation and dynamicism, which is inherent in Clark's theory, as motivated by an idea he has defended in detail—cognition and the mind extend beyond the brain into the environment (Clark, 2008; Clark and Chalmers, 1998). Whether one chooses to defend this claim with its original functionalist commitments (e.g. Clark and Chalmers, 1998), or to reinterpret the constitutive claim evident in the principles above as a non-functionalist account (e.g. Menary, 2007) is unsurprisingly a contested matter. Regardless of how one chooses to defend the account (or which version is the target of criticism), it should be clear why the notion of embodiment is of central explanatory importance, and in what regards it differs from cognitivism. The body and the environment are the brain's partners in constituting cognition, and different types of explanation may be better suited to accounting for different aspects of them. Nevertheless, any explanation that pertains to the sort of varied, adaptive and intelligent cognitive behaviour, which we often attribute to agents such as ourselves, ignores the environment at its peril—symbol processing simply isn't enough.

1.3.3 Proper Embodiment

Although the idea is not addressed by Shapiro (2011), it is also worth considering what Stapleton (2016) calls 'proper embodiment'. This idea aims to more carefully consider the importance of fine-grained, particular details of the organism's body, with emphasis on findings from affective science. The use of the qualifying adjective (proper) draws attention to Stapleton's claim (echoed by (Colombetti, 2013)) that although developments in embodied cognition have acknowledged the importance of modelling organism-environment interactions, it has been slower to acknowledge the importance of affective neuroscience. This latter focus, she argues, is important for uncovering specific details about how our physiology contributes to cognition and consciousness. The thesis of proper embodiment thus states that:

> "[...] (at least some of) the details of our physiology matter to cognition and consciousness in a fundamental way such that (at least some of) the mechanisms of cognition are so fine-grained that specifying the algorithm for cognition would entail specifying parts of the internal body normally considered to be background or enabling conditions for cognition." (Stapleton, 2016, p. 21)

By retaining the functionalist commitments of cognitivism, Stapleton claims some versions of embodied cognition end up retaining some of the neurocentric shortcomings that it was supposed to overcome when turning away from cognitivism. To highlight this, Stapleton breaks the thesis of proper embodiment down into two independent theses: *internal embodiment* and *particular embodiment*. They are defined as follows:

Internal Embodiment: "the internal "gooey" body matters to cognition and consciousness in a fundamental way."

Particular Embodiment: "the particular details of our implementation matters to cognition." (ibid.)

Neither of the definitions make much sense in isolation, so let's expand on them. With regards to *internal embodiment*, the emphasis on 'internal' is to draw attention to the importance of interoception. This term was originally introduce by Charles Sherrington (1947) to refer to the sensation of the visceral body. However, Craig (2002, p. 655) has more recently extended its usage to include other sense such as pain, temperature and light touch, on the basis of shared neural pathways and processing areas. He argues, "interoception should be redefined as the sense of the physiological condition of the entire body not just the viscera." In short, interoception provides the brain with a general sense of how the body is coping.

Stapleton argues that the importance of interoception for cognition lies in uncovering the role of affectively significant sensory signals that originate from the internal environment of the body, and in turn motivate behaviours and provide perceptual states with value. By overlooking this crucial aspect, researchers may fail to appreciate an important aspect of how evolved cognitive systems interact and adapt to their environment.

However, while it is not true that embodied cognition research has completely overlooked interoception, according to Stapleton it has only considered the role that interoception plays in shaping cognition from within a functionalist framework. As such, embodied cognition is committed to certain tenets of functionalism, most notably multiple realisability and supervenience. Although Stapleton takes no issue with these tenets in general, she claims that adherence to these tenets (and more specifically, multiple realisability) has led researchers to overlook the importance of particular implementational details for understanding cognitive systems. This brings us to the second of her theses.

A commitment to functionalism means abstracting away from the messy physical realisers, so that what is important is identifying the causal role that some cognitive process plays within a larger system. Functionalism, with its commitment to multiple realisability, allows the researchers to effectively ignore the messy implementational details, focusing instead on what algorithmic processes are likely to be shared by different cognitive systems performing the same computational task.

In contrast, Stapleton presents an interesting example from research in evolutionary robotics that demonstrates why this functionalist strategy often overlooks important details. She discusses research on GasNets, a class of neural networks that aim to model non-synaptic gasotransmitters such as nitrous oxide, which have long been identified as an important mechanism in neural signalling. In a study performed by Smith et al. (2002), two classes of neural networks were simulated, and allowed to evolve according to equivalent measures of fitness, based on success in the task being studied (see (ibid.) for details). One of these classes of neural networks was designed to only simulate standard synaptic signalling (NoGas), whereas the other was designed to simulate gasotransmission (GasNet). Both networks achieved the same level of functional success, but the GasNet class adapted much faster than the NoGas class. Smith et al. (ibid.) claimed that this flexible adaptivity was a direct result of the gas diffusion mechanisms. Stapleton (2016) argues that this example supports the idea that a *particular physical feature* of an organism's embodiment plays a key role in evolvability, and should therefore be considered as relevant to embodied cognition more generally. A narrow focus on the functional equivalence of the two classes, she argues, is the wrong level to focus on if we wish to understand what is key to each networks ability to succeed in the task environment.

This short discussion of proper embodiment is presented here to provide a more complete (though admittedly patchy) overview of the embodied cognition literature. As we are interested specifically in decision-making, we will postpone critical remarks until later chapters. However, as a preview, a couple of questions can be raised. First of all, does the acceptance of any one theme mean a complete rejection of an alternative, or can a conciliatory approach be achieved? Conditional on the answer to the first, what would count as a satisfactory embodied approach, and can we provide a list of criteria or constraints that would allow us to identify a truly embodied theory? Over the course of the thesis we will favour a more conciliatory approach, motivated by a commitment to explanatory pluralism in the cognitive sciences, but also an acknowledgement that attempting to provide a satisfactory response to the second question seems to be unlikely given the current empirical research. As we will see in the next section, this is no reason to abandon embodied cognition.

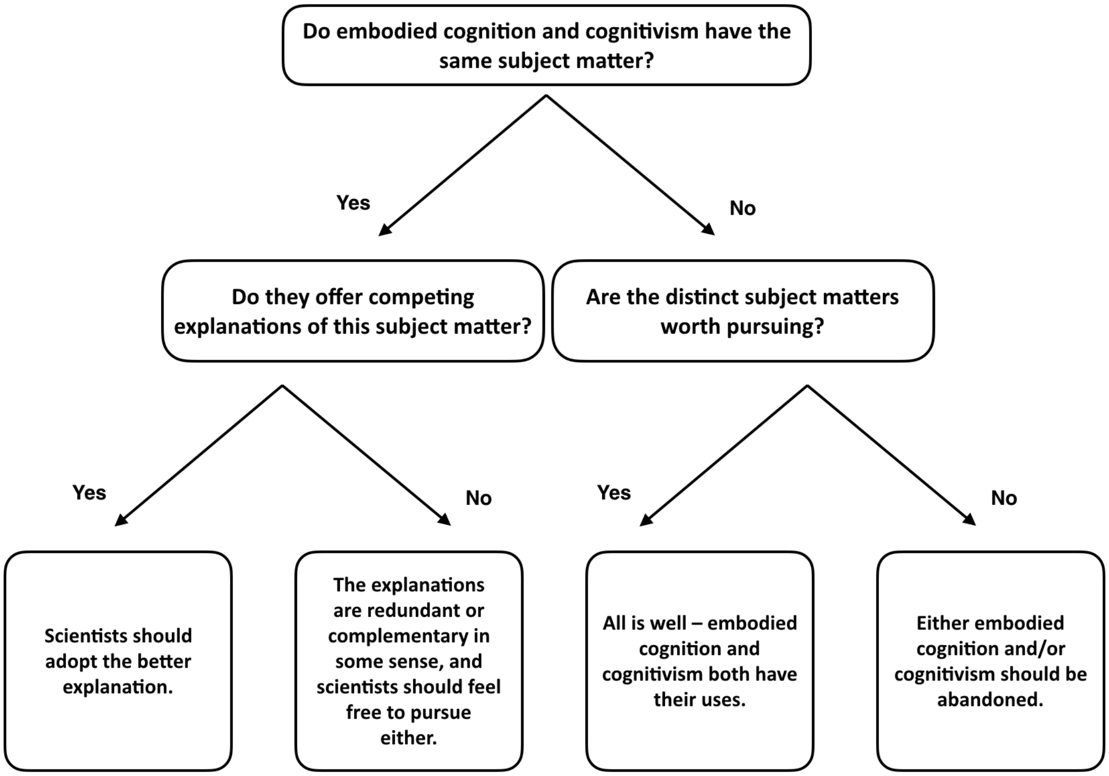

Do embodied cognition and cognitivism have the same subject matter?
Yes
No
Do they offer competing explanations of this subject matter?
Are the distinct subject matters worth pursuing?
Yes
No
Yes
No
Scientists should adopt the better explanation.
The explanations are redundant or complementary in some sense, and scientists should feel free to pursue either.
All is well – embodied cognition and cognitivism both have their uses.
Either embodied cognition and/or cognitivism should be abandoned.

Figure 1.6: Adapted from (Shapiro, 2011, p. 201).

1.3.4 Whatever it is, it's not cognitivism!

As well as providing a useful way of understanding the various commitments of embodied cognition, there is a secondary purpose to Shapiro's (2011) themes—what he terms a 'meta-theme'. The meta-theme is whether cognitivism and embodied cognition offer competing explanations of the same phenomena? He poses the following questions, which we can represent as a decision tree (Figure 1.6). In discussing this decision tree, he quickly argues that the right-hand side of the decision tree can be ignored, as embodied cognition and cognitivism do in fact have the same subject matter (e.g. perception, decision-making, motor control etc.), irrespective of which of the three themes is adopted. This leaves the left-hand side of the tree.

As we saw in the previous sections, each of the three themes is opposed to cognitivism in sometimes shared, and sometimes different manners. If one is committed to a particular theme, then this may require some sort of rapprochement between the aspects of cognitivism and embodiment that are compatible (as with constitution), or attempting to find further theoretical and empirical support to widen the explanatory support (as with replacement).[18] However, there are many commonalities between the themes, which means categorising any particular piece of research can be challenging. For example, *conceptualisation* and *replacement* share an emphasis on the dynamic interaction between body and environment; *replacement* and *constitution* acknowledge a role for dynamical modelling (albeit to a different degree) in explaining an agent's situated behaviour, and *constitution* and *conceptualisation* point to the importance of body-environment interaction as a potentially illuminating source of our conceptual knowledge. As these commonalities between the themes become more intertwined, it can begin to look like taking one of them as the primary defining characteristic of a position is a somewhat arbitrary decision.

This may lead one to think that adopting an embodied framework places any theory on unstable foundations, unless one can explicitly outline all of the assumptions that are being made, and provide independent justification for each of them in turn. Otherwise, the critic could always argue that the disarray caused by the myriad positions threatens the cogency of the position being defended. However, though we will endeavour to make clear over the course of this thesis exactly what commitments to embodiment are being made, explicitly outlining all of our commitments isn't necessary at this stage for a couple of reasons.

Unlike cognitivism, embodied cognition should still be seen as in its infancy, and

[18] See Chapter 7 of (ibid.) for further details relating to each theme.

without an orthodox set of constraints—perhaps this is why (Calvo and Gomila, 2008, p. 3) define embodied cognition as a “post-cognitivist approach”? Despite the point raised at the start of this chapter regarding the insufficiency of a foil to act as a delineating constraint, we are forced to accept it for the time being—whatever embodied cognition may be, it certainly isn’t cognitivism!

Shapiro (2007) has been careful to point out the varied roots of embodied cognition, while at the same time arguing that its somewhat nebulous nature is no reason for one dismissing it. He stresses that for the time being it is best to refer to embodied cognition as a *research programme*, rather than as a *theory*, to avoid the appearance of strict unity of methodological practices. Is this enough to satisfy the cognitivist? Surely not. However, they should pause before celebrating their unitary conceptual framework. For as Menary (2010, emphasis added, p.460) states of the 4E programme:

> “[...] we are in a position of *abundance*, not *disarray*: if one looks at the array of empirical cases that are provided by the, now rich, 4E literature, one finds the need for a battery of different explanatory methods that are suited to the differences in those cases.”

Over the course of this thesis we will discuss a wide variety of these empirical cases, and it will be argued that explanatory pluralism is at present the best methodological stance to adopt in the cognitive sciences. Although we will not take a firm stance on any one of the three themes outlined above, we will acknowledge when significant disagreements arise between them. It may turn out that these disagreements are nothing more than the product of blind scholars grabbing at different parts of an elephant—as illustrated in the famous parable. Alternatively, these disagreements may turn out to be more problematic and thus require resolution in the future. For the time being, and indeed for this thesis, it will suffice to show why the embodied cognition research programme is preferable to cognitivism. Therefore, the focus will be less on whether the varieties of embodied research are competing with one another, and more on whether they collectively provide a genuine alternative to cognitivism. We take the latter to be a more relevant discussion point.

Chapter 2

Predictive Processing: An Introduction

The success of the information-processing approach to cognition should not be understated. As we saw in Chapter 1, the ability to formalise key notions provided a common vocabulary and valuable conceptual tools for the emerging discipline of cognitive science. However, despite the mathematical rigour that this brought to cognitivism, the assumption that information-processing is a bottom-up, serial process (e.g. the classical sandwich model) has recently been challenged by a contemporary framework known as predictive processing (PP).

Whereas cognitivism treats perception as a largely bottom-up process of incremental feature detection, PP overturns this conception, instead placing an emphasis on top-down predictions about expected sensory data (section 2.1). These predictions emerge from hierarchical generative models, which are encoded by the brain in a probabilistic manner (section 2.3), and are continuously modified by bottom-up error signals that communicate mismatches between predictions and actual activity (section 2.2). This initial process is also accompanied by expectations of the precision of incoming sensory data (section 2.4).[1] Each of these claims requires unpacking, but there are two ways we could proceed. On the one hand, the PP framework can be described in a manner that leads to understanding the role of the brain from a neurocentric, internalist perspective (Hohwy, 2013) (section 2.5). On the other hand, the framework can be described in a manner that uncovers a deep affinity with the embodied perspective (Clark, 2016b). Unsurprisingly, we favour the latter, but for expository purposes it is best to consider the former as our starting point.

This chapter proceeds as follows: first, we will outline the contemporary framework known as predictive processing (PP) following the work of Jakob Hohwy. He has claimed that, within the PP framework, many diverse phenomena such as perception, action and attention can be modelled as a form of statistical inference, which in turn may provide a unifying account of the brain's diverse activity. The unifying mechanism, according to Hohwy (2014, p. 2) is known as prediction-error minimisation (PEM), and is claimed to be the "only principle for the activity of the brain". In order to evaluate this claim, we will discuss the main components of the framework (as outlined above), as well as theoretical and empirical research that supports them. We will then look at some of the conceptual challenges with this interpretation. This will set the groundwork for later chapter, where we will argue in favour of an embodied interpretation of PP, and in turn explore an account of decision-making.

2.1 Overturning Tradition

Network-based approaches in cognitive neuroscience view connections in the brain as massively recurrent, and dynamically interacting with other local networks (Sporns,

[1] Although a brief overview of PP is provided in this chapter, for more in-depth overviews and introductions, see (Clark, 2016b; Hohwy, 2013), For formal details, see (Friston, 2010; Seth, 2013)

2011). As such, information does not just feed forward in a serial, incremental manner starting with perception and ending with motor control. Rather, feedback connections exist at multiple levels of the neural architecture, integrating, influencing and inhibiting ongoing activity. The extent of these feedback connections should not be downplayed. As Sporns (ibid., emphasis added, p.150) notes:

> “Even in regions of the brain such as primary visual cortex that are classified as “sensory,” most synapses received by pyramidal neurons arrive from *other cortical neurons* and only a small percentage (5 percent to 20 percent) can be attributed to sensory input. Cortical areas that are farther removed from direct sensory input are coupled to one another via numerous mono- and polysynaptic reciprocal pathways. This prevalence of recurrent anatomical connections suggests that models which focus exclusively on feedforward processing in a silent brain are likely to capture only one aspect of the anatomical and physiological reality. [...] [R]ecurrent or reentrant processes make an important contribution to the shaping of brain responses and to the creation of coordinated global states. This coordination is essential for the efficient integration of multiple sources of information and the generation of coherent behavioural responses.”

How should we model the function of these recurrent connections? A recent proposal known as predictive processing (PP) treats the recurrent connections as encoding top-down predictions about the incoming sensory data, and bottom-up activity as signalling what the predictions got wrong (i.e. an error signal). Figure 2.1 represents this principle by way of a simplified schematic. Each layer in this network encodes what is known as a *probabilistic generative model*, which tries to predict the activity at the layer below it. Furthermore, the system considered as a

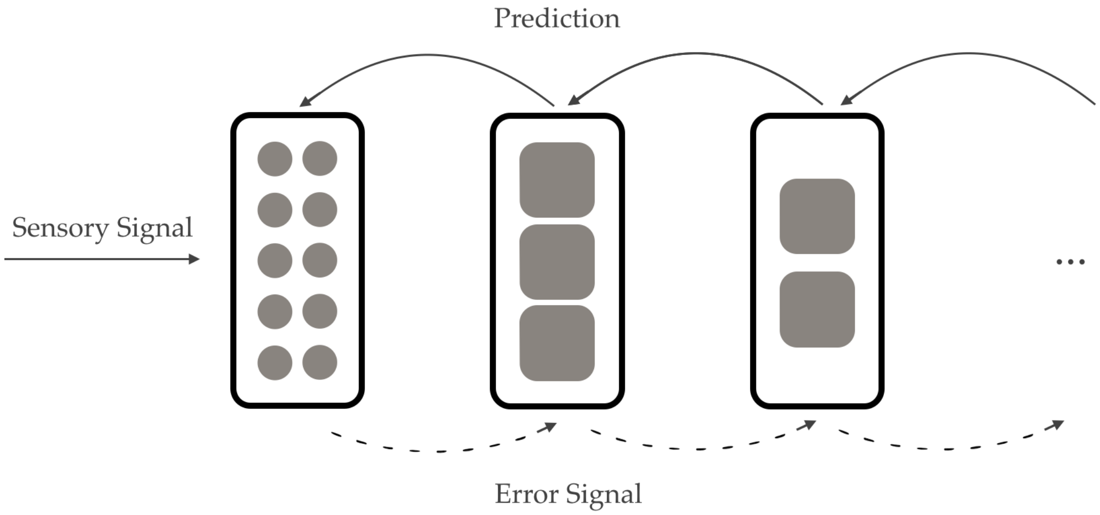


Figure 2.1: A simplified schematic of the principle of predictive processing. Each higher layer generates predictions about the neural activity at a lower layer. Only the residual error (unpredicted activity) is signalled to the higher layers.

whole encodes a *multi-level*, probabilistic generative model that tries to predict the sensory information from the environment.

The notion of a generative model has its roots in machine learning with the famous *Helmholtz machine* and *wake-sleep algorithm* (Dayan et al., 1995; Hinton et al., 1995). It is often contrasted with the notion of a *discriminative* model, which is constructed by a neural network on the basis of successive training on some data set. In this latter instance, the neural network aims to correctly classify (or discriminate) the incoming data, and the parameters of its models can be iteratively adjusted in order to increase the accuracy of its classifications. This method is appropriate for simple data sets (e.g. sets that are accurately modelled using univariate linear regression), but can perform poorly (if at all) when uncovering the hidden causes of data generated by a large number of non-linearly interacting hidden causes. For example, consider the case of a data set representing house sales, and the task of fitting a model that successfully predicts future house prices on the basis of the data received. A neural network operating with a discriminative model may be able to model the relationship between a number of interacting variables (e.g. plot-size, year of build, quality of schools in 1-mile radius, average price of neighbouring buildings and so on) inherent in the data set, but will likely struggle in cases where the variables interact in unconventional ways. For example, which weighted combination of variables accurately captures 'market demand'?

An alternative strategy is to use a generative model. As the name implies, this strategy allows the network to generate its own data, structured around previously learned expectations, and compare the accuracy of this simulated data against the actual data it receives. These simulations are based on predictions about what the network expects to be the most likely cause of the data it receives, and these expectations are updated by signalling what is known as a prediction error that also acts as a learning signal for the network. This means the latter is not restricted solely to detecting pre-classified patterns in the data it receives, in order to refine the parameters of its models. As (Clark, 2016b, p. 20) highlights:

> "The Helmholtz Machine was an early example of a multilayer architecture trainable without reliance upon experimenter pre-classified examples. Instead, the system 'self-organised' by attempting to generate the training data for itself, using its own downwards (and lateral) connections."

A system that can effectively use a generative model in this manner is one step closer to effectively representing the hidden causes of the sensory data (in section 2.2 we will explore the second necessary component). However, the world is an uncertain and multifaceted place, and the same input is often consistent with a number of causes (e.g. a number of houses could be the same price but for different underlying reasons), some of which may themselves be an emergent product of interacting causes (e.g. market demand). Therefore, to maintain predictive accuracy, a generative model should be *hierarchical* and *probabilistic*, such that the *most likely* cause (or as we will see shortly, the one with the highest *posterior probability*) should be selected by the system as the true cause of its input. The motivation for adopting a multilevel or hierarchical setting, reflects both the efficiency of predictive architectures (e.g. the predictive coding example discussed shortly), but also a recognition that the world is equally composed of highly-structured causes that need to be understood. As Clark (ibid., pp. 24-25) notes:

> "This is important since structured domains are ubiquitous in both the natural and human-built world. Language exhibits densely nested compositional structure in which words form clauses that form whole sentences that are themselves understood by locating them in the context of even larger linguistic (and non-linguistic) settings. Every visual scene, such as a city street, a factory floor, or a tranquil lake, embeds multiple nested structures (e.g., shops, shop doorways, shoppers in the doorways; trees, branches, birds on the branches, leaves, patterns on the leaves). Musical pieces exhibit structures in which overarching sequences are built from recurring and recombinant sub-sequences, each of which has structure of its own. The world, we might reasonably suggest, is known by us humans (and doubtless most other animals too) as a meaningful arena populated by articulated and nested structures of elements."

In order for agents such as ourselves to understand the dynamic complexities of the world, this property seems indispensable, and applied to neurobiological phenomena, this idea finds empirical support from recent work by Bastos et al. (2012), Kanai et al. (2015), and Mumford (2003) as well as work in predictive coding.

2.1.1 Evidence from Visual Cortex

An early attempt to model neural systems using a hierarchical predictive architecture was put forward by Rao and Ballard (1999) for the case of visual cortex. This model had the additional virtues of a) being able to independently accommodate the existence of extra-classical *receptive-field* effects, which had been detected in several visual cortical areas, and b) demonstrating an efficient method of information-processing that could be implemented by the brain.

The receptive-field (RF) of a neuron is the region of sensory space to which the neuron is optimally tuned, such that when a relevant stimulus is present in that region of space it will trigger the firing of that neuron. It is possible to construct a probabilistic representation of the neuron's receptive field, known as its 'tuning curve', which takes the form of a probability density function over the relevant stimulus parameters. The *extra-classical receptive-field effect* that was investigated by Rao & Ballard is known as the "endstopping" effect. It refers to the initial presence of a tuned response to an optimally oriented line segment, which is reduced (or eliminated) when the same stimulus happens to extend beyond the neuron's classical receptive field (e.g. a line segment that extends beyond the peripheries of the visual field).

Extra-classical effects have some interesting properties. For example, rather than increasing activity, many in fact inhibit or suppress activity. Though some proposals had been put forward prior to their paper, Rao & Ballard claimed that they failed to generalise outside of the visual cortex. However through the development of an alternative account based on the principle of predictive coding (PC), Rao and Ballard came to the interesting finding—by developing hierarchical PC models of visual cortex—that these extra-classical receptive field effects may be a direct consequence of the brain's use of hierarchical predictive coding.

PC claims that neural networks need only signal deviations from the expected statistical regularities in the sensory input to higher levels for processing, subject to an internal generative model being able to generate predictions that flow downwards through the network. Importantly, they add "[t]his reduces redundancy by removing the predictable, and hence redundant, components of the input signal." (1999, p. 79)

With regards to efficiency, note that only the residual error, or unpredicted activity is signalled to the higher layer, which in turn reduces the redundant signalling of information, increases efficiency, and provides a hierarchical structure to the generative model encoded by the network.[2]

As an illustration, consider the case of compressing a RAW photographic image file (depicting the French flag) into a format like JPEG. Large portions of this image will contain pixels whose value will be strongly correlated with the value of its closest neighbours (i.e. large sections of blue, white and red). Where significant deviations exists, they will be representative of features such as edges (e.g. the edge between the blue and white segments). Therefore, encoding only the unexpected variation (e.g. the cases where the actual value was not predicted by the generative model) allows the network to only transmit the difference between the prediction and the actual data (the *prediction error*), which is a more efficient method than attempting to transmit large swathes of data to each layer of the system.

Further evidence of efficient coding was explored by both Mumford (2003) and Hosoya, Baccus, and Meister (2005). The latter explored the contextual effects on retinal ganglion cells, under the assumption that they implement an efficient coding

[2] The mathematical properties of hierarchical generative models means that they can be composed of many additional nested generative models. The whole hierarchy could thus be considered as one single generative model, and the whole hierarchy minus the bottom levels could also form another generative model, and so on (Friston, 2008).

scheme such as predictive coding, but also display contextual effects based on higher-level priors that rapidly modulate the expectations of the lower-levels. Clark (2013c, p. 184) highlights two important findings from their work:

> "Putting salamanders and rabbits into varying environments, and recording from their retinal ganglion cells, Hosoya et al. confirmed their hypothesis: Within a space of several seconds, about 50% of the ganglion cells altered their behaviors to keep step with the changing image statistics of the varying environments [...] there are neuronally plausible ways to implement such a mechanism using amacrine cell synapses to mediate plastic inhibitory connections that in turn alter the receptive fields of retinal ganglion cells so as to suppress the most correlated components of the stimulus. In sum, retinal ganglion cells seem to be engaging in a computationally and neurobiologically explicable process of dynamic predictive recoding of raw image inputs, whose effect is to "strip from the visual stream predictable and therefore less newsworthy signals."

2.2 Predictions and Prediction Error in the Brain

Intuitively it seems obvious that we are able to generate predictions about future events, which range from the trivial (e.g. what I expect to find in my fridge), to the potentially transformative life experiences (e.g. what life will be like if I have a child). Recent theories have begun to entertain the idea that the primary function of the cortex may be the prediction of such future states in our environment (Bar, 2011a; Clark, 2016b; Hohwy, 2013; Kveraga, Ghuman, and Bar, 2007). Ouden, Kok, and De Lange (2012) provide a comprehensive list of the various applications that these family of ideas have been applied to, including visual and auditory processing, somatosensory perception, motor control, language, memory, cognitive control, and motivational value processing.

Since the environment is constantly in flux, there is an ineliminable source of uncertainty that an agent must deal with if they're to successfully interact with the world. Therefore, if an agent is to maintain accurate inner models, she must have a way of determining when they go wrong. As well as sharing a commitment to the importance of prediction, by necessity these theories also share a commitment to a second notion—the possibility of prediction error.

We can define prediction error as the mismatch between an agent's prior expectation and the actual state of affairs in the environment. As a simplified illustration, consider the following: you are trying to bake a cake. You have a slice of the cake you wish to reconstruct, and have been given all but two of the ingredients (in the right quantities) to make the cake. You are competent enough of a baker to determine that only two ingredients (baking powder and vanilla essence) are missing, but do not know what their quantities are. The complete set of ingredients (with the right quantities) can be considered the *hidden cause* of the cake that you are trying to reconstruct (we denote this as θ). Given that this is the most delicious cake you have ever tried, you decide to bake several cakes, varying the quantities of baking powder and vanilla essence across the different trials. You end up with six different cakes which you label $(P_1, P_2, ..., P_6)$ respectively—these represent your *predictions*. We can also think of the set of hypothesised ingredients, independently from their particular quantities, as akin to the generative model (and its parameters) from which we derive the individual predictions. You proceed to compare the cakes against the original slice (the sensory input). The first cake (P_1) has the same height and texture as the original, but is a lot more bland. You conclude that recipe P_1 has too little vanilla essence—there is a corresponding error generated by the difference in height of the two cakes. Next you compare cake P_2, but this time find that although the flavour is correct, it has not risen as much as the original. You may be able to conclude at this point that you have sufficient information to determine the complete set of ingredients: the amount of baking powder from P_1 and the amount of vanilla essence from P_2. However, it is also possible that there is a further undiscovered relation between the two ingredients when their quantities are changed. Therefore, you continue to test the additional cakes and eventually find that in order to mask the bitter taste caused by increasing the baking powder to get the right height, you now need slightly more vanilla essence than was initially used in recipe P_2—this is corrected by the quantities used in P_6. By forming predictions in this manner, and comparing them with the original cake, you eventually manage to minimise the prediction error between your prediction of θ, and its corresponding effect (the original cake).[3] This example is reminiscent of hypothesis-testing in science, whereby a scientist forms individual hypotheses and tests them against some set of data and attempts to find the closest fit. The similarity is no accident, as we shall see shortly (section 2.3).

The continual testing of different predictions should also reflect the brain's ongoing attempt to learn from its prior experience. In PP, the prior expectations (or predictions) are generated by an agent's model of the environment, which as we will

[3] There are some important differences between this example and PP. For one, the set of ingredients is considered finite, whereas there is a possibly infinite number of hidden causes that could generate the input our brains receive. Secondly, in PP there is a difference in kind between the inner generative model and the hidden cause (θ) in the environment, which is overlooked in this example in favour of a more personal-level description. The inappropriate use of personal-level terminology is highlighted in case one mistakenly worries that we are further attributing inappropriate terminology to the brain (see section 2.3). Finally, there is also the overlooked case of the brain's ability to control action. Therefore, we could extend θ to include the actions taken by the original baker (e.g. mixing and baking) within our predictions. As we will see later, this is an important component that needs adding, but unfortunately we have not yet introduced enough material to elaborate further at this time.

see is partly constrained by the structure of its neural circuits, and partly shaped by the statistical regularities inherent in the flow of sensory inputs that the agent experiences over the course of its lifetime. The latter is derived from the fact that a prediction error, by signalling an inaccuracy in part of the agent's inner models, calls for an update of the model's parameters to take place. This is why authors such as Hohwy (2014, p. 2) claim that:

> "[...] the brain is an organ that on average and over time continually minimizes the error between the sensory input it predicts on the basis of its model of the world and the actual sensory input."

The inclusion of "on average and over time" is important, as it points to a need to consider a sufficient amount of flexibility in the models to avoid the problem often referred to in statistics as *overfiitting*. To be predictively successful a model should not be either too complex (i.e. containing too many parameters such that it reacts to minor fluctuations or noise by generating significant error), nor too general (i.e. unable to spot an underlying trend in the incoming data) (cf. Hohwy, 2013, chapter 2, for a simple illustration). Although each prediction is trying to account for the evidence in as accurate a manner as possible, the global performance of the system takes priority, to avoid running into tricky problems such as the *dark-room problem* (cf. Friston, Thornton, and Clark, 2012). In short, this is the issue of how to explain why an agent trying to minimise prediction-error does not simply go into a dark-room and predict to experience nothing at all. As we will see, in chapter 4, resolving this requires understanding the predictions being generated in an organism-relative manner, one which acknowledges the agent's phenotype.

2.2.1 Evidence of Prediction Errors

A recent review paper by Ouden, Kok, and De Lange (2012) considers what empirical evidence there is for the idea that the brain generates prediction errors. Not only do these authors provide an extensive review of recent prediction error research in the cognitive sciences, but they also point to two important experimental paradigms that are commonly used throughout the literature: the *oddball* and *omission* paradigms.

The former refers to the presentation of a deviant (or oddball) stimulus in a sequence of repeated standard stimuli, and the expectation that the presentation of an oddball will elicit larger neural activity over the relevant sensory areas. The latter refers to the instance where a subject is primed to expect a subsequently withheld (or omitted) event, and a corresponding neural response is measured. Both can be quantified as a measure of *surprisal*, which is an information-theoretic measure that refers to how improbable (or surprising) some outcome is, conditioned upon a model.[4]

In the case of the oddball paradigm, researchers aim to measure what is known as the mismatch negativity (MMN) component of an event-related potential (ERP): the electrophysiological response that results from the presentation of an odd sensory, cognitive or motor event in a sequence of events. To try to dissociate MMN from repetition suppression (i.e. the decrease in activity as result of the repetition of the same stimulus), Tervaniemi, Maury, and Näätänen (1994) played subjects a series of initially non-repeating, rising auditory tones, with the inclusion of a single repeated identical tone (the unexpected oddball) at an unknown stage in the sequence. Their

[4] Formally, it is the negative log of how probable the outcome of an event is ($-logP(e)$), the long-term average of which can be considered the entropy of a random variable. This measure is intuitive when one considers that as the probability of an event goes up the negative log value goes down—an event that is highly probable is unsurprising.

findings included an observed MMN at the time of the repeated tone, suggesting that there was a violation of the agent's expectations.

Perhaps more interestingly for PP are the robust findings supporting the omission paradigm. This is because the increased activity is hard to account for using standard bottom-up accounts, as there is no stimulus to evoke a response. However, PP naturally accounts for this, due to the ubiquity of prediction error transmission that it assumes. As Kok et al. (2011) state:

> "Since predictive coding theories state that the response in sensory cortex is largely determined by the violation of predictions, it may be expected that the failure of a predicted stimulus to appear would similarly evoke a response (prediction error) in the relevant sensory cortex, even though no physical stimulus is presented."

Indeed, a number of studies using the omission paradigm, reported by Ouden, Kok, and De Lange (2012, p. 4), measured evoked responses in the absence of expected physical stimuli, which in connection with the other evidence they summarise, leads the authors to claim that, "PEs appear ubiquitously throughout the brain, lending support to the notion that coding of PEs is a general neural coding strategy."

Returning to Figure 2.1, we can see that the role of bottom-up information in PP is the transmission of prediction-error to the higher-levels of the hierarchy, signalling how accurate the higher-level predictions were at accounting for the sensory evidence.[5] A claim made by proponents of PP is that this process occurs at each

[5] The exact manner in which we measure this accuracy (or inaccuracy) is still an open question that depends on the acquisition of further empirical evidence. At present, authors such as Hohwy (2013) and Friston (2010) favour a measure known as the Kullback-Leibler divergence (or relative entropy) because of certain properties it has (i.e it is always non-negative and non-symmetric).

layer of a multi-level hierarchical system. For example, each of the layers depicted in Figure 2.1 encodes a model, which generates predictions pertaining to the expected neural activity at the layer below, and are continuously updated by the ongoing flow of predictions errors. If translated into a formal model, such a schema implements a version of Bayesian inference often referred to as empirical Bayes or variational Bayes (see section 2.3).[6] As Hohwy (2014, p. 4) states:

> “Computationally, perception can then be described as empirical Bayesian inference, where priors are shaped through experience, development and evolution, and harnessed in the parameters of hierarchical statistical models of the causes of the sensory input. The best models are those with the best predictions passed down to lower levels, they have the highest posterior probability and thus come to dominate perceptual inference. Error is minimized through some minimization scheme such as gradient descent, expectation maximization, or variational Bayes.”

This close connection with Bayesian statistics provides Hohwy with the formal support for what he describes as the hypothesis-testing brain.

These properties make the KL-divergence a suitable measure for the PP framework and a more generalised notion known as the free-energy principle, but other authors such as Clark (2016b) point to the wealth of possible variant architectures for PP that are currently under investigation, and the many ways of conceiving of the notion of prediction-error (Ouden, Kok, and De Lange, 2012, cf.). For the time being, these matters will be put aside as the formal details are not necessary for our discussion.

[6] Though these schemas are considered to be Bayes-optimal, unlike true Bayesian inference, the use of sensory information to update posterior beliefs proceeds is only approximated (Friston, 2010). This is considered preferable due to the computational intractability of trying to estimate hidden variables in the sort of sensory data indicative of real-world systems.

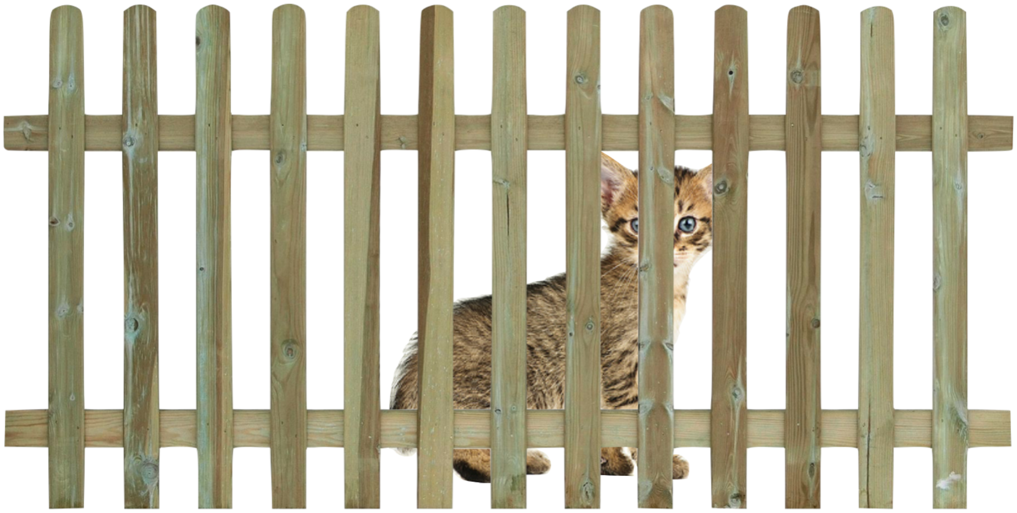


Figure 2.2: The above image could have been produced by a cat occluded by a fence, or a series of cat slices placed opportunely between the bars of a fence. The problem of perceptual inference can be seen as the task of determining which of these two possibilities is responsible for the sensory evidence received by the organism.

2.3 Hypothesis Testing

Hohwy (2013) introduces the PP framework through the analogy of hypothesis-testing. In the case of perception, this view has origins in the work of Helmholtz and Gregory, but Hohwy also notes that it dates as far back as ca. 1030 with the work of Ibn Al Haytham who stated that "many visible properties are perceived by judgment and inference" (quoted in Hohwy, 2016, p.1). Though sharing roots with these authors, the PP framework is a more modern example of what has recently been termed the Bayesian Brain hypothesis (cf. Deneve, 2008; Doya et al., 2007).

The Bayesian Brain hypothesis (BB) defends the claim that the brain implements processes that approximate the rational method of weighing new evidence against prior beliefs (i.e. conditionalisation), by using Bayesian methods to successfully model the functional activity of the brain.[7] As an analogy, consider the scene depicted in Figure 2.2. You need to determine what is behind the fence; is it a cat standing still or a carefully placed series of cat slices designed to trick you? This basic perceptual task is akin to the inferential task faced by the brain according to BB. There are hidden causes in the world that are responsible for the perceptual state currently instantiated in the brain, and the brain has to determine which of the possibilities is most likely given the sensory evidence it is receiving. Each of these possibilities can be referred to as a *hypothesis*, and thus the task is to determine which of these hypotheses is the correct one. This turns perception into an inferential problem; how is the right hypotheses shaped and selected? Unsurprisingly, advocates of BB state that the problem should be approached using Bayes' rule.

$$P(H_i \mid E) = \frac{P(E \mid H_i)\, P(H_i)}{P(E)} \tag{2.1}$$

In the current example, the evidence (E) refers to the sensory signal received by the visual system, and the hypotheses (H_i) would be either a cat or a series of cat slices. Whichever hypothesis has the highest *posterior probability* $P(H_i \mid E)$ is the one that is selected by the brain. However, this simplified account poses a number of challenges. For example, what does it mean to say that the brain is performing

[7] Clark (2013c) highlights an important comment made by Spratling (2013) in response to his overview of predictive processing, which also acknowledges BB. Spratling calls PP and the BB hypothesis examples of intermediate-level accounts. They do not specify implementational details, and instead opt for identifying the "common computational principles that operate across different structures of the nervous system and across different species", and seek "integrative explanations that are consistent between levels of description". By doing so, Spratling (ibid., p. 232) claims "they provide functional explanations of the empirical data that are arguably the most relevant to neuroscience."

Bayesian updating; who sets the prior $(P(H_i))$, and what evidence is there that perception is actually like Bayesian updating (let alone the brain in its entirety)?

With regards to the question of what it means to say that the brain is performing Bayesian updating, and whether the brain in some sense knows Bayes' rule, Hohwy responds by stating that although examples such as the one above are useful heuristic devices to convey the idea that perception is inferential, it is more appropriate to state that perception is *unconscious inference* in the sense put forward by Helmholtz (cf. Hatfield, 2002). Rougly speaking, this is the idea that the phenomenal content and nature of perception is produced by inferences or judgments, which are unnoticed or unconscious by the agent in question. This idea was also developed by psychologists such as Gregory (1980), who further argued that the notion of hypotheses should be considered in a non-propositional manner in order to exploit the tools of information theory for modelling purposes, and more closely draw an analogy with hypothesis-testing in science. Echoing the sentiments of Helmholtz and Gregory, Hohwy (2013, p. 23) argues that the application of Bayes' rule to the brain carries with it the risk of *neuroanthropomorphism*, which he defines as "inappropriately imputing human-like properties to the brain and thereby confusing personal level explanations with subpersonal level explanations." Instead, Hohwy argues on functionalist grounds that in order to understand how a brain engages in unconscious perceptual inference we must also be able to understand how neurons realise the functional rule set out by Bayes' rule. This is a well-rehearsed issue in philosophy of science and philosophy of mind, and throws up a number of questions such as, the autonomy of functional descriptions, the nature of realisation in general, and questions about whether we should adopt a realist stance towards models of this kind.[8] Given that we will be

[8] Colombo and Seriès (2012), for example, argue that currently we should have an instrumentalist attitude towards Bayesian models in neuroscience. They state that we can hope to learn that perception is Bayesian inference, or that the brain is a Bayesian machine, only to the extent that

largely favouring an alternative conception that eschews the hypothesis-testing gloss, it is not necessary to delve into this matter further.

In the previous section, it was mentioned that PP implements what is known as Empirical Bayes. By appealing to Bayesianism, advocates of PP (or indeed BB in general) are thus required to say who sets the priors. Far from being a tedious mathematical requirement, it also reflects a longstanding commitment in the cognitive sciences, which states that in order to effectively engage the world, an agent must be able to incorporate constraints based on the statistical regularities inherent in the environment. In the case of probabilistic schemas such as BB, this means tuning your priors to reflect the underlying regularities in your sensorimotor input, and in turn implicitly embody tacit knowledge of the structure of the world (Feldman, 2013). In the case of the above example, this means tuning your priors to a world full of cats; not cat-slices.

We saw an initial reason why the hierarchical structure of PP was important in section 2.1. A further reason is that in the case of empirical Bayes, priors are extracted from higher-level models (in the form of top-down predictions) that have been shaped by previous experience. This schema allows for the brain to learn and adapt to the current experiential context by estimating the priors from the data through the iterative process of PEM previously outlined, maintaining accurate models that can be subsequently used as the basis for future priors. It has been argued that many of these priors could have been formed through long-term exposure to the sort of sensory signals inherent in an organism's developmental environment,

these models will prove successful in yielding secure and informative predictions of both subjects' perceptual performance and features of the underlying neural mechanisms. However, they argue that Bayesian models in neuroscience do not provide mechanistic explanations, and are only useful devices for predicting and systematising observational statements about people's performances in a variety of perceptual tasks.

but also that some priors may have been hard-wired over an evolutionary time-scale (Hohwy, 2012). If this is the case, and it is certainly speculative at this stage, then it would be expected that different priors will be revisable to different degrees based on an organism's history. Nevertheless, Empirical Bayes is certainly well suited to modelling a hierarchically-organised system such as the brain, for as Friston notes:

> "Empirical Bayes harnesses the hierarchical structure of a generative model, treating the estimates at one level as priors on the subordinate level. This provides a natural framework within which to treat cortical hierarchies in the brain, each level providing constraints on the level below. This approach models the world as a hierarchy of systems where supraordinate causes induce and moderate changes in subordinate causes. These priors offer contextual guidance towards the most likely cause of the input." (Friston et al., 2015, p. 822).

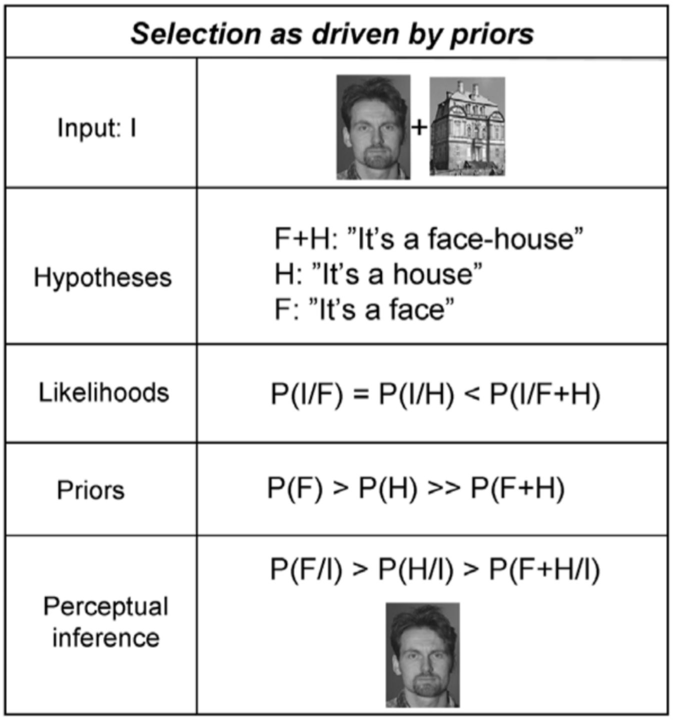
This provides a further compelling reason for adopting Empirical Bayes; by extracting priors from higher levels, predictions at lower levels will be subject to contextual modulation. For this to be effective, and to support learning, the hierarchy should thus be structured according to an increasing spatio-temporal scale, such that higher levels are tuned to the larger and slower statistical regularities in the environment.

A couple of examples will be illustrative at this point. Bar (2011b) and Kveraga, Ghuman, and Bar (2007) have shown how novel visual scenes trigger rapid ascending projections of low spatial frequency to allow the brain to get the "gist" of the scene before the arrival of the higher spatial frequency information, which in turn provides additional detail. Bar (2011b, p. 7) argues that the low spatial frequency version could be responsible for rapidly activating what he calls a "prototypical context frame" in memory, which is "sufficient in most cases to generate rapid predictions that guide our pressing goals, such as navigation and avoidance." In the case of PP this context frame would be the higher-level predictions that contextualise the information expected by lower-levels. Imagine, for example, looking for a lost golf ball in tall grass. When initially trying to find the ball, it is far better to be attentive to the low spatial frequency information (i.e. the roundness of the ball), rather than the higher spatial frequency details such as any text printed on it.

Additionally, Kiebel, Daunizeau, and Friston (2008) explored how hierarchical modelling of birdsong could be used to uncover multiple scales of temporal information inherent in the signal, which could be used by other birds to recover information about the bird that is singing. As examples, the authors state that longer time-scales may be used to measure how long a bird has been singing, providing information of the bird's fitness, whereas at shorter time-scales, the amplitude and frequency spectrum inherent in the dynamics of the birdsong could reflect the bird's strength and size. Although their birdsong models are offered as proof of principles, the authors also reviewed evidence that supports the idea of a hierarchical organisation of the cortical hierarchy, which is reflected in the increasing spatiotemporal scales of their models. They argue that regions of the brain that are farther away from primary sensory areas, encode representations of the environment that change more slowly than the rapid fluctuations at more peripheral layers.

2.3.1 Evidence from Binocular Rivalry

The idea of the hypothesis-testing brain receives wide-ranging theoretical and empirical support from the BB hypothesis (see Chater et al., 2010; Ernst and Banks, 2002, for some examples), with particular emphasis being given to the idea that neural populations can encode probability distributions (Pouget et al., 2013). However, perhaps the most striking (and certainly less technical) example comes from the phenomenon of binocular rivalry. Binocular rivalry occurs when subjects have dif-



Selection as driven by priors	
Input: I	+
Hypotheses	F+H: ”It’s a face-house” H: ”It’s a house” F: ”It’s a face”
Likelihoods	P(I/F) = P(I/H) < P(I/F+H)
Priors	P(F) > P(H) >> P(F+H)
Perceptual inference	P(F/I) > P(H/I) > P(F+H/I)

Figure 2.3: A simplified account of binocular rivalry explained in Bayesian terms. Reprinted from (Hohwy, Roepstorff, and Friston, 2008).

ferent images presented to each of their eyes, by using some sort of specially adapted headset. When this is done correctly the subjective visual experience of the subject continues to alternate between the two images, rather than settling on one of the images. Hohwy, Roepstorff, and Friston (2008) argue that this phenomenon can be understood as the brain engaging in probabilistic unconscious perceptual inference about the causes of its current sensory input. This is illustrated in Figure 2.3, where each of the candidate hypotheses the brain is said to entertain is outlined (i.e. the sensory input is a) a face, b) a house or c) a face-house). The authors argue that rivalry occurs because there is no single hypothesis that from a Bayesian perspective consistently enjoys both high likelihood $P(E \mid H_i)$ and a high prior probability $P(H_i)$. Although one of the hypotheses may temporarily explain the sensory input to one eye, at the same time it fails to capture the incoming evidence from the other, leaving a significant portion of the bottom-up signal unaccounted for. Over time, the instability in the perceptual state rises forcing a transition to the rival hypothesis. The authors add that the reason a conjunct of a face and a house is not perceived (despite the occasional gradual transition between the two images) is due to the low prior that a hypothesis such as 'a face-house' would have in our world—how often do you see a face superimposed on a house?[9]

Though this example is insufficient on its own to fully account for the claims made by the BB hypothesis or PP, Hohwy, Roepstorff, and Friston (ibid.) argue that it is able to jointly explain factors about binocular rivalry (i.e. alternation and selection) that were hitherto accounted for separately, and in ways that were often difficult to reconcile. Therefore, the virtue of their explanation's simplicity, when situated in the wider explanatory scope of PP more generally, offers a compelling reason to take

[9] Interestingly, this may also explain the regularity with which we report seeing things such as faces in inanimate objects (also see Clark, 2016b; Hohwy, 2013, for a discussion of how PP can account for other perceptual illusions).

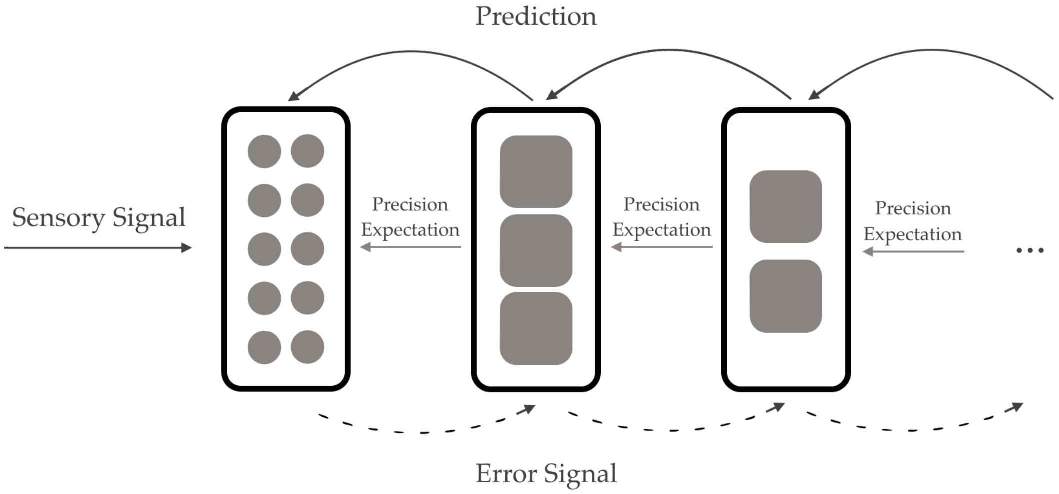


Figure 2.4: A simplified schematic of the principle of predictive processing with precision-weighted expectations about the incoming sensory evidence.

such a view seriously.

2.4 Precision-Weighting

An important component of the PP framework was missing from Figure 2.1, which is added in Figure 2.4. The importance of *precision expectations* is best seen when we note that not all prediction errors are created equally.

As prediction errors are responsible for the updating of generative models, it is important that those which are unreliable have a smaller impact. What does it mean to say that an error signal is unreliable? Consider the following scenario: you are in a noisy room trying to converse with a friend, and struggling to hear what they are saying. In this instance the auditory signal is less informative, but if you are adept at lip-reading you may be able to determine what is being communicated by paying more attention to the visual information. Conversely, imagine that you are in a darkened room. In this situation, it would be better to rely on touch than vision if trying to navigate to some region of the room.

The use of the term 'attention' in the above examples is no accident. In PP, attention is considered to be a process by which the brain increases the gain on prediction errors that are estimated to be the most informative (Feldman and Friston, 2010; Hohwy, 2012).[10] Those that are noisy (e.g. visual signals from dark room, or auditory signals from noisy room) carry greater uncertainty, and should not lead to drastic model revision. This is what it means to say that a prediction has low uncertainty (high precision); it is more informative, *ceteris paribus*, than a prediction that has high uncertainty (low precision) over a range of possible states. Importantly, this noise or uncertainty will be state-dependent, and therefore the precision-expectations should be conditioned on higher-level expectations of the current environment. Whether the sensory signal is a suitable indication of the actual state of affairs in the environment determines to what extent the models are updated.

How can a system learn and employ these precision-weightings? The answer is to again appeal to the hierarchical generative models. Firstly, given that these models are encoded in the brain as probability density functions, we can appeal to the variance of the functions as a measure of the expected precision of the sensory data—the inverse of variance is precision. This idea also receives support from the

[10] Admittedly, attention is a complex and multifaceted phenomena, and one may worry that by equating it with precision-weighting, some important nuances are missed. For example, how are covert and overt shifts of attention explained in terms of precision-weighting? How does the framework accommodate local versus global forms of attention (e.g. blocking out background stimuli in order to narrowly attend to a subtle stimulus as in mindfulness practice, versus a global situational awareness of many disparate stimuli as is reported in police and bouncers)? Unfortunately, a more detailed discussion on these points would be too tangential, and we therefore point the interested reader to (Hohwy, 2012; Ransom, Fazelpour, and Mole, 2016) for two different perspectives.

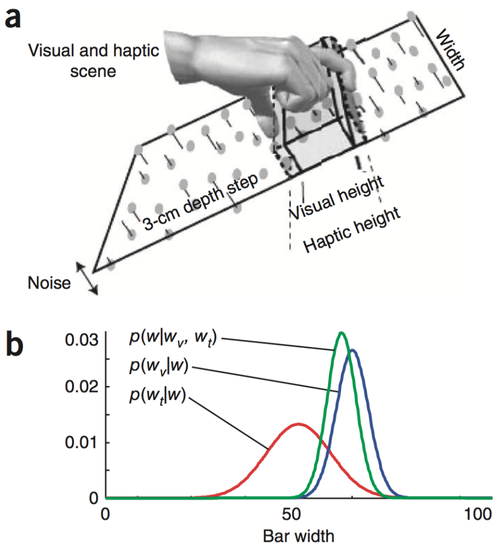


Figure 2.5: (A) In the study performed by (Ernst and Banks, 2002) subjects were required to estimate the width of a bar that could be touched and looked at. (B) The combined distribution over the estimated width of the bar (green curve) is a product of the visual (blue curve) and haptic (red curve) estimations. The combined distribution is shifted toward the more reliable (smaller variance) input (i.e. vision). Reprinted from (Pouget et al., 2013).

BB approach to cognition, and specifically from work on *optimal integration*.

Ernst and Banks (2002) found that humans are able to optimally combine different sources of sensory input, which vary according to how precise the information is from each sense modality. Figure 2.5 depicts a simple example where two distributions representing different sources of sensory information are integrated into an estimation of a single variable (in this example the width of a bar). Each initial estimation is weighted according to the reliability of the information source. In the example depicted, the distribution corresponding to the haptic information has a greater variance, and is therefore considered less reliable.

In the case of PP, an analogous situation occurs when a prediction is compared with a corresponding error signal. Each generative model encodes additional precision expectations of how precise the error signal is expected to be in order to optimally combine the predictions with the incoming error signals. However, as we just discussed, whether a certain input (or indeed sense modality) is reliable is state-dependent (i.e dependent on the type of environment the agent is in). Therefore, it is also important that these precision expectations are conditioned upon higher-level expectations, which provide contextual constraints on the sorts of precision expectations selected at each level in the hierarchy. As with before, the PP proponent appeals to the empirical Bayes schema that is implemented by hierarchical generative models, where the higher layers act as hyperpriors[11] that flow top-down through the hierarchy, contextualising the lower layer precision expectations. In the case of higher-level expectations concerning low visibility, precision expectations of incoming sensory data from the visual system should be adjusted accordingly, in order to

[11] Not to be confused with hyperparamaters, which are parameters of prior distributions. In contrast hyperpriors are prior distributions on a hyperparameter. In PP hyperpriors are employed as higher-level priors regarding precision-expectations, whereas hyperparameters are higher-level parameters that act as predictions for lower levels.

avoid unncessary model revision. This addition of precision expectations to the PP framework should not be taken as shifting an emphasis entirely onto the precision of error signals—variability is not the only factor relevant for model revision. The brain must carefully balance the predictions, precision expectations and error signals in order to minimise prediction error most effectively. Nevertheless, the inclusion of this additional mechanism is an indispensable component of the PP framework, and we will have a lot more to say about it in chapters 4 and 5. It is vitally important for an agent to be able to determine whether its predictions fail to account for the sensory inputs because they are disconfirmed by it (i.e. genuinely inaccurate) or because the sensory inputs are too noisy. Reliable belief revision should only be made on the basis of the former. By including estimates about the reliability of an error signal, and weighting them accordingly, an agent can more effectively modulate its learning and future interactions with its environment.

2.4.1 Evidence from Neuromodulation

It was stated in the previous section that by weighting the precision of prediction errors, attention is able to modulate the influence that they have for ongoing inference and learning. A number of studies have recently shown that this is equivalent to the alteration of synaptic gain on specific sensory neurons (here understood as encoding prediction errors) (Feldman and Friston, 2010; Ouden, Kok, and De Lange, 2012). Moreover, a number of studies performed by Kok, Jehee, and De Lange (2012) and Kok et al. (2011) have shown how the silencing of upwards propagating error signals by successful predictions can be reversed by increased attention to those same regions, which by contrast enhances the activity in those same sensory regions. Which mechanisms could be responsible for this attentional enhancement?

Friston et al. (2012) have proposed that the variance or uncertainty associated with a prediction error could be encoded by synaptic gain, and that key neurotransmitters such as dopamine may play an integral role in modulating this gain. In effect, this means that the dopaminergic system contributes to controlling the precision of sensory cues that are responsible for model revision, and as we will see later, engendering action (section 2.6). Given that we will be exploring this notion in significant depth in chapters 4 and 5 we will postpone any further discussion or empirical evidence until then.

2.5 Self-Evidencing

Generative models that are successful in explaining away the sensory signals (i.e. minimising prediction error) can be said to generate their own evidence for their success—they are *self-evidencing*. This is illustrated in an example Hohwy adapts from (Lipton, 2004). Suppose you look out from your window on a snowy morning and observe footprints in the snow that has settled on your lawn. In attempting to explain the occurrence of the footprints, you form the hypothesis that a burglar attempted to break in during the night. If someone were to ask you what evidence you have for this hypothesis, you would be justified in pointing to the occurrence of the footprints, despite the fact that this is the very evidence that initially led to the formation of the hypothesis. Though it has the appearance of circular reasoning, this form of inference is a common epistemic practice according to Hempel (1965), and is what he describes as a *self-evidencing* explanation.[12] It is also an important component of the notion of hierarchical Bayesian inference that Hohwy (2014, p. 6)

[12] This phrase is initially confusing, and may give the appearance of conflating the notions of hypothesis and evidence. However, it should not be interpreted as arguing that the hypothesis provides evidence for itself. Instead, the evidence that supports the hypothesis is also the same evidence that leads to the production of the hypothesis in the first place. As such, the hypothesis and the evidence are still distinct.

claims characterises PP:

> “The internal model that generates hypotheses that over time makes the evidence most likely, and does so most precisely and simply, will have its own evidence maximized. That is, as a model generates hypotheses that explain away occurring surprising evidence (i.e., minimize prediction error) it maximizes the evidence for itself. Prediction error minimization thus constitutes self-evidencing.”

We here begin to see the roots of the neurocentricism at play in Hohwy’s account of PP. According to Hohwy (ibid.), the hidden causal structure of the world is always being inferred by the brain from within what he terms the “Evidentiary Boundary.” It is the existence of this boundary, in combination with the emphasis on self-evidencing that entails Hohwy’s neurocentricism. In addition to the aforementioned support this picture receives from BB, the idea of an evidentiary boundary finds additional theoretical support, as well as a mathematical generalisation, from a theory known as the *free-energy principle*.

The free-energy principle states that any (ergodic) self-organising system, which can be described in terms of a Markov blanket, will appear to model and act on its world to preserve its functional and structural integrity. This unfolds in virtue of the minimisation of an information-theoretic measure (free energy), which bounds surprising sensory states (see section 2.2.1) for the system, and in turn leads to homoeostasis (e.g. Friston, 2010, 2013). It has been formally shown how the theory can provide a unifying account that bridges many disciplines (e.g., Bayesian inference, expected utility, information entropy, and optimal control), and it should also be noted that Hohwy (2014) has acknowledged the importance of the free-energy principle in providing theoretical support for the PEM account. This is because under simplifying assumptions, free-energy minimisation can be reformulated as prediction-error

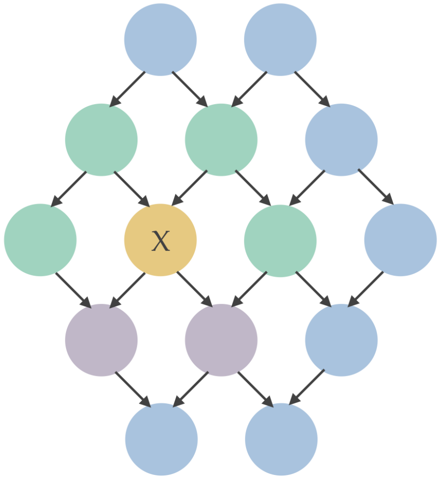


Figure 2.6: A Markov blanket defined over the node X. The Markov blanket consists of the parent nodes that X is dependent on (green nodes), the child nodes that are dependent on X (the purple nodes), and the remaining parent nodes of X's children. The "inferentially secluded" or independent nodes are the blue nodes that are separated from X by the aforementioned Markov blanket.

minimisation (Hohwy, 2013, p. 52). In what follows, many of the technical details have been omitted, and we refer the reader to key papers (e.g. Friston, 2010, 2013) for further information.

The fundamental notion to look at is Hohwy's reliance on a Markov blanket. If the future value of the state of a system can be determined based solely on the value of the present state of a system, and no further knowledge of the past states would change this value, we can say that such a system satisfies the *Markov property*. Now consider a complex system composed of many interacting nodes (variables). Pearl (1988) demonstrated how the Markov property could be extended to these more complex systems (e.g. a Bayesian network), leading to the notion of a Markov blanket. The graph in figure 2.6 depicts a highly simplified network comprised of nodes (coloured circles) and connecting edges (directed arrows), which represents a set of random variables (the nodes) and the *conditional dependencies* between them (the edges). There is also a quantitative component that represents the strengths of the conditional dependencies (not included). Within a Bayesian network, a Markov blanket is defined over a node X; the set of nodes that comprise its parents (i.e. the green nodes that X is dependent on); its children (the purple nodes that are dependent on X); and the other parents of all of its children (the remaining green nodes). Any nodes in the network that fall outside the scope of the Markov blanket are *independent* of X when conditioned on the set of nodes that comprise the Markov blanket. A Markov blanket thus creates a partition of states into inner states and external states, such that learning information about any of the external states will give no further information about the internal states. In short, the Markov blanket is defined for some given node X, such that the value of X is fully determined (and could be predicted) by knowing just the values of the nodes in the Markov blanket.

The notion of a Markov blanket helps to make precise Hohwy's commitment to a neurocentric boundary for the mind, or what he terms an "evidentiary boundary", as any state within the Markov blanket is "inferentially secluded" from the states on the other side. He states his claim in two ways. Firstly:

> "[...] the mind begins where sensory input is delivered through exteroceptive, proprioceptive, and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord." (Hohwy, 2014, p. 18)

Then, in a footnote to the above quote:

> "In slightly more technical terms (Friston, 2013), the sensory input and output at this boundary forms a so-called Markov blanket (Pearl, 1988)

> such that observation of the states of these parts of the system, together with observation of the prior expectations of the system in principle will allow prediction of the behavior of the system as such. Causes beyond this blanket, such as bodily states or external states, are rendered uninformative once the states of the blanket are known." (Hohwy, 2014, p. 25)

The parameters in the models of the brain are thus considered inner states, whereas the hidden states of the environment (including the body) exist on the other side of the boundary that is induced by the Markov blanket. According to Hohwy, by describing the brain in terms of a Markov blanket (or Evidentiary Boundary), the picture of the mind that falls out is one that appears to be "neurocentric" (ibid.). Anything outside of the brain must of necessity be deemed "inferentially secluded" from the internal models, and is treated as a "hidden cause" that must inferred by the brain. We can construct Hohwy's argument as follows:

1. The existence of a Markov blanket entails an evidentiary boundary between the inner states of a system and its external environment.

2. An evidentiary boundary requires the inner (generative) models of a system to be self-evidencing (i.e. to generate their own evidence).

3. If the brain is a self-evidencing system, then it must infer all external causes about the incoming sensory information from within the evidentiary boundary.

4. The evidentiary boundary defines the mind-world relation, opens the door to skepticism, and entails a neurocentric perspective where the mind is inferentially secluded from its environment.

5. PP implies that the brain is a self-evidencing system that generates hypotheses about the world from within an evidentiary boundary.

6. The mind is therefore inferentially secluded from the world, and forces us to resist conceptions of the mind where it is embodied or extended.

There are a number of implicit assumptions in the above formulation, and therefore, a number of areas to take issue with. The first, and perhaps most obvious objection to the above is to undermine the notion of a Markov blanket as employed by Hohwy. At present, Hohwy's commitment to the notion is based largely on the theoretical work of Karl Friston and the free-energy principle (FEP) (Friston, 2013). Though persuasive, inasmuch as it rests on some compelling theoretical modelling that demonstrates the wide explanatory scope of the FEP, it is not without its conceptual worries. Some of these worries have even been expressed by Friston himself. For example:

> "[...] is there a *unique* Markov blanket for any given system? [...] a system can have a multitude of partitions and Markov blankets. *This means that there are many partitions that—at some spatial and temporal scale—could show lifelike behaviour.* For example, the Markov blanket of an animal encloses the Markov blankets of its organs, which enclose Markov blankets of cells, which enclose Markov blankets of nuclei and so on [...] there are probably an uncountable number of Markov blankets in the universe. (ibid., p. 10, emphasis added)

In the case of PP, this issue is particularly pressing due to the framework's commitment to hierarchically-organised, generative models.

Many of the models PP posits exist at specific levels in the hierarchy and only model the neural activity at the level below them (Friston, 2008). As a result, the overwhelming majority of modelling is intra-neural. Only the most peripheral layers of the hierarchy directly model anything beyond the brain, and these operate at extremely small spatial and temporal scales (Kiebel, Daunizeau, and Friston, 2008).

As such, all that they can be said to model (or predict) are fleeting moment by moment impacts on small regions of our sensory receptors. As Hohwy (2014, p. 15) acknowledges, in principle we could isolate the entire system minus the most peripheral layer, and we would still have a prediction-error minimising system, complete with its own evidentiary boundary that separates it from the external world plus the peripheral layer. This process could be repeated, leading to a proliferation of nested hierarchical models, each with their own evidentiary boundaries. This is problematic for Hohwy's account, as it requires him to provide a reason for privileging any of these possible boundaries as the one that defines the mind-world boundary. His favoured solution is to:

> "[...] rank agents according to their overall, long-term prediction error minimization (or free-energy minimization): the agent worthy of *explanatory focus* is the system that in the long run is best at revisiting a limited (but not too small) set of states. It is *most plausible* to think that such a minimal entropy system is constituted by the *nervous system* of what we normally identify as a biological organism: shrinked agents are not able to actively visit enough states, and extended agents do not maintain low entropy in the long run." (ibid., p. 16, emphasis added)

We can respond to this suggestion in a number of ways. As we have already seen, Hohwy favours a neurocentric perspective, where "[...] the mind begins where sensory input is delivered through exteroceptive, proprioceptive, and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord." However, one may argue that any attempts to delineate the mind from the world, or the cognitive from the non-cognitive are simply doomed to failure at the outset. An example of such a view comes from Ross and Ladyman (2010, p. 156), who claim that there is simply no scientifically credible basis for delineating a cognitive from a non-cognitive system as in the proposal above. They state:

> "Modelers will and should draw system boundaries in whichever ways maximize efficient capture of local phenomena. Of course, as models are aggregated into more general theoretical perspectives, local optima should often be expected to be sacrificed for the sake of more parsimonious and powerful global models. But this is compatible with the suggestion that even a fully general theory of cognition—as information processing by relatively autonomous goal-driven systems—need incorporate no single overarching account of limits on the boundaries of cognitive systems. A cognitive system might simply be anything described by the hypothetical fully general theory, and be open to limitless cross-classification with respect to biological or chemical (etc.) principles for system identification."

Ross and Ladyman view their position as being opposed to any thesis that attempts to locate the mind, whether it be within the head, outside the head, or dynamically shifting across the skin-skull boundary. Their justification for this is that "composition in real science, as opposed to in metaphysics and stylized science, is usually a dynamic and complex idea that does explanatory work by reference to distinctive features of specific applications." (ibid., p. 160). An example of a "dynamic and complex idea" that they cite is the identity relation 'water is H2O'. Instead of being identifiable as a synchronic relation, water is composed by oxygen and hydrogen in *various* polymeric forms that are constantly forming, dissipating and reforming over short time scales. Only in this more dynamic manner, and from a diachronic perspective, do the familiar macroscopic properties of the kind *water* arise. The synchronic description, they argue, therefore misses a rich (albeit currently incomplete) scientific picture. The restrictive boundary advocated by Hohwy suffers from the same problem—the concept of the mind, as with the multi-disciplinary approach of the cognitive sciences, calls for more than one overall explanatory perspective, and by proxy no single physical boundary.

A further concern is that despite providing an appealing answer to the aforementioned worry of nested agents, there are a number of problematic assumptions with Hohwy's favoured solution. Firstly, he states, "It is *most plausible* to think that such a minimal entropy system is constituted by the *nervous system* of what we normally identify as a biological organism". However, no justification is given for why we should agree with the "most plausible" qualifier. Hohwy simply points to an argument by Friston in support of the claim. Interestingly, in the cited paper, Friston (2013) raises similar worries about the answer to whether there is a unique Markov blanket for any given system. Although, in line with Hohwy, he appeals to the statistics of the Markov blanket to speculatively claim that the system with the lowest entropy is perhaps the agent of interest, he equates this with the biological organism, rather than the nervous system. It is unclear, therefore, why we should accept Hohwy's claim that the states that are revisited most over time are those of the nervous system, rather than those of the body.

Secondly, Friston (ibid., p. 10) acknowledges that "a system can have a multitude of partitions and Markov blankets. This means that there are many partitions that—at some spatial and temporal scale—could show lifelike behaviour", and, therefore, "minimum entropy is clearly not the whole story". Taking each of these points in turn, it is important to first see what is meant by the claim that a system can have a multitude of partitions. It is increasingly common in the cognitive sciences to see the employment of formal methods that were initially developed in systems biology. This is particularly helpful in the case of evaluating formal models that aim to capture specific cognitive phenomena. In this manner, one begins by appealing directly to systems biology to uncover and identify the sorts of variables (or states) that are relevant to the modelling of a situated agent in question. As an example

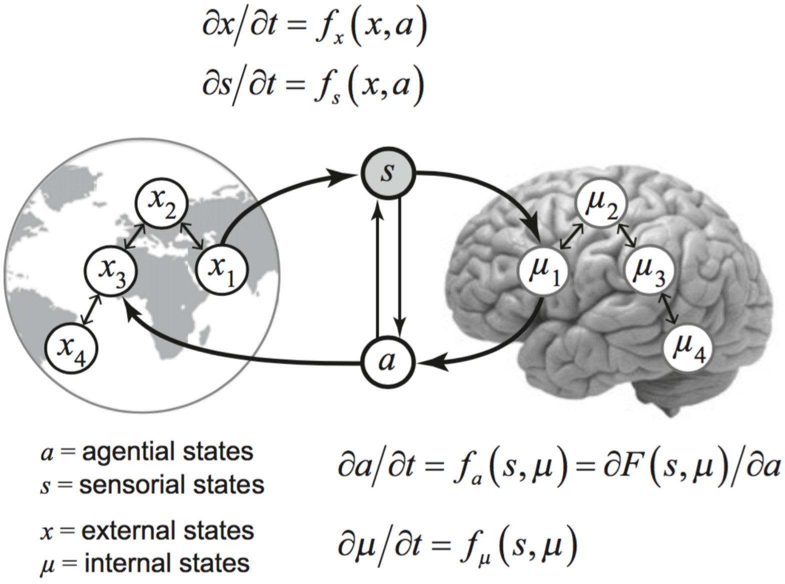

$\partial x/\partial t = f_x(x,a)$
$\partial s/\partial t = f_s(x,a)$
s
x_2
x_3
x_1
x_4
a
μ_2
μ_1
μ_3
μ_4
a = agential states
s = sensorial states
x = external states
μ = internal states
$\partial a/\partial t = f_a(s,\mu) = \partial F(s,\mu)/\partial a$
$\partial \mu/\partial t = f_\mu(s,\mu)$

Figure 2.7: A partition of states for a system that acts on its environment. Reprinted from (Kilner et al., 2016, p. 164).

of this, Kilner et al. (2016) offer the partition depicted in Figure 2.7. This partition captures an agent that acts on its environment (α). It considers the distinction between external states of the world (x), which are hidden from the internal states of an agent (μ) by the sensory states (s), in the same manner as Hohwy's account of PP. Once a partition of states has been identified, it is then possible to make use of various optimality principles to define rational "as if" theories of cognition, which themselves are concerned with one of the states being optimised—there will typically be multiple, sometimes competing rational theories for any given situation. Finally, each theory leads to a number of hypotheses that realise the optimisation by way of certain processes, and these process models are tested according to the empirical behaviours that they predict.

With this process laid out, it is clear that the neurocentricism inherent in Hohwy's account can be traced to the initial partitioning of states, which leads to the formation of a Markov blanket, and thus the separation of inner states (i.e. neural states and processes) from outer states (i.e the world). However, we can again ask what the justification for this initial partitioning is, and whether there are alternatives that are consistent with the PP story. Kilner et al. (ibid., p. 163) claim that the partition depicted in figure 2.7 is “necessarily implied by a system that is acting within its environment”, but again do not consider whether it is appropriate to focus only on the brain. Nevertheless, there is an important piece of the picture that we have hitherto been missing that they raise—we have said nothing of how PP accounts for action. However, as we will see adding action to the picture does not help Hohwy's case for neurocentricism, but in fact opens up the path to a truly embodied account of PP.

2.6 Adding Action to the Picture

In PP, perception, cognition and action are unified by the underlying imperative of prediction-error minimisation (PEM). PEM can be understood in a number of ways, two of which are noteworthy here. According to Hohwy (2013) either the system can update the parameters of the inner model in order to generate new predictions about what is causing the incoming sensory data (what he refers to as ‘perceptual inference’), or it can keep the generative model fixed, and resample the world such that the incoming sensory data accords with the predictions (what he refers to as ‘active inference’). Why is this?

In a particularly lucid account of the mechanisms underlying PEM, Hohwy presents PEM by comparison to scientific hypothesis testing. To begin, he demonstrates the inadequacy of passive evidence accumulation (taken on its own) for hypothesis selection by drawing a parallel with the debate between associationist statistical inference and causal inference. The former observes mere associations in data (e.g. between two random variables), but is unable to distinguish whether X causes Y, Y causes X or if they have a third common cause Z. The latter by contrast sees intervention as a fundamental tool for discerning causal relations between two variables, e.g. if intervening on X has an effect on Y, but not vice-versa, then X is a cause of Y (cf. Woodward, 2003). It is in this latter manner of hypothesis-testing that Hohwy sees a natural place for action.

Action, according to Hohwy, is a form of intervening on the interacting hidden causes in the world in order to test perceptual models, and is therefore a necessary companion to perception, which is otherwise "hostage to the whims of the incoming sensory data" (Hohwy, 2013, p. 76). Moreover, in PP, action is accommodated as a form of statistical inference in its own right, known as active inference, which assists in the overall process of prediction-error minimisation by resampling the world to further test the inner generative models. As Hohwy states:

> "Action makes decent inferences better. For example, I am more confident I am looking at a man's face after successful active sampling of the world according to this hypothesis. This helps decrease uncertainty especially in cases where the winning hypothesis did not have a very much higher posterior than its competitors at the outset." (ibid.)

For Hohwy then, action and perception are intimately related in respect of the underlying imperative to minimise prediction-error. Insofar as PEM is concerned, perception equates to forming or selecting better hypotheses about how the world is on the basis of sensory evidence (perceptual inference), and action equates to intervening on the world to better test and select competing hypotheses (active inference). What is the support for such a picture as it applies to the brain and motor control?

The first point to make is that it is not merely the world that causes our sensory inputs. Our actions in the world have important effects on the changes in sensory input as well. For example by moving my head relative to the objects that are situated on the desk in front of me, previously hidden features come into view (e.g. the initially occluded handle on my mug). In this manner, behaviour can be seen as the control of perception, to borrow a phrase from Powers (1973). Recall that in the case of PP, perceptual experience is determined by successful predictions of sensory input (e.g. binocular rivalry). Importantly, the sensory input that the brain receives is not merely exteroceptive (originating from the outside world), but also extends to include proprioception (sensation of the position and movement of the body) and interoception (sensation of the internal physiological states of the body) that will be affected by action. Any unexpected (or surprising) sensory input, regardless of its source, generates prediction error that propagates upwards through the hierarchy, and the primary task of the brain is to minimise this prediction error generated by all types of sensory input (exteroceptive, proprioceptive and interoceptive). Friston has provided a formal basis for this picture, starting from the premise that adaptive agents must necessarily occupy a limited set of states as defined by their phenotype (Friston, 2010, 2013; Friston et al., 2010). These states are essentially a bounded region (or attractor in dynamical systems theory) of all the possible states an agent could be in, and in order to maintain homeostasis, and crucially avoid death, the agent should revisit these states most frequently. The most important surprising states (or those which generate the most prediction error) in terms of homeostasis are those that reflect unwanted changes in the organism's internal milieu:

> "The fixity of the milieu supposes a perfection of the organism such that the external variations are at each instant compensated for and equilibrated [...] All of the vital mechanisms, however varied they may be, have always one goal, to maintain the uniformity of the conditions of life in the internal environment [...] The stability of the internal environment is

> the condition for the free and independent life." (Bernard, 1974, quoted in Friston et al., 2010, p. 231)

However, a system that can only minimise prediction error passively (i.e. by updating its models) can do nothing to avoid those sensory states that indicate maladaptive situations (e.g. a fish out of water). By contrast, an agent that is able to actively navigate its environment, can utilise behaviour to control perceptual states, and thus minimise prediction error by resampling its world and avoiding states that are least desirable. One may worry that this addition implies that an agent should therefore simply navigate to a darkened room where no further sensory signals are received. However, two points can be made in response to this worry. Firstly, sensory input is here taken to include interoceptive information, which means that there will be an increasingly urgent signal from the body informing the agent to obtain food and water if it is to remain alive. Secondly, it is possible for this fact to be learned, such that higher-level contextualising predictions may override the lower-level tendencies to simply seek out a dark-room, and the agent may not expect itself to inhabit these types of environment (Friston, Thornton, and Clark, 2012). In short, real 'dark-rooms' simply do not exist in nature (aside from a state of death). Putting aside the validity of these assumptions, it is important to note that for Hohwy's claim of PEM as a unifying mechanism to be warranted it is vital that the mechanisms responsible for implementing perceptual inference (e.g. predictions, error signals and precision-expectations) apply equally to active inference.

To demonstrate why this is in fact the case, it is first helpful to note that the motor system is also structured hierarchically, in much the same way as visual cortex. This allows organisms with the appropriate neural structures to control behaviours in a similarly hierarchical manner, with higher levels specifying more abstract plans (e.g. make a cup of coffee) that can be unpacked into more finer-grained motor behaviours at lower levels (e.g. grab kettle). Recent neuroanatomical evidence paints an interesting picture of how this process unfolds in the brain. For example, Adams, Shipp, and Friston (2013, p. 1) argue that "descending projections from the motor cortex are, anatomically and physiologically, more like backward connections in the visual cortex than the corresponding forward connections." Furthermore, (Friston, Mattout, and Kilner, 2011, p. 138) state:

> "The primary motor cortex is no more or less a motor cortical area than striate (visual) cortex. The only difference between the motor cortex and visual cortex is that one predicts retinotopic input while the other predicts proprioceptive input from the motor plant."

What this means is that top-down signals in both visual and motor cortex are functionally similar, and within PP this translates to a shared commitment to the prediction of incoming swathes of sensory information (albeit from different sources). But how could motor control be determined by predictions?

The view defended by proponents of PP, resembles the ideomotor theory attributed to William James, and developed more recently under the guise of the theory of event coding (Hommel et al., 2001). Ideomotor theory claims that thoughts or mental representations, when unimpeded by other factors (e.g. inhibitory mechanisms), can cause a corresponding muscular action by activating reflex arcs. PP makes use of this principle in a novel way. Under the previously mentioned label of active inference, PP claims that descending predictions in motor cortex aim to predict the incoming sensory data from ascending proprioceptive signals. However, in the case of desired movements (i.e. those not yet currently obtained), the error signal will obviously be high as the incoming sensory information (error signal) will not correspond to the desired state. Consider the following: if I predict that I am holding a mug, but the mug is in fact on the desk in front of me, I am doing a poor job of predicting the proprioceptive (and indeed exteroceptive) signals. The trick, according to PP, is to then minimise the prediction error, not by updating the internal models (perceptual inference), but by allowing the descending motor predictions to cause the necessary motor behaviours that will bring about the desired sensory state that matches the agent's expectations. To bring an action about, motor cortex responds to the incoming error signals by temporarily down-weighting the associated precision expectations for proprioceptive feedback, and responding with the desired (and previously learned) control trajectories that lead to the desired state (more will be said about this in chapter 4). These predictions thus have a subjunctive nature—they don't merely make predictions about what probably will happen, but make predictions about various things that *would* happen conditional on an array of *possible* actions (i.e. what perceptual states are expected if this behavioural routine is performed).

The consequence of this view, as argued by (Adams, Shipp, and Friston, 2013, p. 4) is that the "perceptual and motor systems should not be regarded as separate but instead as a single active inference machine that tries to predict its sensory input in all domains: visual, auditory, somatosensory, interoceptive and, in the case of the motor system, proprioceptive." We here see the dissolving of any clearly delineated computational boundaries between perception and action, although as both (Hohwy, 2013) and (Clark, 2016b) acknowledge, there remains an important difference in direction of fit. Nevertheless, with the dissolution of these boundaries, there begins to emerge an obvious challenge to the classical cognitivist picture introduced in chapter 1. However, before we explore how PP overcomes this challenge by connecting with work in embodied cognition, we shall turn in the next chapter to look specifically at decision-making. This will provide the main focus for discussion in later chapters.

Chapter 3

Is Decision-Making Embodied?

> "[...] the concepts of separate perceptual, cognitive and motor systems, which theoretical neuroscience inherits from cognitive psychology, are not appropriate for bridging neural data with behaviour." (Cisek, 2007, p.2)

The primary concern of the current chapter is to explore some different proposals in decision theory regarding how we make decisions, and the mechanisms by which we do so. Decision theory is often considered an interdisciplinary project to which philosophers, economists, psychologists, neuroscientists and statisticians, among others, contribute. It can also be separated into *descriptive* and *normative* approaches, where the first is viewed as an empirical approach that aims to provide an account of how decisions are made, and the second is understood as providing prescriptions for what decision-makers are *rationally required* to do (Peterson, 2009). These two approaches are often considered independently of one another, and as we are interested in the mechanisms that underlie decision-making, the focus of this chapter will be on *descriptive* decision theory.

Section 3.1 begins by highlighting some of the limitations of adopting a traditional decision-theoretic account for modelling real-world behaviour due to underlying cog-

	Heavy Traffic (30%)	Light Traffic (70%)
Route A	24 minutes	14 minutes
Route B	18 minutes	17 minutes

Table 3.1: Traffic Example

nitivist assumptions. These limitations have led some researchers to turn away from the underlying cognitivist assumptions inherent in classical decision theory, towards a notion of embodied decisions (Cisek and Pastor-Bernier, 2014; Cisek, 2012; Lepora and Pezzulo, 2015). In section 3.3, we will explore this more recent embodied approach, which views decision-making as inextricably intertwined with sensorimotor processes, and is contrasted with the neuroeconomics approach outlined in section 3.2.

3.1 The Traditional Cognitivist Account of Problem Solving and Decision-Making

Consider the following decision: you must choose between two routes to work. Route A takes you through a city that has a high risk of heavy traffic, but is short in distance. The other route is less likely to be affected by the increased congestion, but is longer than the former. Suppose you know from previous experience that, given the time you are leaving, it is more likely that the traffic will be light, and your preference is always for the shortest time spent travelling. What should you do?

Table 3.1 represents a decision under risk. Here, the agent has full knowledge of the available options and probabilities attached to the relevant states. In situations like this, deciding what to do is relatively straightforward, and a number of decision rules exist that provide guidance for what is rational to do in these situations. For example, the principle of *maximising expected utility* (MEU) would suggest taking Route A, as the following demonstrates that it has the shortest expected duration (and therefore the greatest expected utility, assuming that utility is a negative linear transform of duration):

Expected Duration of Route A = 0.3 x 24 + 0.7 x 14 = 17
Expected Duration of Route B = 0.3 x 18 + 0.7 x 17 = 17.3

Savage (1954, p. 16) famously referred to these situations as 'small worlds', where it is possible to "look before you leap", by which he meant an agent has knowledge of the states of the world and all of the options available to her. Even in cases where the probabilities attached to the states are unknown (decisions under uncertainty), many decision-theoretic norms (e.g. dominance and subjective expected utility maximisation) exist to help guide this process. However, unlike small worlds, the real world is not so neatly circumscribed. Instead, most everyday decisions can be viewed as 'large worlds'.

Unfortunately, there are a number of assumptions that hold in small worlds, that might not hold in large worlds. For instance, in small worlds, the agent has knowledge of the options available to here (e.g. Routes A and B). In addition, there is a clear set of possible worlds (e.g. set of Heavy Traffic worlds and Light Traffic worlds), and the agent knows for sure that she falls into one or the other of these, but not both—they form a partition of states. Finally, the agent knows the utilities assigned to each of the cells within a given world. In large worlds, any of these assumptions may fail. For instance, recall the example in the introduction of choosing an action to perform in light of the increasing feeling of tiredness while writing the paper. As we saw, there was always the possibility that some other unconsidered option exists, which may or may not have a higher utility than those considered. That is, you were unable to come up with an exhaustive partition, including the set of possible worlds and their corresponding utilities. Simply tagging on the state ‘something else’ doesn’t solve the issue, as such a state is likely heterogeneous—that is, it could contain worlds that are incommensurable with one another. This level of uncertainty is a serious challenge for decision theory, as the possibility of framing a genuine decision problem requires that an agent already has options to deliberate over. Even hallmarks of rationality such as Bayesianism have been criticised as inapplicable in these types of large worlds (Binmore, 2008).

It may be argued that this is not really a problem for decision theory per se. On this line of thought, the issue of determining options is a problem for the perceptual system to initially solve, whereas decision-making, which is decomposable into a process of *deliberation* (i.e. calculating the values of the relevant decision variables) and *commitment* (i.e. selecting an action) merely evaluates the presented options. As such, the brain is faced with the task of constructing a representation of features of the environment, which can then be used as the basis for making decisions (along with abstract representations of related decision variables such as expected gains or potential risk). Furthermore, behaviour is simply the means by which a decision is reported, and can be used to reveal an agent’s preferences (Sen, 1971). We wish to resist this characterisation.

This account of decision-making is based on a number of cognitivist assumptions, which are nicely captured by Hurley’s (1998, p. 401) critique of the “classical sandwich model” of the mind (Figure 1.4), which we first saw in Chapter 1. Recall, in this model, the outer slices of perception and action are peripheral to the inner filling of cognition, and thus separate from one another. They are also separate from cognition, which interfaces between perception and action. First, perception builds a reconstructive representation of features of the external world. These discrete, abstract representations are then transformed by cognitive processes into a motor plan for action, according to the agent's beliefs and desires, and subsequently carried out by the motor system. Within this model, decision-making would reside within the middle box, and deliberation and commitment could take place in some 'central executive region' such as the prefrontal cortex, which could integrate relevant information from other systems such as working memory (Baddeley, 1992).

Hurley saw a number of problems or limitations with this account, and a similar set of problems can be uncovered by exploring the more general notion of problem solving.[1] Kirsh (2009) discusses the traditional view of problem solving, in which agents first delineate the problem to be solved by constructing a representation of it—a *problem space.* The problem space could be represented in an abstract manner by using a graph with nodes and edges that determine the possible states of the problem, and the connections between them. Solving the problem is then understood as the deployment of various rules to *search* the possible paths in the problem space, moving from an initial state (or the current state) to some desired goal-state.

This method was recognised by Simon and Newell (1971) as a fruitful way of visualising the task faced by agents in idealised situations. Mindful that science often starts from idealisation, Simon and Newell sought to place problem-solving on the same firm-footing, with the intention of generalising from well-defined problems to a broader class of ill-defined cases. They took games and puzzles to be a hallmark type of problem solving that could be treated as the well-defined type of problems to be studied, due to a number of salient properties they possess. Firstly, the rules

[1] We assume here that many decisions can be viewed in terms of problem solving. For example, the decision introduced at the start of this thesis regarding what to do about the unfinished paper is easily recast as the search for the optimal action that leads to the most productive solution. This is not to deny important differences between problem solving and decision-making, but by exploring their commonalities, we stand to gain a greater understanding of some of the challenges faced by the traditional cognitivist picture.

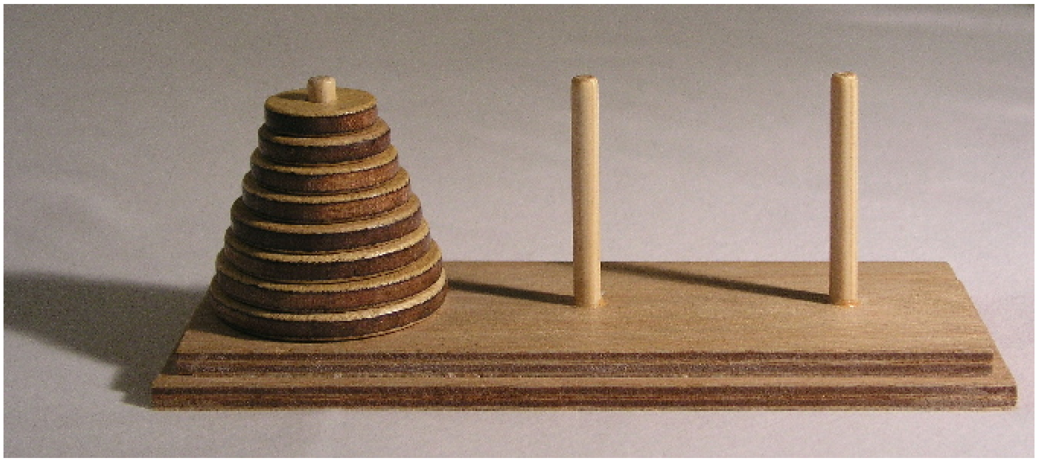


Figure 3.1: Tower of Hanoi Game

Image reprinted from Wikimedia Commons under Creative Commons Licence 3.0: `https://commons.wikimedia.org/wiki/Category:Tower_of_Hanoi` [Accessed: 16/08/16]

of games and puzzles are self-contained, generating a well-defined object of study for the experimenter, which Simon and Newell (ibid.) termed a *task environment*—an abstract structure that corresponds to the problem space of the agent.[2] Secondly, puzzles and games are easy to represent abstractly, and can often be instantiated in various physical forms. For example, consider the well-known Tower of Hanoi puzzle, in which a stack of discs must be moved from one peg to another, such that the order of the discs in the initial state is replicated at the goal-state (see Figure 3.1). It is relatively easy to formulate an abstract representation of this task (see Figure 3.2).

From the perspective of the experimenters, the abstract structure of the task environment allows for specific behaviours to be deemed irrelevant in the experimental setting. For example, an agent scratching their head is not considered to be a task-

[2] Interestingly, as Kirsh (2009) notes, the use of the term 'environment' was selected to acknowledge that subjects who improve their performance, are in some sense *adapting* their behaviour to the constraints of the task environment.

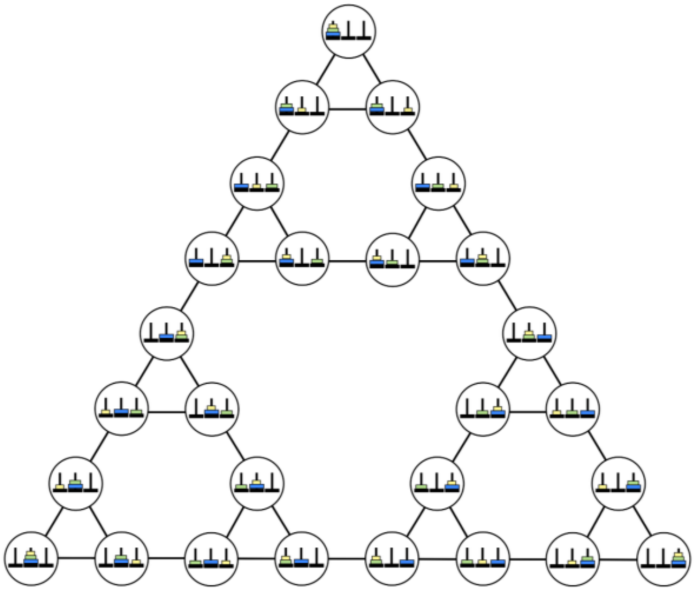


Figure 3.2: Tower of Hanoi Problem Space - each node represents a possible state of the game and the edges denote the legal moves between them.

Image reprinted from: `http://www.suffolkmaths.co.uk/pages/images/Hanoi.png` [Accessed:16/08/16] relevant behaviour, based on the structure of the task environment. Rather, what is studied is the method by which agents search the problem space, and the operations or methods they perform to move towards the goal-state. Thus, problem-solving is understood as the method of search that is performed once an adequate representation of the problem space is generated by the agent (Kirsh, 2009), in much the same way as the deliberation and commitment stages in a decision task occur once the decision problem is represented. However, why should we think that this is a) the most interesting or essential part of problem solving or decision-making, and b) why should we think that it is encapsulated from other processes that both precede and succeed it?

Consider another problem. One of your colleagues is on holiday, and has handed an urgent administrative task over to you to complete in their absence. The task had been started prior to their departure, but is left unfinished. The problem for you is that they have not specified where in the process they got to before leaving. Which is the harder of the tasks: (a) determining where in the process they were before their departure, or (b) continuing with the process once you know where they were?

Arguably the former is harder, or at least on a par. And yet the former doesn't seem to be a case of searching a pre-defined problem space in the traditional sense, but is rather understood as merely the framing of the task environment before proceeding with the main task.

To highlight this, Kirsh (ibid.) labels the various modular components of a problem solving task as follows:

Framing: determining which states or processes in the world are salient to the task.

Representation: constructing an abstract structure of the problem to be solved.

Search: finding a (potentially optimal) solution to the task.

Registration: reinterpreting the results of the task and connecting the solution back to the physical world.

By decomposing a problem solving task into these modular components, we are able to explore some of the issues with the cognitivist conception of decision-making. For example, it is clear that before you can even begin the administrative task—perhaps by following a pre-specified set of rules—you must first determine whereabouts in the task your colleague left you.

3.1.1 Criticisms of the Classical Cognitivist Picture

The routine described above is informative as it exposes a number of issues with cognitivist conceptions both of decision-making and problem solving in general. Recall that, according to the classical sandwich model, perception and cognition are encapsulated from one another. In the above problem-solving routine this translates into the separation of framing and representation (purported constituents of the perceptual processes), from search and registration (purported cognitive tasks). Action would again be understood as the agent's way of reporting the solution in whatever manner is appropriate for the task. There are a number of issues with this picture:

1. How to account for cases of problem solving or decision-making that are ill-defined.

2. The assumption that the process of problem-solving or decision-making is serial and modular.

3. No explanation given for how agents solve the problems of framing and registration.

4. A failure to recognise additional behaviours or resources that may be part of the agent's method of problem-solving.

We will explore each of these points in turn.

Ill-Defined Problems or Decisions

By focusing on the well-defined task environments of games and puzzles, Simon and Newell (1971) wished to demonstrate that the classical theory of problem solving has a formal elegance, which lends itself to clearly defined experimental procedures. Experimenters could learn a lot by starting with the clear cases, prior to moving outwards to the ill-defined ones. However, the move towards ill-defined cases generates problems that expose some flaws of the traditional picture, as many real-world problems are difficult to represent abstractly for a number of reasons. Firstly, many problems have multiple goal-states, and no unambiguously right answer (e.g. getting from point A to point B may have multiple paths of equal distance). Secondly, some problems do not have a well-specified goal-state in advance of beginning the task, and part of the problem-solving routine may therefore be to find adequate solutions (e.g. creative or design problems such as cooking, music or painting.) Finally, other problems may have some vaguely defined goal-state, but no clear set of rules or operators to define the problem space (e.g. novel tasks that employ new methods). In many of these cases, it may be inappropriate to consider the problem solving task as a simple search for the right solution. Instead, focus should be drawn to the manner in which an agent decides to frame and represent the task environment.

Serial and modular

As mentioned previously, the classical cognitivist account takes framing and search to be separate, modular processes, where the latter is considered to be the real core of the problem-solving task. However, what is the justification for doing so? From the perspective of the psychologist studying the behaviour of a subject during the performance of some task, it may be convenient to break problem-solving down into

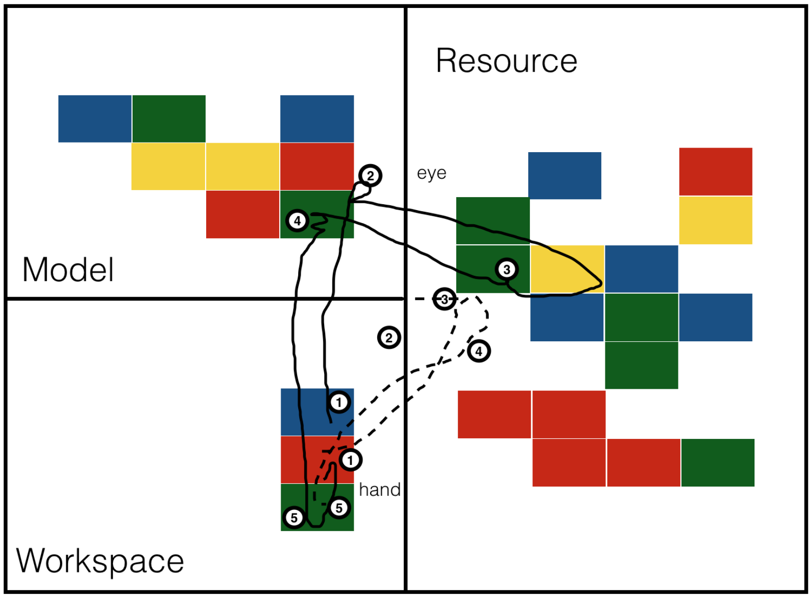


Figure 3.3: Subjects are required to copy the model by moving the blocks in the resource area over to the workspace. Numbers correspond to points in time. The dashed line responds to the subject's hand movement. The solid line responds to the subject's eye movements. Figure adapted from (Ballard, Hayhoe, and Pook, 1997).

modular, sub-routines, in order to isolate relevant variables for study. This perhaps explains why the increased interest in the novel formal methods originating in the 1960s (see chapter 1) made 'search' an easy target for cognitive science, thanks to an ability to construct algorithmic models that could explain how the process could be performed *efficiently*. However, this narrow focus on abstract symbol systems likely contributed to a failure to overlook important agent-level behaviours that play a significant role in the process of problem solving and decision-making.

The first thing to note is that in the absence of an external memory source (e.g. a pen and paper), searching an abstract representation of a problem space in working memory can be very demanding. However, depending on the resources available to an agent, there may be an alternative to constructing a highly-structured mental representation in the first place. A famous experiment that explored this was performed by Ballard, Hayhoe, and Pook (1997), who had subjects move blocks (in a computer program) from one area to another in order to construct a replica of a model that was displayed on screen (see figure 3.3). During the performance of this task, eye-tracking technology recorded where the subjects were looking. The classical problem-solving routine assumes an agent formulates a representation of the problem before executing the plan. However, the study performed by Ballard et al. recorded behaviours that were inconsistent with this account.

Their study found that numerous saccades to the model were made during the performance of the task, both before and after picking up a block. This suggests that the subject is only storing a minimal amount of information at any one time, either the colour or the position of the block to be copied. To test this, Ballard et al. switched the colour of one of the un-copied blocks while the subject was looking elsewhere (determined using eye-tracking technology). The assumption is that if a representation had been formed prior to the execution of the task (and consulted by working memory processes), then the end model would be inaccurate after a change of colour. However, the subjects were not found to make this mistake, suggesting again that only the current/next block is stored in working memory, and regular saccades are made between the areas throughout the task.

In this experiment, the usual process of 'search' can be viewed as an interactive process that makes regular call-backs to the world, challenging the idea that it is separate from representation and/or registration. Rarely do people solve problems like this in their head and then announce the solution all at once. Instead, they interact with the world to break the task environment into multiple sub-tasks, trialling different stages as they go along. Maintaining the strict demarcation of the purported modules is strained by examples such as the one offered by Ballard, Hayhoe, and Pook (ibid.), but is not entirely refuted. It is still possible to maintain that the serial and modular process occur in a cyclical fashion, looping through the various stages repeatedly. However, this is challenged by another criticism offered by Brooks (1991).

Brooks (ibid.) offered a criticism of what he termed the ‘sense-model-plan-act’ (SMPA) model of robotic intelligence. The idea that Brooks wished to challenge was that if a robot was a) required to gather information from its environment (sensing), in order to b) build a richly reconstructive representation (model), with which to c) formulate a plan of reaching some desired goal-state (plan), before d) effecting the necessary movements (act), then outside of a carefully designed and controlled laboratory setting, such a serial process would be insufficiently dynamic to cope with the time pressures of a constantly changing environment. In the time taken to build a model, the environment may have changed (e.g. the colour of a block may have changed), which would render the current model (and any actions based on it) inaccurate. Utilising the SMPA model in ecologically-valid scenarios would mean either the robot would incur an accuracy cost (subject to the environment changing) if it were to pass through the stages once before completing the full action plan, or it would incur a drastic speed cost if it cycled through the stages performing incremental, but carefully controlled-actions. Instead, Brooks’ suggestion was to implement a more straightforward sensorimotor coupling approach, where the internal models were replaced with a more direct sensitivity to the environment, and the environment directly elicited certain behaviours with no need for mediating representations. In his own words, “The world is its own best model.” (ibid., p. 15)

This worry about the urgency of performing an action in ecologically-valid scenarios is particularly pressing when applied to the case of decision-making. In traditional decision theory, models of decision-making do not incorporate the time constraints of agents, and therefore fail to account for a number of additional pressures that the agent is faced with. Gigerenzer and Todd (1999) refer to these types of models as instances of 'unbounded rationality' or 'optimisation under constraints'. The first finds its clearest expression in traditional forms of expected utility maximisation where an agent is expected to perform the full calculations required by rationality assumptions, and show “little or no regard for the constraints of time, knowledge, and computational capacities that real humans face” (ibid., p. 7). Of course, no one really defends the claim that humans indeed have the sort of Laplacean superintelligence that is required for these calculations, but rather defend them on either a normative basis, or on the basis that humans act *as if* they were unboundedly rational. By contrast, optimisation under constraints acknowledges the importance of search as an external process, and thus looks to implement rules, which allow the agent to determine when enough information for the decision problem has been acquired. For example, an agent could implement a rule that translates to “stop search when costs outweigh benefits” (ibid.). This appears at first glance to acknowledge the sorts of ecological constraints that unbounded rationality overlooks, but as Gigerenzer and Todd note, these types of models can still be incredibly computationally demanding. For example, imagine you are considering possible options in the thesis writing example (see introduction), and have written down two possible options. Before proceeding to write down a third, you will have to calculate whether the benefits of continuing search will outweigh the possible costs, and this latter step requires consideration of all the possible options available to you in order to estimate their utilities and probabilities. This step would need to be repeated each time another option is considered, and so as Gigerenzer and Todd state, “constrained optimisation invites unbounded rationality to sneak in through the back door.” (ibid.)

It is worth reiterating the point regarding application again. These models are not held up as examples that capture the actual mechanisms that underlie human decision-making, and almost all will undoubtedly appreciate their limitations as descriptive models, thus restricting their application to the status of ideal models. There is nothing inherently wrong with the pursuit of ideal norms, unless it is used as an unreasonable measure of human (and non-human) intelligence.

Framing and Registration

So far, we have seen how the classical cognitivist theory of problem solving begins by assuming one of the hardest aspects is already dealt with, i.e. the framing of the problem. However, framing is something that real-world agents undertake prior to search, and determines what states or processes are salient to the task. In real-world situations, framing brings a host of biasing preconceptions, specific to the agent about what is salient to the problem, and thus how the problem space will be represented. For example, consider how an expert mathematician may be more adept at recognising the abstract structure inherent in a problem, or how a builder approaches a construction task. The manner in which they frame the problem to be solved will undoubtedly be different to the manner in which non-experts approach the same tasks. However, the construction of an abstract task environment blurs this distinction, and potentially closes off a fruitful investigation into the importance of biasing inputs (e.g. affective signals) and prior learning.

These problems have long been identified as a class of related problems often brought together in discussion of the notorious *frame problem*. Originating in the fields of robotics and artificial intelligence, the frame problem is concerned with the question of how it is possible for a machine to know which of a potentially infinite number of possible actions is relevant at any particular time, without running an infinitely long checking procedure that consults them all. It is obvious that humans (and many non-human animals) have in some regards solved the frame problem, though unfortunately our pragmatic ability to do so on a daily basis does not translate into an understanding of how in fact we achieve this feat. This is a well-known problem, and is neatly described and discussed in (Dennett, 1984). In the next chapter we will begin to look at a possible solution to it by casting it as a problem of deciding between multiple action opportunities. By doing so, we will also see a more natural solution to the problem of registration.

Recall that registration is the problem of how to reinterpret the results of the problem solving task and connect the solution back to the physical world. Note that this problem is only an issue for accounts that separate the aforementioned processes from real-world interactions by making the agent's interactions depend on an indirect mental representation that encodes knowledge in an abstract format. This issue emerged in chapter 1 when we considered the symbol grounding problem, but it was argued there that one of the motivations for defending an embodied account was in order to directly ground the content of mental representations in body-environment interactions (at least for those accounts which make explanatory use of them).

Environmental Resources and Simplifying Behaviours

Finally, recall that one of the purposes of constructing an abstract task environment was to separate relevant task behaviours from irrelevant ones (e.g. the movement of a chess piece versus scratching your head when playing chess). Attempts to delineate something like task-relevancy at the outset are often challenged by some embodied theories who emphasise the importance of acknowledging the whole situation that the agent is embedded in when interpreting observed behaviours (Robbins and Aydede, 2009). This is because the situation and local resources available can alter how the agent will frame the problem. The setting and local resources activate what Kirsh (2009) calls an 'interpretive framework', which is a way of conceptualising the task that primes agents to approach their environment in activity-specific ways, biasing what they see as problematic and what they see as functional. This draws our attention to the famous notion of bounded rationality and Simon's analogy of the

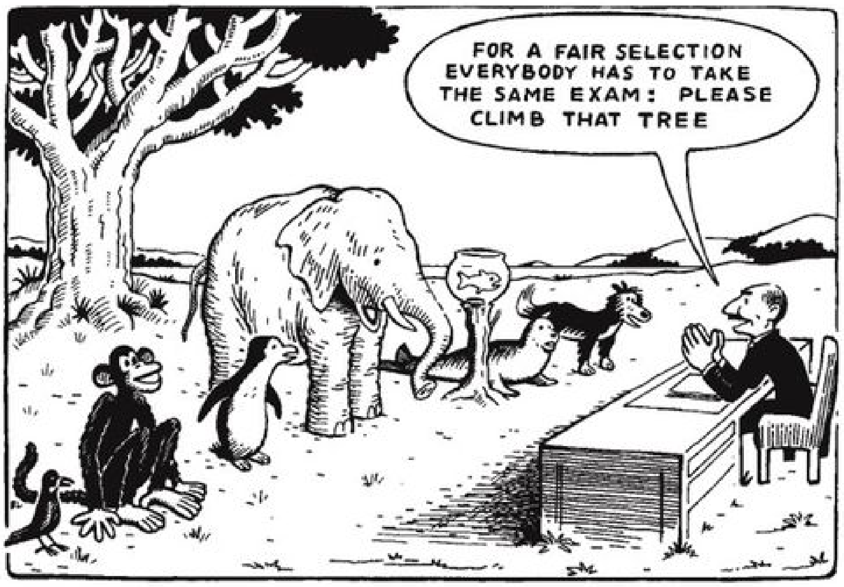

FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE

Figure 3.4: Our Educational System

Image reprinted from:

`https://marquetteeducator.wordpress.com/tag/micah-russell/` [Accessed: 29/08/16]

scissors in which, "rational behaviour is shaped by a [pair of] scissors whose blades are the structure of task environments and the computational capabilities of the actor." (Simon, 1990, p. 7)

Simon's proposal was in effect that without paying due attention to (a) the real task environment, shaped by the physical structures in an agent's world, and (b) the capabilities afforded by the agent's physiology, we would be unable to get a handle on whether some action was in fact rational in some bounded sense. Figure 3.4 demonstrates, in an admittedly tongue in cheek manner, how a failure to acknowledge these two sides of the scissors in cases of cross-species comparison can result in inappropriate ascriptions of (ir)rationality.

Simon's initial proposal of bounded rationality as satisficing, spurred a large research literature that aimed to uncover alternative mechanisms behind decision-making behaviours—most notably the work of simple heuristics proposed by Gigerenzer and Todd (1999). Although this is an interesting literature in its own right, many in the embodied cognition would argue that it is important to begin by looking at the evolutionary environment, in order to uncover opportunities for so called *epistemic actions*: interactions with the environment that simplify the task and offload some of the cognitive demands. By uncovering these cases first, we can determine where and when heuristics (or alternatively more knowledge-rich structures) are necessary.

As a way of demonstrating the efficacy of this strategy, Wilson and Golonka (2013) use the famous case of the Outfielders Problem as an illustration.[3] Approaching this from a perspective that aims to uncover the abstract structure of the problem would begin by describing how a baseball in flight follows a parabolic trajectory, affected by numerous variables (e.g. the angle of the ball as it is struck by the bat, speed and direction of the wind etc.). How would the traditional cognitivist describe the problem for the outfielder?

On this picture, we would begin with perception gathering information about the necessary variables such as initial direction, velocity, and angle, as well as other relevant local factors such as wind speed. These variables can then serve as inputs (representations) into some inner simulation of the actual world. Once computed, the task is then to predict where one needs to run to, and how fast, in order to successfully catch the ball. One of the problems with this picture is that at the usual distances involved, the optical projection of the baseball is tiny, and usually moving quite fast. If we consider that in ecologically-valid situations, cognition is time-pressured, the cognitivist picture begins to seem even more far-fetched, and this

[3] For those unfamiliar with the Outfielder Problem, the example refers to the task faced by a baseball outfielder who has to figure out where to move to in order to successfully catch the ball that is hit by the batter (the flyball).

is before we factor in the time that is required for the outfielder to actually run to the predicted spot, and on the basis of highly uncertain perceptual information.

In contrast, the embodied view can appeal to a wider range of resources. Not only does the embodied view have the brain, the body and the environment at its disposal, but perhaps even more importantly, it has the relations between these things as well (e.g. how our bodies interact with the environment). Wilson and Golonka (ibid., p. 3) propose that a task analysis for the observing cognitive scientist should begin with an exhaustive list of resources available that could contribute to solving the task, and importantly should be approached from the perspective of the subject rather than the observer, beginning with perception and action, and postulating more complex cognitive resources only once the capabilities of the other resources have been exhausted. What does such a task analysis look like in the case of the Outfielder Problem?

To identify the resources available to the outfielder, we first need to understand the nature of the flyball event as a process that unfolds over time. This event produces kinematic information (i.e. information about the objects motion independent of any underlying forces), which is available to an observer. If the observer were simply to remain passive, trying to determine where the ball would land on the basis of this information would be too computationally demanding. However, the embodied view can appeal to further resources such as the body and the environment and the interactions between them. Of particular interest here is the close coupling between action and perception, and how certain movements of the outfielder change the perceptual input of the ball in parabolic motion.

Fink, Foo, and Warren (2009) discuss two strategies for solving the Outfielder problem that are known as Optical Acceleration Cancellation, and Linear Optical Trajectory (LOT). OAC involves the outfielder running in a particular alignment with the ball so as to cancel the vertical acceleration of its optical projection in the visual field, which results in the ball appearing to move with a constant velocity. LOT requires the outfielder to move laterally, so as to keep the apparent trajectory of the ball linear and appear to trace a straight line. Neither option requires the outfielder to predict in advance where the ball will land. Instead in both cases, the very movement of the outfielder is harnessed to bolster the otherwise limited kinematic information. By using the wider resources of a body-environment relation, the outfielder can solve the task in a far less computationally demanding manner, simply by running in a particular way.

We will see another example that exploits this ecological approach in the final chapter. For the time being it is sufficient to highlight (a) the limitations of the cognitivist approach, and (b) the explanatory scope of accounts that adopt exclusively knowledge-rich or knowledge-lean accounts.[4] Although the outfielder problem eschews knowledge-rich explanations successfully, few would argue that a complete account of human cognition can be accommodated by similarly knowledge-lean explanations—many of which are likely to be domain-specific (e.g. heuristics). It may be possible to argue that 'experience' could be recast as familiarity with relevant 'search heuristics', but much of an agent's success surely has to do with an understanding of how to translate domain general knowledge to new tasks in the first place. These concerns will be returned to in chapters 5 and 6.

[4] By knowledge-rich, I am referring simply to the idea that certain problems require the positing of richly reconstructive mental representations that aim to accurately reflect the structure of the world, and are subsequently used as the basis of inner cognitive processing that is detached (at least during key stages of processing) from sensorimotor regions. By contrast, knowledge-lean solutions would emphasise close coupling with the environment, and the problem-solving may make use of sensorimotor regions in a constitutive manner. This does not need to imply an entirely anti-representational stance, although some situations may lend themselves to such an explanation. Ultimately the distinction admits of degrees, and does not map onto the representational/ anti-representational divide cleanly.

3.2 Neuroeconomics

Some may worry that the criticisms in the sections above were directed at a strawman, or at least at a traditional account that is no longer seriously defended. In this section, we will briefly outline some more recent empirical work from the field of *neuroeconomics*, which demonstrates the continued adherence by some working in the cognitive sciences to certain cognitivist assumptions.

As a movement in its own right, neuroeconomics has emerged relatively recently, following the development of techniques such as functional magnetic resonance imaging (fMRI) in the early 1990s. However, as a combination of two pre-existing approaches (neuroscience and economics), its history extends further back into the origins of these pre-existing disciplines. Although exploring the history of neuroscience would be fruitful for understanding the motivation behind the merging of these disciplines, it is perhaps more fruitful (for reasons that will be made clear) to explore a brief history of economic theory.

As many historians of economics would acknowledge, the birth of the classical period of economic theory began with the publication of Adam Smith's *The Wealth of Nations* in 1776. In addition to the many insights into the causes behind a nation's prosperity, Smith explored a number of phenomena that he believed were integral to understanding choice behaviour—in effect providing *psychological* insights that were first explored in his earlier work *The Theory of Moral Sentiments* (see Ashraf, Camerer, and Loewenstein, 2005, for a discussion). This trend was continued by later economists, often unperturbed by an inability to experimentally test these psychological models.

In the 1930s, economists (e.g. Samuelson, 1938) attempted to develop more rigorous mathematical models that explained choice behaviour by appealing to a number of primitive assumptions about an agent's preferences. Although this was a departure from the earlier economics of Smith, the approach had a precursor in the form of Daniel Bernoulli's observations of people's behaviour in games of chance. Bernoulli (1738) noted that people's behaviour regularly failed to maximise expected monetary value. This was famously illustrated by the well-known St Petersburg Paradox, which today enjoys the status of a well-confirmed empirical fact (Okasha, 2015). In response, Bernoulli suggested that people are instead maximising expected *utility*, the function of which is the logarithm of monetary value. The importance of these observations should not be understated, for as Okasha (ibid.) notes, Bernoulli's suggestion, despite being influential, failed to provide any explanation of *why* an agent should maximise utility.

Once Savage (1954) and von Neumann and Morgenstern (1944) had developed Bernoulli's initial argument—demonstrating how an agent's subjective utility function can be determined on the basis of observable preference relations between lotteries—Bernoulli's unexplained assumption was given little attention by economists. Following the axiomatisation of expected utility theory, it was possible to demonstrate, by means of a representation theorem, that any agent whose preferences satisfied reasonable axioms (i.e. transitivity, continuity and independence) would behave *as if* they were maximising some expected utility function (see Glimcher and Fehr, 2014a; Okasha, 2015, for introductions). The question of why agents acted like this appeared to be of little concern once a rigorous mathematical structure had been provided. This effectively divorced economics from psychology, as economists needed only to concern themselves with observable, and easily quantifiable, choice behaviour. Questions regarding the psychological processes underlying this behaviour were extraneous, so long as assumptions regarding the consistency of certain axioms were maintained.

These developments in neoclassical economics proved to be incredibly popular, in spite of examples such as the Allais paradox (Allais, 1953) that challenged the plausibility of axioms such as independence, and were subsequently backed up by empirical observations (Tversky and Kahneman, 1986). However, the popularity came at a price. As more and more counter-examples accumulated (e.g. Ellsberg's Paradox (Ellsberg, 1961)), economists were compelled to weaken the normative force of their models by defending weaker axioms, or by setting restricted boundary conditions on the descriptive validity of the models. This latter move was the choice made by Simon with his proposal of *bounded rationality* that we discussed in the previous section. Eventually, a group of psychologists (most notably Daniel Kahneman and Amos Tversky) who were studying the foundations of choice behaviour, presented a range of phenomena, which diverged so drastically from the models of expected utility theory, that the descriptive validity of the expected utility approach was radically undermined (Tversky and Kahneman, 1981).

A particularly noteworthy effect that was observed by Tversky and Kahneman (1981) is the *framing effect*, which shows the influence of context on value-based choice. For example, subjects are observed to prefer riskier choices if they are presented in terms of a potential loss rather than a potential gain. This observation was important for a number of reasons, but is particular noteworthy for the present purposes because of a connection with later material regarding how neural systems encode value (section 3.2.2). As these experiments tentatively suggest, agents may not necessarily represent value in an objective or stable manner (e.g. by means of a utility function) but may use some other method (e.g. heuristic). Alternatively, if researchers wish to maintain the psychological reality of utility functions, given that an agent's utility function must have certain properties (see Okasha, 2015), there must also be additional corresponding mechanisms that are postulated in order to account for the observed divergences from rational behaviour.

Bringing us to the present, these alternatives provide researchers with competing hypotheses to explore, and a number of ways of attempting to account for apparent violations. On the one hand, behavioural economists[5] often argue that good decisions aim to maximise expected utility over the *short-term* or the *long-term* (Glimcher and Fehr, 2014b). Violations can be put down to inappropriate framings, or mistaken assumptions regarding the task environment. Alternatively, behavioural ecologists argue that organisms are instead aiming to maximise their fitness, rather than some abstract utility function, and that the basic goal for any biological agent is primarily survival and reproduction (Stephens, 2008). These definitions may align in some cases, but sometimes they will diverge, and these latter cases can be incredibly fruitful for gaining an understanding of the mechanisms behind decision-making (Summerfield and Tsetsos, 2015). To help decide between these competing theories, neuroeconomists propose that neuroscientific research should be undertaken in order to gain a more tangible grasp on the inner mechanisms that underlie our decision-making capacities.

The literature surrounding neuroeconomics is vast and continues to grow rapidly, which means we must unfortunately restrict ourselves to a small number of cases. Therefore, it is important to highlight at the outset that any criticisms should be understood with a sufficiently narrow scope, rather than attempting to undermine the general movement, or cast it in overly monolithic terms. In spite of this caveat, it is often the case that neuroeconomics research aligns with the aforementioned strand of economic theory that treats human choice behaviour as predominantly aiming at the normative prescription to maximise expected utility (Glimcher and Fehr, 2014b). Many defenders of neuroeconomics are happy to acknowledge that this commitment to using economic methods for understanding neural behaviour means

[5] Glimcher and Fehr (2014a, p.xix) define behavioural economics as a discipline which seeks to propose "models of limits on rational calculation, willpower, and self-interest, and seeks to codify those limits formally and explore their empirical implications using both mathematical theory, experimental data and analysis of field data."

working within well-known constraints of expected utility theory (e.g. measuring utilities on an ordinal rather than a cardinal scale):

> “To those coming from the natural sciences, it can come as a shock to discover that economists shy away from assigning cardinal meaning to numerical utilities. Economists look askance at those who would assign any but the most qualitative of meanings to these utility numbers. A higher number means no more and no less than that an option is preferred. How much higher one number is than another is seen as essentially meaningless, largely thanks to Pareto. This is an absolutely central feature of economic thought that must be understood by anyone who interacts with economists.” (Caplin and Glimcher, 2014, p. 7)

However, one of the strengths of the neuroeconomic approach can be traced back to the methodological approach of Samuelson (Samuelson, 1938), who in effect argued that rather than merely assuming as if subjects maximise utility with their choice behaviour, economists should figure out how to test the hypothesis that certain choices are consistent with the approach. Just as Samuelson’s approach was influential in axiomatising expected utility theory, so too neuroeconomists hope that the same methodology can be applied to cases in neuroscience. One instance where this is particularly notable is in the case of the *Reward Prediction Error Hypothesis*.

3.2.1 Reward Prediction Error Hypothesis

The reward prediction error hypothesis (RPE) explores the role that the neurotransmitter dopamine plays in encoding a teaching signal that guides reward-based reinforcement learning (particularly in the case of midbrain dopaminergic neurons) (Glimcher, 2011b). The general idea is that these dopaminergic neurons signal a prediction error that can be used to update predictions that correspond to expectations regarding certain options that an agent desires. The relevance to neuroeconomics is that these predictions are hypothesised to correspond to something like the lotteries of subjective expected utility theory, while the errors correspond to the discrepancy between the anticipated lottery and the actual prize (Caplin and Glimcher, 2014). This is because, if there is a reward, there must be something the agent desires (i.e. prizes); if there is a prediction, there must be subjective beliefs concerning expected prizes (i.e. lotteries), and if there is an error, it must be possible that the actual prize does not align with the agent's belief (i.e. outcomes). Neuroeconomists who favour the RPE hypothesis propose that the amount of dopamine that is released would be proportional to the prediction error, and could also account for the subjective value that an agent assigns to the expected lotteries (Glimcher and Fehr, 2014b). This in turn affects the probability that a corresponding action will be chosen. Before assessing the empirical validity of this scheme, it is important to highlight its close links to economic theory.

Caplin and Dean (2008) have argued that any model that supports the RPE hypothesis can be tested by developing a number of axioms (similar to the axioms of expected utility theory), and then performing experiments that place subjects in situations that mimic decisions under uncertainty. These axioms are summarised by Caplin and Glimcher (2014) and connected with three elements of the RPE hypothesis:

(Reward) Coherent prize ordering: Holding the probabilities of rewards fixed and varying their magnitude in an order-preserving manner (e.g. more money or more juice (for monkeys)) should not result in different ordering observed at the level of neural activity.

(Prediction) Coherent lottery ordering: Fixing rewards and varying the probabilities of obtaining them should result in coherent orderings across trials with different prizes but similar probabilities.

(Error) No-Surprise Equivalence: When prizes are perfectly anticipated (i.e. no surprise), the dopaminergic response should be identical across all predicted outcomes.

All of these axioms refer to predictions regarding the correlated dopaminergic response, and have precise mathematical definitions given by Caplin and Dean (2008). Subsequent experiments using fMRI have found that these axioms are maintained in some regions of the brain (e.g. ventral striatum), violated in some (e.g. insula) and are ambiguous in others (e.g. prefrontal cortex) (cf. Glimcher, 2011b; Glimcher and Fehr, 2014b, for a review of the studies). Regardless of any subsequent criticism, it should be noted that this approach (and its corresponding methodology) is an incredible achievement, and upholds the scientific ideal of formulating rigorous mathematical models that make empirically testable predictions that are potentially falsifiable. Unfortunately, at present the RPE hypothesis itself rests on a number of unstable conceptual and empirical foundations.

The first challenge is that the RPE hypothesis is controversial in neuroscience, with many pointing to alternative influencing factors that are strongly implicated as influencing dopaminergic neurons, and which are only weakly related to reward prediction error (e.g. reward-neutral properties such as surprisal or salience (Knutson and Peterson, 2005), discovery of new actions (Redgrave and Gurney, 2006) and modulatory roles in precision-weighting (Friston et al., 2014)). Also, within these roles, dopamine release can function differently depending on the timing. If it is released *following* some salient behaviour, it can play the role of updating the subjective probability of future choices, but if it is released *prior* to behaviour it appears to act as the gating mechanism that enables both cognitive and behavioural mechanisms (e.g. updating plans in working memory or enabling motor control) (Landreth and Bickle, 2008, p. 423).

The second challenge is that the RPE hypothesis makes strong functional assumptions regarding the format of the value representations, which must have discoverable neural correlates if the hypothesis is to be vindicated as anything more than a behavioural theory. For example, whereas the sorts of natural stimuli that are used as prizes in the experiments (e.g. juice or money) can be easily quantified such that more is obviously better, the history of economics demonstrates why it is unwise to translate this into a similar quantifiable measure in the case of subjective beliefs. The axiomatised RPE hypothesis sidesteps this issue to some extent by beginning with weaker assumptions. However, it is still limited to the claim that the dopaminergic response correlates with objective, quantifiable features in the external world, rather than something that may be more salient to the organism and is measurable on an entirely different scale (e.g. affective significance), which perhaps results in a problem of underdetermination. This issue can be clearly seen by turning to another example.

3.2.2 Common Currency and The Futile Search for True Utility

In the previous section we noted that the responses of dopaminergic neurons are implicated in a number of cases, some of which only weakly correspond to reward. Although this undermines the RPE hypothesis, some may worry that it also undermines the search for an unambiguous neural signal that encodes subjective value, which seems to be required to vindicate neuroeconomics' search for the neural basis of utility. However, we may wonder whether this search is even well-defined in the first place.

Recall that in traditional decision theory, utility is taken to be a mathematical representation that is inferred from simple choices that meet certain consistency axioms. This is quite different from identifying utility with either (a) an experienced hedonic value or pleasurable feeling, or (b) an agent's expected reward. Fumagalli (2013) calls these latter two 'experienced utility' and 'neural utility' respectively, and to contrast them with the decision theoretic notion, groups them together under the label 'true utility'. We will here focus on the notion of neural utility, given that it is the notion that is advocated by many neuroeconomists, who often claim that it should replace traditional notions of utility as a mathematical representation (e.g. Camerer, Loewenstein, and Prelec, 2005). They argue that the "“as if” approach made good sense as long as the brain remained substantially a black box" (ibid., p. 10). However, developments in neuroscience mean the brain is now ripe to be explored and understood by incorporating many of the constructs of economic theory to modelling the behaviour of interacting neurons and neural populations.

As Fumagalli (2013, p. 329) defines it, *neural utility* relates to patterns of neural activity in certain regions of the brain, where "desirability is realized as a concrete object, a neural signal in the human brain, rather than as a purely theoretical construction". He cites a number of advocates of this idea, whose views reflect differing degrees of support. For example, Park and Zak (2007, p. 50) claim that "the utility calculations that people were assumed to do really happen in the brain". Whereas Glimcher (2011a, pp. 133-134) supports a slightly more nuanced view that states when a subjects' behaviour accords with the predictions of expected utility theory, it is "because they neurally represent something having the properties of utility—a neural activation that encodes the desirability of an outcome in a continuous monotonic fashion".

In support of the latter claim, Levy and Glimcher (2012) present a meta-analysis of neuroimaging studies in humans. These studies appear to demonstrate how the neural encoding of subjective values, in a number of brain areas (most notably ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC)), reflects a common value scale for comparison of options that is required by expected utility theory. This is the so-called *common currency hypothesis*, and is motivated by a belief that in order to make rational decisions agents must evaluate the costs and benefits of available options using an independent 'currency' that is able to compare otherwise incommensurable options. Montague and Berns, who have also worked on the neuroeconomic notion of a common currency, define the term as follows:

> A currency is an *abstract* way to represent the value of a good or service. For our purposes in this paper, it possesses an important property: it provides a common scale to value fundamentally incommensurable stimuli and behavioral acts. For example, suppose we want to understand the relative value of 17 coconuts and 41 sips of water. There is no natural way to combine coconuts and sips of water; however, each can be converted to their valuation in some currency, and the values can be combined in any number of ways. This kind of abstraction is so common in our everyday world that its biological substrates go virtually unnoticed. Without internal currencies in the nervous system, a creature would be unable to assess the relative value of different events like drinking water, smelling food, scanning for predators, sitting quietly in the sun, and so forth. To decide on an appropriate behavior, the nervous system must estimate the value of each of these potential actions, convert it to a common scale, and use this scale to determine a course of action. This idea of a common scale can also be used to value both predictors and rewards. (Montague and Berns, 2002, p.276, emphasis added)

Neuroeconomists have set themselves a goal of determining which brain mechanisms are responsible for the evaluation and deliberation of this common currency.

We can critique this goal by way of several questions:

1. Does the brain encode a single *abstract* currency?
2. If so, where in the brain does this happen?
3. If not, how many currencies does the brain compute?
4. Are there significant differences between these currencies?

With regards to the first question, Levy and Glimcher (2012) may appear as if they are arguing strongly in favour of an affirmative response, while also providing a response to the second that supports Fumagalli's (2013) portrayal of neuroeconomics' search for true utility. For example, they claim on the basis of the aforementioned meta-analysis that:

> "Quite a few studies have now demonstrated that a subregion of the vmPFC/OFC [...] represent subject-specific reward value in a common neural currency, the expected subjective value of Neuroeconomic theory." (ibid., p. 1035)

If this were the case, it would seem to support a strict cognitivist reading of the common currency hypothesis by which the vmPFC/OFC encode an abstract representation of subject-specific value. However, it wouldn't be an *entirely* fair characterisation, and would also ignore a whole host of other evidence that seems to weaken this claim substantially. First of all, as Levy and Glimcher (ibid.) themselves note:

> "[...] there is no evidence to support the claim that the neural common currency of value arises *only* in this subregion of the vmPFC/OFC. Any common currency observed in the brain must reflect the activation of

> multiple brain areas. [...] Indeed, the evidence reviewed here suggests that portions of the striatum and perhaps the insula also participate in this process."

However, whereas they argue that these other regions are active, and contribute to a distributed task of encoding a single common currency, there is still the assumption that this distributed activity (representing an agent's deliberative process) is somehow integrated into an abstract representation that allows for commitment to take place in some central executive region such as vmPFC/OFC. Others have argued in favour of a similar approach. For example, Platt and Padoa-Schioppa (2009) focus on not only the OFC, but also the lateral intraparietal area (LIP) and the posterior cingulate cortex (CGp), and argue that value representations differ significantly across these areas. In the case of the LIP (a region commonly associated with eye movement), this region has been heavily implicated in decision-making tasks, but is often described as encoding a more reward-neutral signal (i.e neither immediately rewarding or aversive stimuli) such as behavioural salience (Cisek and Kalaska, 2010; Freedman and Assad, 2011; Landreth and Bickle, 2008; Treue, 2003). Sugrue, Corrado, and Newsome (2005, p. 367) describe this point clearly when they state that activity in the LIP encodes "information that is pertinent to the selection of future shifts in gaze or attention." With regards to the CGp, Platt and Padoa-Schioppa (2009) claim that studies of this area—typically associated with learning, and strongly connected with parietal cortex, an area implicated in planned movement—support a number of distinct roles in decision-making, including risk-evaluation, motivational significance and temporal discounting. Although they summarise this under the label of 'behaviourally salient value', the key point is that this region undoubtedly plays myriad roles in what appears to be a distributed set of mechanisms underlying decision-making.

Why would the brain encode so many disparate currencies? One answer is that although the environment affords multiple simultaneous action opportunities, agents also have individual needs (e.g. thirst, hunger or tiredness). For each of these needs, certain outcomes may be more effective at satisfying the current desire of the agent (e.g. water vs. fruit for quenching thirst). The brain may use multiple currencies to rank outcomes and actions as a function of the initial need that motivated the decision task, calling upon different regions of the brain as necessary. However, it seems to leave unanswered how the brain deals with options that are incommensurable in other respects, and possibly lead to radically different (but equally valid) solutions to the same problem (e.g. what to do in the case of the thesis writing—also see section 3.1.1). This seems to necessitate a return to Levy and Glimcher's claim that regions of the brain such as the OFC may in fact act as some sort of executive region, co-ordinating the other decision-making systems; an executive which perhaps fails from time to time to effectively integrate the competing information arising from these distinct systems, leading to the sorts of irrational decisions that plague rational choice theory. As they rightfully ask:

> "[W]hat happens in the brain when we need to choose between a large amount of water and a single apple? [...] What we need to do is to take into consideration many different attributes of each option (like color, size, taste, health benefits, our metabolic state, etc.), assess the value of each of the attributes, and combine all of these attributes into one coherent value representation that allows comparison with any other possible option. What we need, at least in principle, is a single common currency of valuation for comparing options of many different kinds." (Levy and Glimcher, 2012, p. 1027)

Whist we are sympathetic to the idea that more frontal regions of the brain play some sort of co-ordinating role (see chapters 5 and 6), we resist the idea that the correct representational description of this region is the one described by Levy and Glimcher, or indeed in any cognitivist style description that posits a disembodied, abstract decision-making system that is ultimately responsible for integrating activity from other ancillary decision-making systems. As we will see in the next section, we do not believe it to be helpful to assume that commitment (in the decision-theoretic sense) occurs only once all relevant sensory information has been integrated, as it has the unwanted effect of separating decision-making from sensorimotor regions.

It would be easy to pass this off by claiming that most of the confusion arises in part due to multiple, diverging uses of the term 'decision' in the cognitive sciences. For example, there is a distinction in the cognitive sciences between *economic decisions* on the one hand, and *perceptual decisions* on the other. The former involve choosing among alternative, discrete options associated with different rewards, while the latter require subjects to "choose" between competing percepts on the basis of ambiguous or noisy sensory evidence, in order to categorise objects in the world, and perhaps choose some relevant associated actions (Freedman and Assad, 2011). However, while this undoubtedly accounts for some of the variety in the decision-making literature, it does not vindicate the common currency hypothesis entirely.

For a start, although we have said nothing about the notion of *experienced utility* that Fumagalli (2013) introduces, it should be clear that the notion of value that has been proposed by neuroeconomists is too narrowly defined to be able to successfully accommodate all of the interesting phenomenological and conceptual differences between these notions. Therefore, as Fumagalli rightfully argues, we should resist arguments that attempt to collapse the two into a single unitary concept such as neural utility. However, what Fumagalli does not consider is whether we can account for the findings of neuroeconomics within an alternative framework. One that is perhaps able to acknowledge these distinctions without appealing to detached abstract representations, and at the same time retain the admirable goal of explaining how the brain is able to deal with the competing sources of information arising from multiple decision-making systems. The remainder of this thesis turns to consider this very possibility.

3.3 Embodied Decisions

> "[...] studies on the neural mechanisms of decision making have repeatedly shown that correlates of decision processes are distributed throughout the brain, notably including cortical and subcortical regions that are strongly implicated in the sensorimotor control of movement. Neural correlates of decision variables appear to be expressed by the same neurons that encode the attributes of the potential motor responses used to report the decision, which reside within sensorimotor circuits that guide the online execution of movements." (Cisek and Kalaska, 2010, p. 270)

The cognitivist view of problem solving and decision-making leads to a tendency to think of sensorimotor control in terms of the transformation of input representations into output representations through a series of well-demarcated, encapsulated processing stages. It also often leads to the assumption that key decision variables are encoded in some central executive region, as an abstract value (Levy and Glimcher, 2012; Padoa-Schioppa, 2011). Deliberation and commitment thus proceed in sequence, and importantly are separate from sensorimotor regions. As an example of the standard account, Cisek (2012) cites the goods-based model (Padoa-Schioppa, 2011), which suggests choice behaviour is governed by integrating all relevant factors (e.g. expected gains, possible risks etc.) into a single subjective value. This value, which is associated with some corresponding action, is compared with the set of alternative options, and the one with the highest expected value is selected. This commitment occurs prior to movement onset.

Cisek and Pastor-Bernier (2014) argue that this picture is hard to reconcile with a growing body of neurophysiological data. They discuss three instances of this conflict, which taken together represent a considerable challenge to the traditional account. Firstly, the traditional account predicts that motor behaviour only begins once an option has been selected. However, this is challenged by multiple studies (reviewed by (Cisek, 2012)) that demonstrate how neurons in motor regions represent multiple potential targets and actions prior to the agent selecting between them. Secondly, additional studies (Cos, Bélanger, and Cisek, 2011) show that when humans were required to freely choose between two reaching actions with equivalent reward values, the subjects unsurprisingly favoured the one with a lower associated biomechanical cost. However, the studies importantly ensured that the difference in biomechanical cost only exists during the later stages of the movement, therefore requiring that the brain represents information about future biomechanical costs before deciding between them. Finally, the traditional account fails to account for the wide spread existence of decision-related modulatory effects in sensorimotor regions (see below). [6] To accommodate this otherwise anomalous data they propose a notion of embodied decisions. Embodied decisions have a number of properties that are quite different to the kinds of decisions modelled by traditional decision theory.

3.3.1 Decision-Making as a Distributed Consensus

Cisek proposes the affordance competition hypothesis (ACH) as a model that aims to explain both the cognitive and neural processes implicated in decision-making (Cisek, 2007; Cisek and Kalaska, 2010). According to the ACH, decisions emerge from a distributed, probabilistic competition between multiple representations of possible actions in sensorimotor circuits. To expound this view, a number of components

[6] As we will see, this is something that PP is well-equipped to handle.

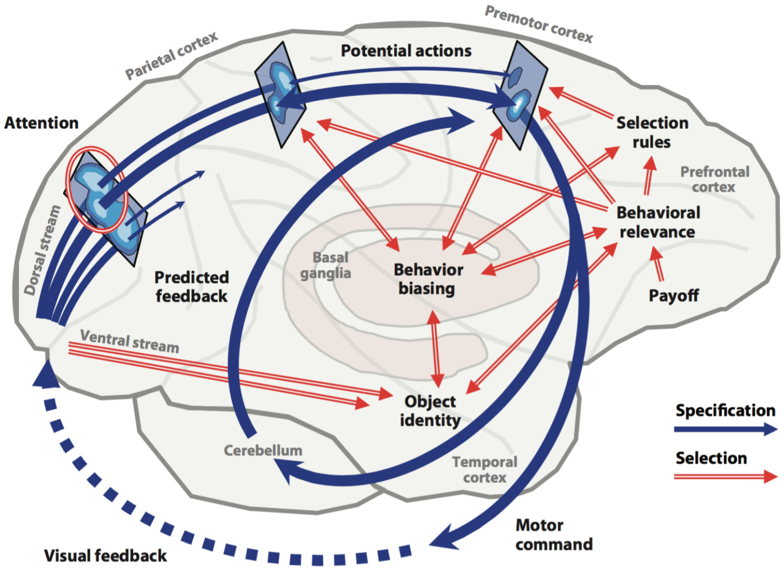


Figure 3.5: A sketch of the Affordance Competition Hypothesis. Reprinted from (Cisek and Kalaska, 2010, p. 278).

require clarification.

Cisek's focus on the distributed manner of decision-making stands in contrast to the earlier cognitivist framework, and also to other models that propose that decision-making occurs downstream of the integration of multiple sources of information, which yields a common representation of abstract value (Padoa-Schioppa, 2011). Instead, according to the ACH (see Figure 3.5), the sensorimotor system is continuously processing sensory information in order to specify the parameters of potential actions, which compete for control of behaviour as they progress through a cortical hierarchy, while at the same time other regions of the brain provide biasing inputs in order to select the best action (Cisek and Kalaska, 2010). These processes of specification and selection occur simultaneously and continuously, and are not localisable to a specific region. Rather, the competition occurs by way of mutual inhibition of neural representations, which specify the parameters of potential actions, until one suppresses the others and a distributed consensus emerges. At this point, movement onset commences, and Thura and Cisek (2014) propose that the point when the competition between actions is resolved within the motor system constitutes the voluntary commitment to an action choice. Note here that the commitment is to an *action choice*, rather than a more abstract state of the world. This will be important in later chapters.

Integral to this process is the role of continuously biasing influences (i.e. rule-based inputs from prefrontal regions, reward predictions from basal ganglia, and a range of further biasing variables from sub-cortical regions). Each of these biasing inputs contribute their votes to the selection process. As the authors state:

> “[...] the decision is not determined by any single central executive, but simply depends upon which regions are the first to commit to a given action strongly enough to pull the rest of the system into a ‘distributed consensus’.” (Cisek and Pastor-Bernier, 2014, p. 4)

Again, this idea stands in stark contrast to the cognitivist picture, where the perceptual system merely processes information in order to construct a perceptual representation, which provides the evidence about the environment needed to make decisions. Rather, here we have the beginnings of an account that explains how the relevant options of a decision problem are being selected in parallel with the specification of sensorimotor information:

> “[...] although traditional psychological theories assume that selection (decision making) occurs before specification (movement planning), we consider the possibility that, at least during natural interactive behavior, these processes operate simultaneously and in an integrated manner.” (Cisek and Kalaska, 2010, p. 277)

One of the specific claims made by Cisek and Pastor-Bernier is that as part of the competitive process, the brain is simultaneously specifying and selecting among representations of multiple action opportunities or affordances, which compete within the sensorimotor system itself (Cisek and Pastor-Bernier, 2014). These representations serve as indications of the possible actions available in the agent's environment, rather than as objective, organism-independent properties of the world.

For example, Cisek and Kalaska discuss recordings taken from the dorsal premotor cortex (PMd) in monkeys during a reaching task (Cisek and Kalaska, 2010). In the experiment, monkeys were presented with two potential reaching actions by way of spatial cues, where one would later be indicated (using a non-spatial cue) as the correct choice. During a memory period, where the spatial cues were removed and the future correct choice was uncertain, recorded activity in the PMd continued to specify both directions simultaneously, suggesting an anticipatory nature for the neural activity. When the information specifying the correct choice was eventually presented, activity relating to the respective action was strengthened, and the unwanted action was suppressed.

Importantly, this process occurs within the same system that is ultimately used to prepare and execute the movement associated with the action representations. Furthermore, Cisek and Kalaska state that the task design allowed for the monkeys to exploit a different (cognitivist) strategy, where the target locations are stored in a more general-purpose working memory buffer, distinct from motor representations, and converted to a motor plan after a decision has been made. However, though conceptually possible, the findings did not seem to support this latter view. Instead, the study seems to point to a need for representations that encode predictive (or anticipatory) action opportunities, rather than abstract representations that specify the state of the world independently from an agent's particular goals and capacities.

The ACH also makes key predictions that can be tested in future experiments.

For example, it predicts that actions that are farther apart from one another will show stronger mutual inhibition than those that are closer together. This is because action representations are specified in terms of spatial parameters (ibid.), which means that a decision between similar actions (with overlapping neural representations) can be encoded using a weighted average. This weighted average could evolve over time, initially tolerating some uncertainty between two future actions, whereas drastically different options could not. A prediction made by the ACH is, therefore, that if one records from neural cells related to a given option, while modulating the desirability of a different option, the gain of that modulation will be strongest when the other option is most dissimilar to the one coded by the recorded cell (Cisek and Pastor-Bernier, 2014, p. 5). This opens up a doorway for so-called "weak" long-range connections in the brain, which Park and Friston (2013) and others have argued may play a fundamental role in the global integration of densely connected sub-regions (see chapter 5).

The ACH thus differs from more traditional approaches to decision-making by eschewing abstract representations that capture knowledge about the world independent of an agent's interactions with it. Instead, it is best seen as a functional mixture of the myriad biasing inputs that contribute to the specification and selection process (to be explored in more detail in the following chapters). It thus lacks a clear commitment to explicit perceptual, cognitive or motor representations, opting instead for a blurring of these boundaries. The role of these action-oriented representations is not to accurately reconstruct an inner description of the world, but rather to coordinate adaptive interaction.

3.3.2 Simultaneous Decisions

Traditional accounts of decision-making have focused on decisions pertaining to options that remain stable over time. This emphasis may have contributed to the

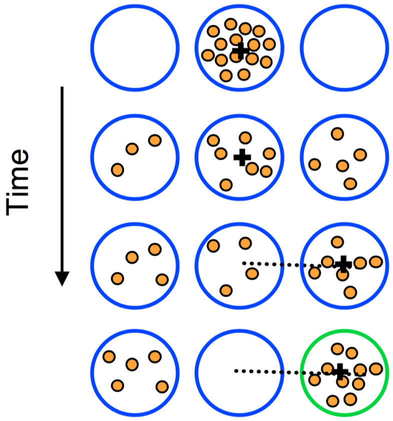


Figure 3.6: Monkeys were required to indicate which of two targets was expected to receive the majority of tokens, and were free to indicate this at any time. Reprinted from (Thura and Cisek, 2014, p. 1402).

postulation of stable, abstract representations in the brain. When deciding between different courses of action in the world, however, sensory information rarely stays fixed, action in the world can open up new possible options, and agents are free to decide on the basis of incomplete information. One of the claims of embodied decisions is that sensorimotor regions not only track the changing state of sensory information in the world, but moreover facilitate efficient action selection by actively contributing to the decision process. Moreover, the sensorimotor system remains receptive to simultaneous action opportunities even once commitment has occurred, in order to keep track of the unfolding consequences of action performance. To further reinforce the claim that sensorimotor regions actively play a role in decision-making, Thura and Cisek (2014) performed an experiment on monkeys, which aimed to replicate this more dynamic approach to decision-making.

Their experiment required a monkey to indicate which of two possible targets was expected to receive a majority of tokens, which moved successively from a central region in 200ms steps. The monkey was trained to indicate their decision by moving a cursor to the respective target, and was free to do so at any point during the trial. In a similar fashion to the previous experiment, neural activity in PMd and also primary motor cortex (M1) was recorded, and approximately 280 ms before the monkey initiates movement, activity in PMd that was tuned to the selected target reached a consistent peak, while M1 activity tuned to the unselected target was simultaneously suppressed. The authors argue that the activity recorded did not support a model of integration of sensory information. Instead they claim that PMd activity tracked the evolving sensory information, but also included a general urgency signal which increased over time urging the monkey to act. In experiments that indicate when the subject is able to respond, there would be no basis for such an urgency signal. However, in ecologically-valid scenarios, opportunities may be lost over time, and thus there will be no *a priori* value for the optimal time to initiate action. Thura and Cisek (ibid.) argue that a growing urgency signal (also biased by modulating inputs) would be preferable in these situations, and could further lead to an optimal (context-dependent) trade-off between speed and accuracy.

In addition to a growing urgency signal, an ability to effectively decide between simultaneously presented action opportunities requires that the agent is able to manage the diverse range of sensory inputs with limited neural resources. To account for this, Cisek and Kalaska review a considerable number of studies on the pervasive effect of attentional modulation, which support the idea that activity in the visual system is strongly influenced by attentional modulation, even in familiar and stable environments (Cisek and Kalaska, 2010). This is usually recorded as an enhancement of activity correlated with the attended regions of space, and a suppression of activity from the unattended regions. For example, studies by Stefan Treue (2001; 2003) show the ubiquitous effects of attentional modulation in primate visual cortex. This attentional modulation results in the enhancement of activity towards behaviourally relevant stimuli, along with a corresponding suppression of those cells tuned to non-attended spatial features. Attending only to those features of the world that are behaviourally salient is likely to be far removed from what is considered rational. However, echoing the sentiments of work in ecological rationality, Treue acknowledges that it is nevertheless “an effective use of limited processing resources.” (Treue, 2003, p. 428)

Despite the attractiveness of appealing to saliency and attention on ecological grounds, without the inclusion of reciprocal communication between affective and sensorimotor regions such an account would remain incomplete. This is because adaptive choice behaviour requires an awareness of the changing demands of both the external and internal environment, in response to the homeostatic demands of the agent—in short what the agent cares about. Although they have pointed to the possible mechanisms involved, at present this is one area that is left underdeveloped by Cisek, Kalaska and Pastor-Bernier. In chapters 4 and 5, we will see how this aspect of embodied decisions can be developed further, by exploring the unique roles of attention and salience within predictive processing, and its emphasis on interoceptive inference.

3.3.3 Dynamic Choice Behaviour

Continuing with the theme of a more dynamic approach to decision-making, Cisek and Pastor-Bernier (2014) claim that the continual processing of noisy or uncertain sensory information after commitment suggests that agents continue to deliberate during the overt performance of a task. This means the agent constantly monitors the overt performance of their actions through sensory feedback (e.g. proprioception). As deliberation is supposed to occur prior to commitment, the existence of this evidence, they argue, requires the revision of some commonly used formal models in decision theory that are unable to account for this post-selection monitoring and alteration.

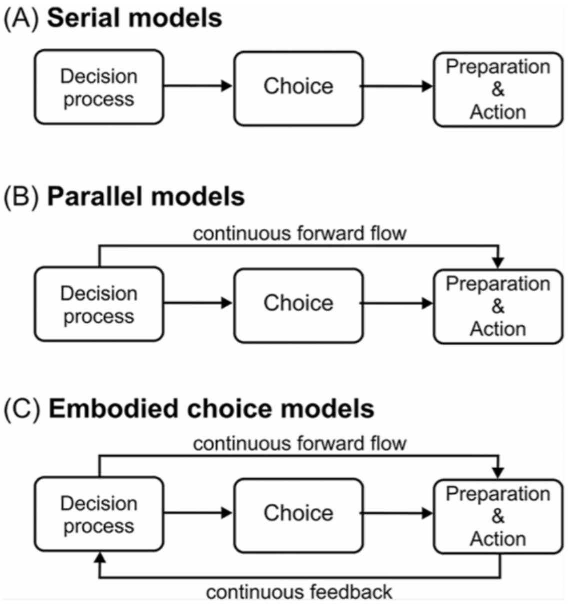


Figure 3.7: A schematic of three models that link decision and action systems. The modularity of the decision process, choice and action is for illustrative purposes only. Reprinted from (Lepora and Pezzulo, 2015, p. 3).

Lepora and Pezzulo (2015) also acknowledge this requirement, and claim that action performance should be considered a proper part of a dynamic model of decision-making; rather than being understood as merely the output of the decision process. As a proof of principle to support this claim, they develop a computational model, which they call the *embodied choice* (EC) model. The most important point of the EC model is the existence of bidirectional influences between action and decisions. Lepora and Pezzulo compare the EC model against two alternative models based on the well-known drift-diffusion model[7] (Ratcliff, 1978).

[7] The drift-diffusion model aims to capture how a subject integrates (noisy) accumulating evidence, for multiple distinct options, in a forced choice task. The model assumes that evidence is integrated at various time steps, until some threshold is reached and a commitment is made to one of the options.

As is depicted in Figure 3.7 the first of these models (a) is represented by a simple serial process, where deliberation fully precedes a choice that commits the agent to the preparation, and subsequent performance, of the chosen action—much in the same way that the 'classical sandwich model' highlights (Hurley, 1998). A parallel model (b) develops this by connecting the decision process to action preparation. This speeds up the agent's performance by anticipating what action will be most likely given the incoming sensory evidence. As evidence in support of one option increases, the agent can begin to make preparations for the respective action, before fully committing to it. Though the latter model gains a speed increase, it does so at the expense of accuracy. It could easily turn out that evidence that initially supports one option is overshadowed by later competing evidence, leading to inaccurate or clumsy actions. To deal with this speed-versus-accuracy trade-off, Lepora and Pezzulo develop the EC model (c), which, in addition to the parallel feed-forward connection, has a feedback connection that allows action dynamics (e.g. current trajectory and kinematics) to influence the decision-making process. Whereas the previous models consider decisions to be independent of ongoing action (only allowing for influence from prior experience), EC considers action as an integral part of the decision-making process, with proprioceptive signals feeding into the ongoing deliberative process to provide information about the biomechanical costs of associated actions.

Lepora and Pezzulo argue that the EC model accounts for this greater balancing of speed and accuracy by incorporating two key mechanisms. Firstly, unlike the serial model, the EC model enables what they term *action preparation* strategies, which allow an agent to alleviate delays when enacting a choice. For example, rather than waiting for a bound to be reached before commencing action, the parallel model and the EC model allow the agent to trade-off accuracy for speed, by starting an action on the basis of incomplete evidence. However, unlike the parallel model, the EC model allows for action performance to feedback into the decision process, and thus where the action dynamics alter the value of certain prospects, they create what Lepora and Pezzulo call *commitment effects* to the initially preferred choice. To highlight this, Lepora and Pezzulo (2015, emphasis added, pp. 4-5) discuss an example of a lion that has begun tracking a gazelle, deliberating over whether to switch and track another:

> "[...] if the lion waits until its decision is complete, it risks missing an opportunity because one or both gazelles may run away. The lion faces a decision problem that is not stable but dynamic. In dynamic, real-world environments, costs and benefits cannot be completely specified in advance but are defined by various situated factors such as the relative distance between the lion and the gazelles, which change over time as a function of the geometry of the environment (e.g. a gazelle jumping over an obstacle can follow a new escape path) and the decision makers actions (e.g. if the lion approaches one gazelle the other can escape)."

They continue:

> "[...] action dynamics in all their aspects (i.e. both their covert planning and their overt execution) have a backwards influence on the decision process by changing the prospects (the value and costs of the action alternatives). For example, when the lion starts tracking one of the gazelles, undoing that action can be too costly and thus the overall benefit of continuing to track the same gazelle increases. This produces a *commitment effect* to the initial choice that reflects both the situated nature of the choice and the cognitive effort required for changing mind at later stages of the decision."

A couple of comments are necessary. First of all, by being receptive to ongoing action, the EC model can consider changing biomechanical costs that are salient to the current decision. Although the serial and parallel models can incorporate action costs as well, they must do so *a priori*, as there is no way for the ongoing action to feedback into the deliberative process. Critics may argue that part of the developmental process for any organism is learning about the body, and associated biomechanical costs, which are not going to change that drastically, given the limited number of states that the body can be in. Therefore, prior knowledge of biomechanical costs can be incorporated through learning. This is surely correct, but is also incomplete. As the gazelle example should highlight, biomechanial costs are also partly dependant on the evolving state of the environment, and where other agents are involved, are unable to be precisely evaluated in advance.

Second, commitment effects make it harder to change your mind once an action is performed, because the later sensory information must outweigh the initial commitment that arises from having started an action. Situated agents that are receptive to subjective commitment effects may gain an important adaptive advantage, especially if the agent is able to learn about them for future interactions (see chapter 6).

Finally, Lepora and Pezzulo restrict the discussion of commitment effects to metrics that are relevant to simply visually-guided decisions. For example, a change in the trajectory of a mouse cursor, which represents the evolving choice of a subject to one of two targets, indicates a change of mind that only occurs once sufficient conflicting sensory evidence has accumulated. However, it is possible that commitment effects may also contribute to the existence of apparently irrational behaviour in more complex tasks (e.g. sunk-cost fallacy). We will pick up on this suggestion in chapter 6.

As well as dovetailing nicely with the embodied decisions account, Lepora and Pezzulo found their EC model to perform better in terms of speed and accuracy than the alternative models. Initially, the models were evaluated in two simulation studies representing a two-alternative forced choice task (see ibid., for details), which on its own stands as an interesting proof-of-principle. However, they also compared their models with empirical evidence from human studies, and found that the EC model was a good fit with human behaviour.

Taken together, the aforementioned properties of embodied decisions stand in contrast to the cognitivist assumptions of traditional decision theory. To reiterate, the cognitivist perspective of decision-making is strictly separated from evidence accumulation in perceptual systems, and the control of action in motor systems. However, embodied decisions view deliberation as a continuous competitive process within sensorimotor circuits, modulated by relevant biases from cortical and sub-cortical regions. It is hard to maintain the traditional functional separation of perception, cognition and action if we are to appreciate this process fully.

A number of issues remain. Firstly, although there is mention of 'continuously biasing influences' in the embodied decisions research, there is little explicit mention of the role of affective signals in the aforementioned work. This is of vital importance; an agent should have some way of determining which action opportunities it cares about most.

Secondly, the work in embodied decisions (Klaes et al., 2011; Pastor-Bernier and Cisek, 2011, see also) suggests that at least part of a prototypical cognitive process (decision-making) is inextricably intertwined with sensorimotor control, suggesting a blurring of the boundaries between perception, action and cognition. This view stands in contrast to decision-theoretic accounts that model humans as making decisions between different options by integrating the relevant factors into a single variable, such as subjective utility (Levy and Glimcher, 2012). For example, we saw in section 3.2 how some have argued that the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC) could integrate the relevant information and encode such an abstract value (Padoa-Schioppa, 2011). This conflict may appear to suggest that we should adopt one view or the other. However, as Cisek himself notes, “we are capable of making decisions that have nothing to do with actions, and in such situations the decision must be abstract.” (Cisek, 2012, p. 927) Therefore, instead of a straightforward conflict, it may be that the contrast between embodied decisions and neuroeconomics suggests a need for a two-systems approach, with different domains for the two approaches, rather than a strict incompatibility.

Before addressing these issues directly, we will explore how predictive processing shares many of the same motivations as embodied decisions. By doing so, we hope to uncover where the two frameworks can offer mutual development.

Chapter 4

Dissolving Boundaries

In this chapter we turn to explore how the research from the last chapter on embodied decisions connects with the PP framework. We will explore how an embodied account of PP blurs the boundaries between perception, cognition, action and emotion, and why this is relevant to understanding how PP connects up with embodied decision-making. We start by looking at how perception and action are intertwined.

4.1 Active Inference

An embodied account of PP eschews the idea that perception is a passive accumulation of sensory evidence with the purpose of reconstructing some detailed inner model of the world (Burr and Jones, 2016; Clark, 2016a). Instead, according to embodied PP, perception has the function of guiding actions that keep the organism within homeostatic bounds and maintain a stable grip upon its environment (Friston et al., 2010). We will unpack this claim more fully across this section and section 4.2.

Clark (2015) sees this version of the PP framework as a contemporary expression of many of the key motivations highlighted by the theory of interactive vision

(Ballard, 1991; Churchland, Ramachandran, and Sejnowski, 1994). This theory, as expressed by Churchland, Ramachandran, and Sejnowski (1994), took issue with an idea they dubbed the ‘pure vision’ strategy. According to this idea, vision passively reconstructs a rich inner representation (percept) from two-dimensional sensory data, which can subsequently be used to perform many different tasks. This reconstructive process also occurs largely independently of other sensory modalities, previous learning, goals, motor planning, and motor execution, and is reminiscent of the classical-sandwich model mentioned earlier.

In contrast to this model is the ‘interactive vision’ picture, which has come to be known simply as ‘active vision’. One of the motivations behind the active vision theory is that a perfect internal recreation of the organism’s world is not just unnecessary, but also computationally intractable and maladaptive (also see chapter 3, section 3.1.1). They base this argument on several claims, which include the idea that vision has its evolutionary rationale in motor control, and as such vision only needs to partially represent the most salient information, where salience is determined by an organism’s interests, goals and additional factors relating to the properties of a stimulus. To defend this position, they argue that vision is inherently exploratory and predictive, aided by learning from previous behaviour, and further governed by simple facts regarding our embodiment (e.g. size and placement of our visual apparatus, including the relations to effectors). This idea points to a neurophysiological picture far removed from that assumed by cognitivism. The idea that the connection between the motor system and the perceptual system is made only once the visual scene has been fully reconstructed and interpreted by distinct cognitive processes is simply false according to active vision.

More recent neurophysiological evidence corroborates this account. Cisek and Kalaska (2010) comment on experiments that show how neural responses in simple visual tasks are observed rapidly throughout the dorsal visual system, and engage motor areas such as the frontal eye fields in approximately 50ms. They state that this is significantly earlier than other visual areas such as V2 and V4. What could explain this shortcut to motor-related areas? One thought, which is sympathetic to the active vision theory, is that these neural responses are not to be thought of as simply visual, but action-oriented. That is, they specify visual information that has the purpose of specifying potential action opportunities. Many of these motivations are also present in embodied accounts of PP (e.g. emphasis on prior beliefs, and vision as a predictive process). Clark emphasises the following role for PEM:

> "[...] it is the guidance of world-engaging action, not the production of 'accurate' internal representations, that is the real purpose of the prediction error minimizing routine itself." (Clark, 2016b, p. 168)

This shift in emphasis requires a reinterpretation of the related notions of perceptual inference and active inference. Recall, these terms refer to the two ways that prediction-error can be minimised. Either the system can update the parameters of the inner model, in order to generate new predictions about what is causing the incoming sensory data (perceptual inference), or it can keep the generative model fixed, and resample the world such that the incoming sensory data accords with the predictions (active inference).

However, although both play an important role in PP, for Clark (ibid., p. 124), the primary role of perceptual inference is to "prescribe action", and as such, he states, that our percepts, "are not action-neutral 'hypotheses' about the world so much as ongoing attempts to parse the world in ways apt for the engagement of that world." This is a thoroughly action-oriented account, and acknowledges the earlier motivation of the active vision theory, which deemed a perfect internal recreation of the organism's world computationally intractable and maladaptive. Importantly, it is also this shift in emphasis that exposes a unity between Clark's account of PP and the insights of the ACH (ibid., p. 181). Recall, one of the claims made by the ACH was that neural representations were of *action opportunities*, rather than organism-independent, objective properties of the world. In addition, Clark views 'active inference' as a more-encompassing label for the combined mechanisms whereby the perceptual and motor systems cooperate in a dynamic and reciprocal manner to reduce prediction-error by exploiting the two strategies highlighted above. Active inference is accomplished using a combination of perceptual and motor systems rather than being confined to the latter that are traditionally associated with action.

There are two other consequences of this shift in emphasis. Firstly, this view diverges from the one introduced in chapter 2 in which perception equates to perceptual inference, and action to active inference. We have argued elsewhere that on this basis it is misleading to simply equate perceptual inference with perception and active inference with action (Burr and Jones, 2016). Although they are importantly linked by their shared role in prediction-error minimisation, there is nevertheless a distinction to be made at both the personal and sub-personal levels. Instead, we take perception to be an active exploration of the environment, involving a continuous (and simultaneous) unfolding of *both* perceptual inference and active inference. Similarly, action involves *both* altering the environment by changing one's bodily state, and monitoring the ongoing changes. In this manner, perception and action, understood at the personal level, involve a combination of both perceptual and active inference at the level of underlying cognitive processing. This is not to reject the important distinction outlined earlier between perceptual inference and active inference. After all it is presumably possible to construct an artificial system that engages in purely passive perceptual inference. However, an important lesson from the theory of active vision (and certain theories of embodied cognition) is that, *for organisms like ourselves*, perception is never merely a process of passive perceptual inference—perception always involves an active exploration of the environment. This is not because passive perception is impossible but because active perception allows us to access more information by exploiting the reliable and predictable bodily relations between motion and sensory input (i.e. sensorimotor contingencies). By intervening on causal relations, an agent can learn, and indeed shape, the causal structure of its environment, all the while testing the accuracy of its inner models. This point connects directly with the second consequence.

The lessons of the active vision theory (Churchland, Ramachandran, and Sejnowski, 1994), research in sensorimotor theory (e.g. Noe, 2004), and now also PP, is that perception and action are not separable in any meaningful sense at the level of cognitive or neural mechanisms. By providing a common underlying imperative to minimise prediction-error, PP goes further by arguing that at the level of cognitive processes, perception and action rely on the same principles of perceptual and active inference. As such any strict boundary between the processes is undermined. Similar views have led some to argue for an anti-representational view of perception (Chemero, 2011; Orlandi, 2014), because of the direct coupling of sensorimotor circuits. Some may worry that this causes a potential problem for PP accounts, which explicitly rely on representational generative models. This topic deserves special treatment in its own right, and although a lot of the material covered in this thesis is of relevance, this topic is not the primary aim of the thesis.[1] It will suffice to state that it is possible to maintain the action-oriented nature of perception without taking the radical step of eliminating representations altogether. Instead, as has been argued previously, one can maintain that perception represents the world in an action-oriented manner (Clark, 1997a; Mandik, 2005). As such, the seemingly representational nature of PEM is no reason to discount the potential significance of action-oriented perception.

[1] We have covered the topic in more detail in (Burr and Jones, 2016).

Finally, as well as following in the tradition of active vision, this view is also supported by recent neuroanatomical evidence that suggests a close relationship in the functional anatomy of the perceptual and motor systems (Adams, Shipp, and Friston, 2013; Shipp, Adams, and Friston, 2013). As we saw in chapter 2, research by Adams, Shipp, and Friston (cf. 2013), Friston, Mattout, and Kilner (2011), and Shipp, Adams, and Friston (2013) collectively supports one of the core claims of PP, which states that action is accounted for by a downwards cascade of predictive signals through the motor cortex to elicit motor activity, in much the same way as predictions descend through perceptual hierarchies. By demonstrating a deep continuity in the functional profiles of sensory and motor systems, this work also supports a dissolution of the boundary between perception and action as realised at the level of cognitive and neural mechanisms.

A deeper point can be teased out of this work. Any of the lower-level predictions will be constrained (and importantly contextualised) by higher-level models that function as multimodal predictions of the sensory evidence arising from both exteroceptive and proprioceptive causes. Although we can describe a particular anatomical region as visual or motor cortex, understanding the region's functional profile requires an appreciation of the current larger-scale dynamics of the brain as a whole, and specifically the networks that a particular region is effectively connected to (more on this in chapter 5). In the case of PP this means an appreciation of how the higher-level predictions contextualise the dynamics of the lower-level regions, but also an appreciation of how the lower level dynamics in turn bias and select the higher-level predictions. Understanding this reciprocal relation between incoming error signals and descending predictions is important for understanding how PP accommodates decision-making.

4.2 Predicting Choices

In PP, choices are made between competing higher-level predictions about expected sensory states. The formal basis for this perspective is based on the *free-energy principle* (Friston, 2010).[2] Friston et al. (2014) extend this account to decision-making in terms of active inference. Friston describes choices as 'beliefs about alternative policies'. A *policy* is defined as a control sequence, which is a trajectory of sensory expectations associated with a sequence of descending proprioceptive predictions that determines which action is selected next.[3] For example, there will be a sequence of sensory expectations associated with the movement made to open a cupboard and grasp a mug. This sequence can also be decomposed hierarchically, with different sequences expected at the corresponding spatiotemporal scale (also see chapter 6). Pezzulo, Rigoli, and Friston (2015) have provided a formal argument for how these policies can be acquired (and optimised) through experientially-based reinforcement learning. Policies are selected under the prior belief that they minimise the prediction error between attainable and desired outcomes, and on the basis of a belief in their expected precision. This line of thought bears a close resemblance to work in optimal control theory, which explores how optimal movement brings about valuable states for an organism.

Within this literature, Daniel Wolpert (2012) has demonstrated the close ties

[2] The relationship between PEM and the free-energy principle is introduced and explored in Chapter 2 of (Hohwy, 2013). For present purposes, it is not necessary to explore the connection in any formal detail. It will suffice to follow the claims of Hohwy that under some simplifying assumptions free-energy minimisation can be recast as PEM, and as such the free-energy principle is a more general and more encompassing framework.

[3] In the case of dynamical systems a trajectory is defined as a path through successive positions in state space (i.e. the space defined by the set of all possible states for the system) (Chemero, 2011).

between optimal control and decision-making, and argues that choice behaviour may be viewed as a problem of maximising the utility of performing some behaviour, where the consequence of this behaviour is associated with an option. This requires the agent to model, among other things, a cost function associated with the behavioural sequence, in order to accommodate the potential cost of performing some behaviour (e.g. expended energy, task uncertainty). This cost function is then minimised in order to select the optimal control sequence. Optimal control theory assumes that movement is caused by the minimisation of this cost function (Körding and Wolpert, 2006). One key difference between the views expressed by Wolpert and PP, however, is the latter's rejection of the separate representation of *cost functions*.

In PP, cost functions are *absorbed* into the generative model, becoming intertwined with the expectations of some policy (control sequence). As these expectations will have been shaped by learning, it is argued that there is already a prior belief about a policy's value or cost implicit in the existence of a generative model—a value based on previous error-based learning and captured by the extent to which it successfully minimises prediction error through action (Friston et al., 2014). Some may worry that this view eliminates too much, and that the need for encoding some measure of the value associated with an outcome is necessary to explain why certain behaviours are preferred over others. Moreover, it seems necessary for agents to represent value independently of beliefs. For example I can believe that it is more busy on the roads during rush hour, but unless I value my safety whilst cycling I may not deem it sensible to wear a helmet. How can we respond to such a worry?

It is important to reiterate that an implicit notion of the cost (or value) of some policy is not absent, but merely subsumed within the generative models, and thus associated with the sensory consequences of some policy and its expected precision. In other words, there is no additional cost function encoded over and above the already existing prior beliefs that are necessary for controlling action. One of the motivations for absorbing the cost functions into the generative model is an acknowledgement that in ecologically-valid scenarios agents must also optimise the behavioural routines that are associated with the desired option, and not just the outcome itself (e.g. smooth trajectories of motion rather than jerky motion). For example, even if A is strictly preferred to B, the sequence of behaviours that bring about A may require a large amount of energy to perform, and this trajectory or sequence of behaviours may itself be what determines the real cost to the agent. An integral part of the learning process for any agent is learning how to most efficiently bring about some desired state, dependent on states of the environment and their internal states.

As these *policies* are decomposable into sub-routines (e.g. getting dressed requires a number of steps, and each of these steps can be done clumsily or carefully), representing the behaviour with a single cost function overlooks the separable control sequences that are likely governed by distinct neural control mechanisms (more on this in chapter 6). To act in an optimal manner, especially in the pursuit of long-term distal goals, requires the careful co-ordination of numerous motor trajectories, and as Friston (2011c, p. 488) notes, "we know from the physics of flow that motion cannot be specified by a single value function."

This is not to deny the possibility of representing the coarse-grained behaviour of the whole agent at some higher-level as implicitly minimising a cost function. However, it seems that in terms of understanding the complex dynamics of choice behaviour and sensorimotor control, such representations are likely to be only instrumentally valuable for an observer, and quite likely to mislead entirely. By positing a separate cost function, which is represented independently of the policies being considered, the value of a goal-state becomes separated from the actions required to bring it about. This means positing additional mechanisms, which can a) encode a representation of an abstract cost function that is associated with some external goal-state, and b) integrate the cost function with the policy in real-time when deliberating over some choice. Although there is nothing implausible about this requirement *a priori*, it stands in contrast to the findings outlined in the previous chapter.

For example, Lepora and Pezzulo (2015) argued in favour of viewing the ongoing perception of bodily dynamics (e.g. through proprioception), during the overt performance of some choice behaviour, as an integral part of the decision process. This has the added benefit of allowing the agent to adjust the implicit value of the available policies on the fly to meet the changing demands of the environment. Until certain actions are performed, it is not possible to consider the myriad ways in which the environment will present or restrict action opportunities that an agent may have anticipated. An agent who is unable to accommodate these changes in a fluid manner will likely be at a disadvantage in an uncertain and constantly changing environment. Therefore, as Clark acknowledges:

> "By re-conceiving cost functions as implicit in bodies of expectations concerning trajectories of motion, PP-style solutions sidestep the need to solve difficult (often intractable) optimality equations during online processing and—courtesy of the complex generative model—fluidly accommodate signalling delays, sensory noise, and the many-one mapping between goals and motor programs. Alternatives requiring the distinct and explicit computation of costs and values thus arguably make unrealistic demands on online processing, fail to exploit the helpful characteristics of the physical system, and lack biologically plausible means of implementation." (Clark, 2015, references suppressed, p.11)

This connects with a second point, regarding the earlier assumption made by control theory, that movement is *caused* by some value representation. Friston (2011c, p. 488) rightfully states that "value is an attribute of states that are caused by movement: it is a consequence, not a cause." However, it is often assumed that an agent must have a representation of this value separate from its beliefs, in order to ground the basic capacity for desire. When combined with a relevant belief, desire provides the jointly sufficient conditions for motivating action according to Hume's belief-desire thesis. Unfortunately it is not easy to determine what Friston's stance is on this thesis. Take the following quote:

> "I can believe I am being drenched by rain and yet place a high cost on this state of affairs. However, if I believe that I will seek shelter when it rains, then I will behave optimally, *provided I act to fulfil these beliefs*. Note that these prior beliefs are not about states of the world but transitions among states (i.e., a policy)." (Friston, Samothrakis, and Montague, 2012, p.524, emphasis added)

Initially it appears as though Friston is rejecting the requirement that desires play any sort of motivational role. Not only does he eschew explicit cost functions, in favour of prior beliefs about policies, but he seems to argue that the desire to avoid getting wet from rain can be accounted for by a string of beliefs regarding transitions among inner control states, which are connected to each other in a manner that will bring about optimal behaviour. However, note the inclusion of the proviso in the quote that optimality requires the agent to "act to fulfil these beliefs". This seems obvious to the point of triviality, but its inclusion may be problematic for Friston as it seems to leave unspecified what the motivation is for the agent to act in the first place. Philosophers will be keen to highlight this problem by reiterating Hume's belief-desire law—this motivation could only come about from the jointly sufficient conditions of both the possession of a belief and the desire to bring about the state represented by the belief. One possible way of understanding Friston's point on this, comes from separate work where he spells out the reasons why the free-energy principle implies embodied cognition. Here, it is worth quoting a passage at length:

> “The free-energy formulation starts with the premise that biological agents must actively resist a natural tendency to disorder. It appeals to the idea that agents are essentially inference machines that model their sensorium to make predictions, which action then fulfils [...] The free-energy formulation generalises the concept of agents as inference machines and considers each agent as a statistical model of its environmental niche (econiche). In brief, the free-energy principle takes the existence of agents as its starting point and concludes that each phenotype or agent embodies an optimal model of its econiche. [...] the statistical model entailed by each agent includes a model of itself as part of that environment. This model rests upon prior expectations about how environmental states unfold over time. Crucially, for an agent to exist, its model must include the prior expectation that its form and internal (embodied) states are contained within some invariant set. [...] Therefore, if the agent (model) exists, it must a priori expect to occupy an invariant set of bounded states (cf., homeostasis). Heuristically, if I am a model of my environment and my environment includes me, then I model myself as existing. But I will only exist iff (sic) I am a veridical model of my environment [...] This tautology is at the heart of the free-energy principle and celebrates the circular causality that underpins much of embodied cognition.” (Friston, 2011a, pp. 89-90)

Though we have seen some of these assumptions (derived from systems approaches in biology) earlier in Chapter 2 (also see Figure 2.7), this quotation needs unpacking.

We can draw out a couple of notable points of discussion, which I will label as follows:

The Autopoiesis Assumption: biological systems are self-producing (autopoietic) systems that occupy a limited range of states.

The Econiche Assumption: autopoietic systems must embody a model of their environment, which must also necessarily be a model of the physical states of the agent (i.e. its body).

The Circular Causality Assumption: any agent that meets these criteria must necessarily exist in order to be able to model itself as existing, which in turn contributes to its ongoing existence.

The first of these points will be familiar from work in the enactivist approach to embodied cognition. This approach seeks to understand the continuity between life and mind by exploring how the latter is brought about (i.e. enacted) through the interactive processes of life as an autonomous, self-organising process. Stemming from its development in biology (Maturana and Varela, 1980), and later discussed in philosophy of mind (Thompson, 2007), the theory of autopoiesis is concerned with the dynamic, *self-producing* processes that sustain life. The neurobiologists Maturana and Varela coined the term *autopoiesis* to stand in as a label for the processes of circular-organisation (see chapter 1), which they believed constitute the basis of life. Autopoiesis, they argued, was necessary to account for the apparent unity that is perceived in living systems in their expression of autonomy, in spite of the continuous perturbations from the external environment that threaten disorder. The key example of an autopoietic system pointed to by Maturana and Varela is a biological cell.

They present a cell as a set of chemical interactions, bounded by a semi-permeable cell membrane. The membrane maintains a favourable inner environment of chemical concentrations (i.e. a limited set of states) by constantly and selectively allowing transportation of chemicals from the extra-cellular environment into the cell, and disposing waste products from the inner processes of the cells interactions. However, the cell membrane is also maintained by the processes that it serves to bound. In this sense, the membrane and the reactions can be treated as an operationally closed set. Although this provides a boundary between the cell and the membrane, the two are closely coupled by virtue of the ongoing processes. By passing waste products to the extra-cellular environment, the cell is acting on its environment, which in turn impacts the cell by altering its chemical concentrations. The cell is an autopoietic system that is coupled with its environment. With this example in mind, Maturana and Varela (1980) outline a number of salient properties of autopoietic systems:

Autonomy: an autopoietic system is organised as a network of processes, which themselves produce the components that interact to sustain and realise the network of processes that realised them in the first place (e.g. the cell membrane). Thus, an autopoietic system is a homeostatic system which has its own autonomous organisation as the fundamental variable which it aims to keep constant (or within a narrow range of parameters).

Unity: by aiming to keep their organisation as invariant as possible, autopoietic systems maintain an identity (or unity) through the specification of their own boundaries in the processes of self-production (e.g. the cell as separate from the extra-cellular environment).

Perturbation: an autopoietic system is not characterised functionally by way of inputs or outputs, but can nevertheless be perturbed by independent external events. These perturbations must be compensated for by internal structural changes (e.g. flushing waste products), in order to maintain homeostasis.

Together these properties help us understand the motivation for the three assumptions. By treating biological agents as inference machines, autopoietic systems can be formally cast as (actively) modelling their environment in anticipation of external perturbations that threaten the autonomous, self-regulating processes that define them. By successfully resisting the tendency to disorder, any self-maintaining system can be treated as actively inferring its environment in order to adaptively respond to external perturbations. Doing this requires an organism to make structural alterations to its inner environment in order to maintain homeostasis, and thus brings the states of an organism's body within the remit of the environmental model.

The formal basis for this assumption comes from work in cybernetics, and more specifically the *good regulator theorem* of Conant and Ashby (1970). This proved that under broad assumptions any successfully self-regulating system embodies a model of its environment.[4] In so far as the free-energy principle treats the brain as a self-regulating (self-organising) system (Friston, 2010, cf.), it follows that the brain must also embody a model of its environment, which is taken to include the physical states of the organism that it controls. This is because adaptive responses for organisms such as ourselves require structural changes that involve motor control (e.g. reaching to grasp food for energy intake; running to avoid predators). This active, dynamic element to adaptive behaviour means that an organism's expectations must equally be dynamic (hence the focus on policies rather than stable goal states), requiring the agent to model itself as in a constant source of fluctuation. Finally, if an agent finds itself in this situation, then it must necessarily have resisted previous external perturbations and successfully maintained homeostasis long enough to be in a situation where it embodies a model of its environment. As Friston (2011a, p. 90)

[4] Importantly, this does not mean that the system trades in inner representations as commonly understood in philosophy of mind, but is a far weaker notion of modelling. See (Burr and Jones, 2016) for discussion.

states “I will only exist iff (sic) I am a veridical model of my environment.”

Friston’s conception raises a whole host of tricky epistemic questions. We have already seen one of these issues in chapter 2, when we explored the motivation behind the selection of a specific partition of states. We stated above that insofar as the free-energy principle treats the brain as a self-organising system, it follows that the brain must also embody a model of its environment. This is true, but overlooks the fact that the same can be said for the entire organism as well. We must therefore appeal to additional factors (as we shall do over the course of this thesis) if we wish to argue that either the brain or the body is the model that we should be interested in.

A further worry is that the embodied model Friston posits requires an observer’s perspective for its content, which raises the problem of how the agent itself has epistemic access to the representational content of the aforementioned models. Many enactivists would argue that this motivates the need for an anti-representational interpretation (e.g Chemero, 2011; Hutto and Myin, 2013) devoid of any contentful interpretations entirely; a position which resists the desire to ascribe contentful states to the agent (e.g. propositional attitudes). This requires careful philosophical analysis to evaluate, but as previously stated, we will not discuss this possibility in any further depth.

Finally, it seems as though this account assumes a proximal explanation, and omits the sort of ultimate explanations pursued by evolutionary biologists. It is common to keep what are sometimes referred to as ‘how’ and ‘why’ questions separate in the cognitive science, with the former often appealing to mechanistic explanations to provide answers for how some phenomenon is produced, whereas the latter attempt to elucidate the adaptive significance of cognition and behaviour. It may be that by eschewing the use of cost functions in PP, defenders are restricting their claims only to ‘how’ questions, and make no claims regarding ‘why’ questions. This seems unsatisfactory, given that if we are to take the replacement of cost functions with expectations over policies seriously, then we appear to be forced to accept that natural selection must have operated on these policies' underlying mechanisms and selected them because of some evolutionary advantage—thus committing ourselves to an answer to the 'why' question regardless.

Regretfully we have no firm answers to these problems in Friston's account, though we will point towards a tentative proposal in Chapter 6. For the time being, it appears that PP is committed to the claim that cost functions are implicit in bodies of expectations concerning policies, rather than explicit, detachable representations. Whether this represents an improvement depends on your perspective:

> "In one sense, active inference replaces a hard optimal control problem with a hard inference problem. Having said this, the nice thing about active inference is that these problems can be solved in a simple and neurobiologically plausible fashion: by effectively equipping predictive coding schemes with classical reflex arcs. Perhaps the most definitive argument in favor of active inference, as a normative model of motor control, is that prior beliefs about behavior emerge naturally as top-down or empirical priors during hierarchical perceptual inference. This contrasts with optimal control, which, at the end of the day, still has to explain how cost functions themselves are optimized. In short, active inference eliminates the homunculus implicit in cost functions." (Friston, 2011c, p. 492)

If the promise of a more integrated and unified framework does not appeal to the reader, Clark (2015) also notes that many working roboticists have turned away from the explicit encoding of value/cost functions for pragmatic reasons, arguing that they are too inflexible and biologically unrealistic due to their computational demands. Instead, they favour approaches that likewise exploit the complex dynamics of embodied agents (e.g. the approach of Lepora and Pezzulo (2015)), which are computationally less demanding. These approaches acknowledge that the physiological constraints of an agent provide implicit means of understanding the value associated with dynamic action performance, without the need for additional abstract neural representations (see section 6.3.1).

In addition, as we saw in the previous chapter, there are a number of debates in decision theory about whether the brain does in fact calculate value, with some arguing in favour of some abstract form of a neural 'common currency' (Levy and Glimcher, 2012). As Vlaev et al. (2011) argues, these views are beset with difficulties both from behavioural studies that explore contradictory, empirically-observed context effects (e.g. preference reversals, prospect relativity and various memory effects), as well as competing neurophysiological studies, which favour alternative approaches (see section 4.3.1). By eliminating the explicit encoding of value from its models, PP avoids these worries.

In these first two areas, the development of the active inference framework has much to offer, and we can begin to see the extent of the unifying scope of PP. However, we also seem to encounter the same problem faced by the embodied decisions research: how exactly does an agent select between the large number of simultaneous action opportunities available to it, especially if it has no direct access to an explicit representation of value? To answer this question, and provide the final response to the above challenge, requires bringing the body more closely within the remit of PP.

4.3 Embodied Emotions

> "[...] in order to have anything like a complete theory of human rationality, we have to understand what role emotion plays in it." (Simon, 1983,

p. 29)

Within the ecological psychology tradition, the environment is considered as providing the agent with a number of possible action opportunities (or affordances), rather than merely as a series of causes that push the agent around in myriad ways. But how does an agent determine which of the myriad action opportunities are salient, or to adopt the terminology of Withagen et al. (2012), how do *affordances* become *invitations*? Recasting this question in terms of how an agent decides between the simultaneous action opportunities present in its environment, allows us to demonstrate how PP is able to extend the notion of policies discussed in the previous section to accommodate a number of additional areas of research. In this section we will explore how PP connects with research in the decision-making literature by exploring the key role of emotion. We then turn to see how an understanding of neuromodulation (chapter 5) can further bolster this picture and strengthen the connection between PP and embodied decisions. However, even after discussing these topics, the picture will remain incomplete, and will require us to look further into the physiological constraints of the agent's body, as well as additional constraints that come from an agent's environmental niche (chapter 6).

The above quote by Herbert Simon highlights the limitations of any revisions to decision-theoretic models that fail to include some important role for emotions. In a recent review of the psychological research on emotions and decision-making, Lerner et al. (2015, pp. 800-801) begin with the previous quote from Herbert Simon, and subsequently claim that "many psychological scientists now assume that emotions are, for better or worse, the dominant driver of most meaningful decisions in life". If their statement is true, then it appears as if Simon's advice was heeded. However, in spite of the clear review that they offer, the statement by Lerner et al. is ambiguous, and perhaps even a little misleading. First of all, although emotions undoubtedly play a role in influencing decision-making, as we will see shortly, it is far from a trivial claim to suggest that they are the "dominant driver of most meaningful decisions in life", rather than being a necessary contributor in a collection of additional biasing factors. As we will see shortly, this is because it is not always easy to separate emotion from cognition or perception, and, secondly, it is not always clear what constitutes an emotional episode in the first place.

Although the full details are beyond the scope of this thesis, a few points should be made regarding the nature of emotions as considered within the psychological and philosophical literature. Let's begin by offering a tentative definition of some key terms: 'core affect' and 'emotions'. Here, I shall follow James Russell and Lisa Feldman Barrett (Russell and Barrett, 1999, 2009) in differentiating the terms as follows:

Core Affect: a neurophysiological state that is consciously *accessible* as a simple primitive non-reflective feeling most evident in mood and emotion. Core affect is experienced constantly as a single feeling, but the nature and intensity varies over time according to two scales. These scales are known as degree of valence (e.g. pleasure versus displeasure) and degree of arousal (e.g. feeling energetic versus enervated). It can be caused by external factors or internal factors, some of which may be beyond the ability for an agent to perceive. For example, in cases of object-free disorders such as depression, core affect can be free floating. Alternatively, core affect can be a component of object-directed emotions and moods. For example, feeling good about oneself is decomposable into the affective feeling of 'good', and the intentional (or cognitive) component that directs the affective state at oneself. An agent is always in a state of core-affect, but does not need to be conscious of it (the state must be accessible though). Furthermore, the state can extend for long periods of times (as in moods), or shorter periods (as in emotional episodes).

Emotion: a "prototypical emotional episode" or occurrence of an emotion is a complex structure often associated with an intentional object that can be real or imagined (e.g. anger towards a person, or fear of the bogeyman). The components of this structure include: (a) core affect, (b) an appropriate overt behaviour (e.g. smiling when happy, frowning when angry), (c) directed attention towards the eliciting stimulus (e.g. shifting gaze or mental attention) with corresponding cognitive appraisal of the stimulus, and attribution of the emergence of the episode to the emotion itself, (d) a consciously accesible experience of the emotion as involving oneself, and (e) the relevant physiological changes consistent with the emotion. Due to the intentionality of an emotion, the cognitive appraisal is often considered a key element of an emotion, and as such allows for constructivists to argue for the influence of sociocultural practices on emotions.

Two points are worth highlighting in these definitions. First, there is the multidimensional nature of emotional episodes, including components such as: changes in neurophysiological states, cognitive appraisal, behavioural response, intentionality and consciously accessible experience or feeling (i.e core affect). Second, there is the emphasis on core affect as a separable aspect of an emotional episode, dissociable from the cognitive appraisal. With regards to the first point, it is important to note that each of these components may be accepted or dismissed by a particular theory of emotions, and some theories may include multiple factors. Theories that emphasise several (so called hybrid theories) are commonplace, though within this collection of theories, some may choose to emphasise one component as the defining feature of an emotional episode.

In the case of the well-known 'somatic feeling theory' of James (1890) and Lange (1885), what characterised an emotional state was the experience of various felt changes in the body. As such, their theory emphasised both the feeling (or experiential aspect) as well as changes in neurophysiology. To motivate this idea, James asks his readers to consider an emotional state, and then to subtract away the phenomenal qualities that are associated with feelings of bodily changes. In his own words:

> "If we fancy some strong emotion, and then try to abstract from our consciousness of it all the feelings of its characteristic bodily symptoms, we find we have nothing left behind, no 'mind-stuff' out of which the emotion can be constituted, and that a cold and neutral state of intellectual perception is all that remains." (James, 1884, p.193, quoted in Prinz, 2004)

James argues that although this runs counter to the common-sense notion that emotions cause bodily changes, we should nevertheless think of emotions as being caused by the perceptual experience of bodily (or somatic) changes (i.e. you're happy because you smile, you don't smile because you're happy). Here 'somatic' is used to refer to any part of the body, including the respiratory system, circulatory system, digestive system and musculoskeletal system.[5]

One limitation of this theory is that by claiming all emotions are associated with perceptual experiences of bodily feelings, we are led to the prediction that patients with spinal cord injuries should therefore experience a subdued range of emotions, being unable to perceive many physiological changes. However, early results investigating this claim came to drastically different conclusions regarding the intensity of felt emotions in spinal cord injured patients (Chwalisz, Diener, and Gallagher, 1988; Hohmann, 1966). How should we construe the veracity of the somatic feeling theory in light of these challenges?

[5] The extent of bodily changes that are included within the set of registrable effects on emotions is not always agreed upon. James and Lange differed on what they included, and a recent extension of their work by (Damasio, 1994) extends the set further. See (Prinz, 2004) for a review.

A particularly notable response can be found in a more recent extension of the theory by Antonio Damasio (1994). Important to note, Damasio claims, is that although nearly every part of the body can send signals to the brain via the peripheral nervous system, which enter the brain at the level of the spinal cord or brain stem, this is not the only mode of influence that the body has over the brain. Additionally, chemical substances arising from bodily activity (notably the endocrine system) are also able to affect the brain via the circulatory system. Moreover, both of these channels provide a medium for the endogenous dynamics of the brain to reciprocally interact with the body. This raises two points. First, spinal cord injuries may diminish emotional episodes, but would not eradicate them altogether due to the possibility of continued influence from the endocrine system. Secondly, the somatic theory of feeling should not be taken as ruling out the impact of endogenous brain dynamics on the generation and co-ordination of emotional episodes.

For Damasio, and unlike James and Lange, this means that emotional responses can occur in the absence of bodily changes if brain centres ordinarily associated with a corresponding brain centre are active. In short, the brain runs what Damasio terms "as-if" loops, whereby the brain triggers *somatic markers*, which are neural representations of the bodily changes. These somatic markers can be used in the online processing of affective information, but can also be used offline (e.g. recalling some previous event). Importantly, these somatic markers are not only able to influence other neural processes, but Damasio argues are integral to our very ability to decide effectively.[6]

What has subsequently been termed the *somatic marker hypothesis* (SMH) has

[6] Colombetti (2008, p. 52) points out that Damasio was not the first to make this claim. Precursors can be found in James' somatic theory, and also in the work of de Sousa who argued that emotions assist 'pure reason' by retrieving relevant information—this is one way that Nature could help an agent to deal with the frame problem.

been incredibly influential in understanding the role of emotions in decision-making. The SMH proposes that these somatic markers, act as biasing signals in key emotional processes in the brain, with particular emphasis on areas of prefrontal cortex (Bechara, Damasio, and Damasio, 2000). Due to the affective nature of these markers (e.g. their valence and arousal), they act as signals for whether certain options are valuable for the agent, or whether some action is salient. Most of the empirical support for the SMH is based on performance in what is known as the Iowa Gambling Task, which was constructed to measure the performance of a subject's decision-making abilities. However, as with previous work, the conclusions drawn from these studies have not gone unchallenged. Some have argued that the very conceptual foundations of the SMH and Iowa Gambling Task have been poorly specified (Colombetti, 2008), and it is also unclear how the somatic feeling theory connects agents to meaningful interactions with the objects in the world that cause feelings in the first place.

Building on earlier work by Magda Arnold, Richard Lazarus (1991) sought to explain these meaningful aspects of emotions by appealing to what he termed their *core relational theme*. Core relational themes were built up of multiple molecular appraisals, and represent the defining characteristic of an emotional episode. For example, anger is defined as "a demeaning offense against me and mine", and is constructed from several molecular appraisals that are representations of organism-environment relations that bear on a subject's well-being (e.g. goal relevance, goal congruence and the agent's coping potential). Other emotions have different mixtures of molecular appraisals, and thus different molar core relational themes. Unlike early somatic feeling theories, appraisal theories such as the one defended by Lazarus, extend emotional episodes to include a key role for cognition, and thus provide a representational role to emotions that extend beyond the body out into the world in meaningful ways. As such, appraisal theories emphasise cognition over other components as the defining feature of emotional episodes. What defines an emotional episode, and differentiates it from others, is the cognitive appraisal of its intentional object (i.e. the object in the world, whether real or imagined) characterised by way of its core relational theme—or in other words, the associated thoughts and concepts that are consciously accessible to the agent.

This brings us to one final noteworthy development offered by Jesse Prinz (2004), known as the the *embodied appraisal* theory of emotions. Prinz argues, in line with the somatic feeling theory that emotions are perceptions of certain kinds of neurophysiological states, but extends this notion such that what is important about these kinds of states is that they reliably track salient conditions in the environment. For example, 'fear' is the perception of certain neurophysiological states that are reliably linked with dangers. To provide the necessary theory of content for his proposal, Prinz extends the idea of Lazarus's core relational themes, so that rather than describing the structure of an emotion, they instead pick out the content of these emotions. Under Prinz's theory, emotions become appraisals of organism-environment relations (e.g. a bodily state represents some intentional object due to its reliable connection with it), harnessed through perception of relevant bodily states. At first glance, this would seem to suggest that Prinz's theory offers a compelling companion for the active inference view discussed above, with its emphasis on sensorimotor dynamics. However, Prinz (ibid., p. 194) rejects the idea that emotions are essentially related to actions, opting for a separation of emotions from motivating tendencies and opting instead for the weaker notion of emotions as "action enabling". This is consistent with his representational account of emotions, but means that the content of emotions is distanced from the more action-oriented forms of representation defended by sensorimotor approaches.[7] As such, Prinz distances himself from action-

[7] Whether this means that Prinz's theory is *incompatible* with sensorimotor theories is another matter that will not be discussed here.

oriented theories of emotions (e.g. Nico Frijda's view of emotions as action-readiness patterns (Frijda, 1987, 2010)), as well as some recent work in cognitive neuroscience.

Reviewing a large body of neuroscientific data, Pessoa (2013) has recently outlined the many ways that cognitive and emotional processing interact and are integrated in the brain. One key area that his work focuses on is the amygdala, a group of nuclei in the limbic system that has long been associated with emotional processes. Pessoa argues that the amygdala's function goes beyond emotional processing, and is involved in shaping selective information processing. He describes the amygdala as a core structure in a system involved in "What is it?" and "What's to be done?" processing (Pessoa, 2010), which contributes to the specification and selection of salient action opportunities for the organism—in line with the affordance competition hypothesis. The processing of this affectively-laden information is constrained by a sort of neural bandwidth, and thus is not independent of attentional processes. Pessoa advocates replacing the cognitivist strategies of functional decomposition and localisation, in favour of a network architecture whereby emotion and cognition will fail to map cleanly into compartmentalized pieces of the brain (see chapter 5). This also leads to a dissolution of the boundaries between emotion and perception, a view defended by proponents of a predictive processing approach. To see why this is the case, let's turn to see how PP accounts for affectively-laden information processing.

4.3.1 Interoceptive Inference

In PP, the predictions generated by the inner models of the brain do not merely attempt to anticipate the flow of sensory input from the outside world, but also the flow of interoceptive inputs (i.e. pertaining to endogenously produced stimuli, e.g. bodily organs), which further constrain the set of viable actions in important ways. For example, deciding to quench one's thirst or sate one's hunger is often more important than allowing oneself to be distracted by other action opportunities.

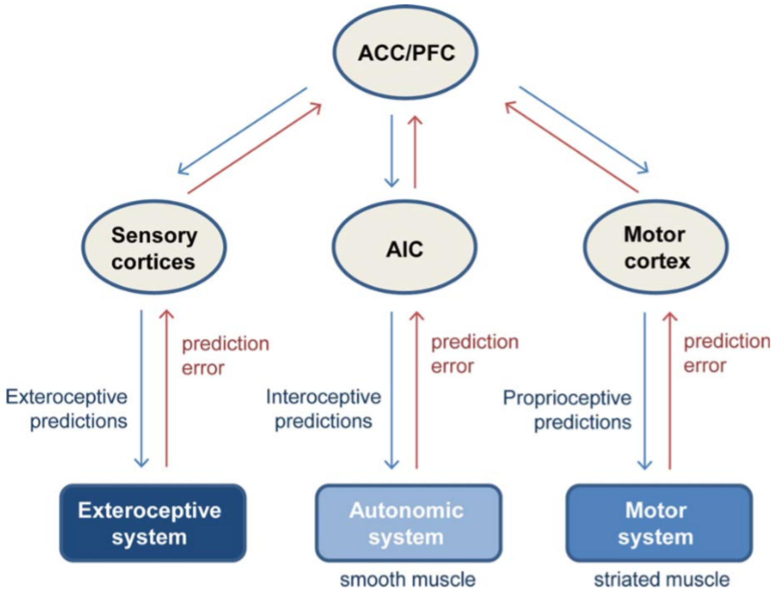


Figure 4.1: A proposed model highlighting a) the role of the AIC in interoceptive inference generating descending predictions sent to the autonomic system via smooth muscles to provide a point of reference for autonomic reflexes and b) the role of top-down predictions from regions such as the anterior cingulate cortex (ACC) and prefrontal cortex (PFC) integrating ascending prediction errors from exteroceptive, proprioceptive and interoceptive causes. Reprinted from (Gu et al., 2013, p. 3382)

Being receptive to the current state of your body, therefore, is fundamental to making adaptive decisions, as it allows us to determine which options have the greatest value relative to our needs.

Anil Seth (2015, p. 9) has argued that on this basis, PP may apply more naturally to interoception than to exteroception. He states that unexpected sensory states that pertain to interoception are more likely to be bad news for an organism (e.g. an unexpected level of blood oxygenation or blood sugar) than external states. Tracking this type of sensory information requires incorporating interoceptive information into the PP framework, and thus integrating affective information within the generative models harboured by the brain. Seth (2013) has argued that active inference should thus be extended to include interoceptive inference, and that key areas such as the anterior insular cortex (AIC) are well-suited to play a central role as both a comparator that registers top-down predictions against error signals, and as a source of anticipatory visceromotor control (i.e. the regulation of internal bodily states).

Here, Seth is developing on a recent meta-analysis of neuroimaging data by Gu et al. (2013), which he states is compatible with the active inference framework. He argues that the study provides initial evidence for the claim that descending predictions generated by the AIC are sent to the autonomic system via smooth muscles to activate autonomic reflexes in a similar manner as earlier described in the case of proprioception. This is important because the goal of interoceptive inference, as with active inference in general, is not simply the perceptual awareness of internal states. If we were to approach interoceptive inference as a case of perceptual inference, this would lead us to the strange conclusion that minimising interoceptive prediction error should be done by simply changing our models to fit the world. However, monitoring important signals such as those originating from our own bodies require adaptive responses as pertinent boundaries are reached (e.g. forage for food when blood-sugar levels are low). As Seth (2015) states: “interoceptive inference can be thought of as predictive control, in the same manner as active inference.” Therefore, interoceptive inference can be brought within the PEM schema, and as depicted in Figure 4.1, these interoceptive predictions can then influence higher-level multimodal predictions generated in regions such as the anterior cingulate cortex (ACC) and the prefrontal cortex (PFC).

This is immensely important for integrating decision-making within the PP framework, as it allows for a consistent understanding of the role that affective information (and possibly emotions) play in guiding our actions. By integrating interoceptive signals within the hierarchical generative models, the downwards predictions that are responsible for generating both perceptual content and motor behaviour also have affective significance.

> "These models instantiate predictions of temporal sequences of matched exteroceptive and interoceptive inputs, which flow down through the hierarchy. The resulting cascade of prediction errors can then be resolved either through autonomic control, in order to metabolize bodily fat stores (active inference), or through allostatic actions involving the external environment (i.e., finding and eating sugary things)." (ibid., p. 10)

This connection has not gone unnoticed by a number of researchers (Lerner et al., 2015; Phelps, Lempert, and Sokol-Hessner, 2014, e.g.). Some of these studies (e.g. Phelps, Lempert, and Sokol-Hessner, 2014) echo the sentiments of the earlier embodied decisions work, but in line with Pessoa (2013), they go further in demonstrating how specific biasing inputs, such as affective information, play a fundamental modulatory role in the competitive process of action selection. This provides an important extension to the embodied decisions research, which was initially left underdeveloped. It shows how affective signals are able to provide a basis for determining the salience of potential actions when integrated within the hierarchical generative models of PP (Barrett and Bar, 2009; Lindquist et al., 2012; Ouden, Kok, and De Lange, 2012).

The inclusion of core affect (see above) provides a way for an agent to know if some action is salient (i.e. good or bad for it), while the contextualising appraisals from top-down influences situate this within a wider web of agent-specific knowledge about bodily-environmental relations. Importantly, this evaluation need not be considered as a separate step in a computational process. Barrett and Bar (2009) argue that activity in OFC is reflective of ongoing integration of sensory information from exteroceptive cues with interoceptive information from the body. They claim that this supports the view that perceptual states are “intrinsically infused with affective value”, such that the affective significance (or salience) of an object (or action opportunity) is intertwined with its perception. Affective information is thus brought within the same scheme that we saw accounting for the rapid visual comprehension in earlier chapters (Bar, 2011b), and provides further evidence against the cognitivist picture.

One worry here is that different sources of sensory information have often been treated as conveying distinct kinds of signals to the agent. For example, perceptual content is often seen as carrying indicative content, whereas affective content has an imperative, or motivating aspect to it. By reducing all types of sensory information to prediction errors, PP may lose the ability to distinguish between these types of content.

A response to this problem comes from Ouden, Kok, and De Lange (2012), who review the neurophysiological evidence relevant to an understanding of prediction errors, and argue that there is support for multiple kinds of prediction errors (PEs) in the brain: perceptual PEs, cognitive PEs and motivational PEs. The first two types are referred to as *unsigned* PEs. These do not reflect the valence of any sensory input, but simply the surprise of its occurrence. The final kind of PEs, however, are known as *signed* PEs, for they reflect whether an outcome was better or worse than expected. They state:

> “Signed PEs play a central role in many computational models of reinforcement learning. These models describe how an agent learns the value of actions and stimuli in a complex environment, and signed PEs that contain information about the direction in which the prediction was wrong, serve as a teaching signal that allows for updating of the value of the current action or stimulus.” (ibid., p. 4)

Having access to multiple kinds of PEs, including those with affective significance, may provide the brain with the means to implicitly compare and evaluate which policy is most desirable based on prior learning. This is important for ecological considerations as not all errors are created equal. To illustrate this, we can turn to Hammerstein and Stevens (2012, p. 9) who describe what is sometimes referred to as the "smoke detector principle":

> "Natural selection will likely favor the avoidance of even small errors if they incur high costs in terms of fitness. In contrast, seemingly large errors (e.g., a male mating with a member of the wrong species) may not face strong selective pressure if they have little impact on fitness."

In short, a false alarm from your smoke detector may be an annoyance, but it is preferable to a smoke detector not triggering in the case of a real emergency. One way to minimise instances of the latter kind is to integrate different sources of information, and in some sense hedge your bets. Therefore, if value is determined through the comparison of multiple PEs, this would allow the agent to determine which of the myriad possible action opportunities is most salient given its current needs, and at the same time minimise risk of selecting inappropriate actions based on the response to a single PE. The comparison could take the form of a distributed competition, in line with the proposal offered by Cisek (2012), with no need for an abstract encoding of value that is generated after the reconstruction of perceptual information.

This connects with a related topic in the decision-making literature, which centres on the question of whether, and how, the brain calculates value? Vlaev et al. (2011) review a range of theories and models and provide the following positions to help capture these commitments:

1. Value-first position: the brain computes the value of different options and

simply picks the one with the highest value.

2. Context-dependent value: the brain computes values, but the choice is heavily context-dependent on the set of available options.

3. Comparison with value computation: the brain computes how much it values options, but only in relation to other values.

4. Comparison-only: choice depends on comparisons without any computation of value.

Micro-debates exist within each of these positions. For example, is value represented on some ordinal, interval or ratio scale, and what objects are represented? Regardless of how these debates turn out, it should be clear that the value-first position is incompatible with both PP and the embodied decisions account. This is because value-first positions maintain that the value of an option is stable, and explicitly represented in some region of the brain such as OFC or vmPFC (Padoa-Schioppa, 2011). We have already seen that the embodied decisions account is opposed to such a view, due to conflicting neurophysiological evidence. In addition, we have seen how PP eschews the explicit representation of value/cost functions altogether. However, it is unclear which of the alternative positions would best describe an embodied account of PP.

Adopting the suggestion offered by Ouden, Kok, and De Lange (2012) of multiple PEs seems to frame PP as either an example of the 'context-dependent value' view or the 'comparison with value computation' view, depending on which additional mechanisms are posited to co-ordinate or integrate the options based on the type of PE considered. For example, although PP eschews talk of explicit cost functions, there is nevertheless a non-trivial sense in which the brain is comparing the expected values of the predictions that stand in place of the cost functions. Given the uncertainty regarding the precise implementational details of an exact architecture for

PP (see Clark (2016b, pp. 298-299) for a list of possible schemas) this could be a possibility. It is also one area where a synthesis between the work on embodied decisions and PP could be mutually beneficial, as the former is presently developing novel computational methods that may help specify architectural details, whereas the latter provides a more developed account of the importance of interoceptive information within a wider framework that unifies perception, cognition, action and now emotion. However, there is another approach, which makes use of the precision-weighting mechanisms discussed in chapter 2 that may frame PP as an example of the ‘comparison-only’ view. We turn to explore this in the next chapter.

Chapter 5

Effective Decisions in the Interactive Brain

In the previous chapter we argued that adopting an embodied approach to decision-making means blurring the boundaries between perception, cognition, action and emotion. This need not result in the conclusion that the brain is an undifferentiated, homogeneous mass of cells. There is a wide space of conceptual possibilities between this extreme, and the other extreme of a massively modular architecture. A promising approach to understanding the complex dynamics of the brain, and its interactions with the body, comes from recent network-theoretic approaches. In this chapter we will explore how the modulatory effects of precision-weighting in PP dovetail with recent work on the interactive brain, and how this leads to a novel approach for understanding decision-making. We argue that effective decision-making emerges from both the brain's ability to flexibly and rapidly alter its effective connectivity to meet the shifting demands of the environment, but also requires longer-term learning (over developmental and evolutionary timescales), and the subsequent redeployment of prior knowledge. PP has the conceptual and theoretical tools to explain this ability, but doing so requires revisiting some of the standard assumptions in cognitive psychology, and extending our explanatory scope out into the body and the world.

We begin in section 5.1 by revisiting the account of precision-weighting offered by PP, with an emphasis on its connection to neuromodulation. This will connect to a discussion in section 5.2, concerning Anderson's (2014) proposal for Neural Reuse and the Interactive Brain, and its consequences (outlined in section 5.4) for traditional cognitive psychology. We will conclude in section 5.5 by discussing what this means for our account of the mechanisms that underlie decision-making.

5.1 Balancing Expectations

Consider the case of learning to play an instrument (e.g. a guitar). During the earliest stages of learning it is likely that you will have low dexterity, and will be slow to move between certain chords due to an unfamiliarity with the positions and movements of your fingers on the strings. In these early stages, it is common to look carefully at your finger placements, while your brain adjusts to the large proprioceptive error signals that are generated by unfamiliar finger-placements and unfamiliar tensions in your muscles. In addition, as you play a chord for the first time and notice the auditory signal, you may carefully and deliberately pluck each string in sequence, in order to ensure you are not inadvertently muting a string due to clumsy finger placement. As you progress into an intermediate stage, becoming more adept with your finger-placement and the feel of the guitar, you will be able to shift your attention from the feeling of the guitar and your fingers, towards the nuances of the sound being generated. This will allow you to uncover more creative ways to play, and perhaps you will accidentally discover new augmented chords by mistakenly placing a finger on an incorrect string. The focus of your attention in these intermediate stages will have drastically shifted away from the slow, deliberate attention directed towards your body when you first picked up the guitar.

Perhaps you continue to practice and become an expert musician. If so, you may be lucky enough to experience what is known as 'flow'. This psychological state is experienced not only by musicians, but also expert athletes and other practitioners of skilled disciplines, and is commonly described as a loss of reflective self-consciousness, and a heightened immersion in the present activity. This is considerably different from the state experienced by the beginner, and one characteristic stands out: the shifting focus of attentional awareness.

In chapter 2 we discussed how attention in PP is identified with the shifting precision expectations that adjust the weight of error-signals (e.g. salience of auditory information in a noisy room). Error-signals arise when predictions are unable to account for particular sensory signals, and thus it stands to reason that part of becoming an expert in some task requires becoming more adept at predicting future states (e.g. a beginner will have little to no idea of what sound will be generated by plucking certain strings when holding an F Major chord, whereas an expert may be so familiar that a slightly out-of-tune string may stand out). This could in turn lead to a more *creative* use of precision-weighting mechanisms as the brain becomes expert at predicting sensory signals due to an increased familiarity with the situation.[1] The situation is not too dissimilar from a phenomena that is well-studied in newborn infants known as 'motor-babling'.

Motor-babbling is the execution of seemingly random movements, which allows the infant to learn about the physiological characteristics of their body through sensing the reafferent information generated by their own movements (Kilner et al., 2016).

[1] A related story is given by (Hobson and Friston, 2012; Hobson and Friston, 2014) in the case of dreaming, where exteroceptive error signals are reduced as the organism falls asleep. In the absence of error-signals, the predictions generated by the brain have no anchor to the stable structure of the external world, which Hobson and Friston claim could be a cause of the particular characteristics of dreaming.

The perceptual information that is generated by the self-produced changes, creates bidirectional associative information (Hebbian learning) between the pattern of neural activity responsible for producing the movement and the activity representing the reafferent information. Initially, this motor babbling will appear uncontrolled, but as the infant learns to associate certain movements with a sense of agency (e.g. this motion was produced by an internal cause, not by an external cause), further opportunities for self-directed actions arise. Just as the musician must familiarise herself with an instrument, an infant must become familiar with the characteristics of its body, in order to effectively interact with the environment, and in turn make adaptive decisions. This happens through repeated interactions, spread out over time. Accommodating this flexible learning in PP requires carefully balancing top-down expectations (representing prior beliefs), with the unexpected sensory signals from the world, in order to determine whether the inaccuracies result from inappropriate top-down expectations and thus represent a learning opportunity where the world can drive the updating of the brain's inner models. As learning happens over extended timescales, and is responsive to interacting nested structures in the environment, the hierarchical commitments of PP are yet again vital.

We have already seen why the hierarchical organisation of the brain is conducive to the idea that the hierarchical generative models are organised over increasing spatiotemporal scales. This structure is important to allow more complex agents such as ourselves to be receptive to similarly nested structures that exist in the world. From this perspective, we could approach the hierarchical-organisation of the brain's models in a synchronic manner, potentially reflecting on the representational content at different layers. However, this should be carried out with caution, as it overlooks an obvious but important point. Like the world, the brain is constantly in flux, flexibly altering its patterns of *effective connectivity* over short timescales, and more slowly adjusting its structural connectivity and morphology over developmental and evolutionary timescales.[2] This is not to deny the need for some robustness and stability in the brain's inner models—too little robustness or stability is equally maladaptive, as it prevents an organism from relying on the deployment of prior knowledge to its own advantage—but an uncertain and changing world is not always a congenial environment for an entirely inflexible system. Finding the right balance between utilising prior knowledge and seeking new opportunities to learn, seems to be key in making effective decisions. In the next section we will look at some of the mechanisms that support this picture.

5.1.1 Neuromodulation and Effective Connectivity

PP is considered by many to be a functional-level description, and therefore can be considered independent of implementational details (Hohwy, 2015; Spratling, 2013). However, some have nevertheless proposed specific mechanisms for certain components. Here we focus on precision-weighting.

A number of claims have been made regarding the mechanisms that support the

[2] A brief note on terminology: we can distinguish between three types of neural connectivity, which are known as structural, effective and functional. Structural connectivity is the most common, and refers to the gross anatomical connections that exist between neural cells allowing them to interact and communicate (perhaps in conjunction with extra-synaptic mechanisms). Functional and effective connectivity refer to the activity that is estimated by neuroimaging techniques such as fMRI or EEG and does not necessarily identify a complete chain of anatomical connections. Functional connectivity is defined as the temporal correlation between two regions given some task, but does not provide any information concerning the directionality or causality between the regions. However, effective connectivity uses models of neural interactions to infer directionality. It is commonly understood as the influence that one neural system exerts over another, possibly through the use of extra-synaptic mechanisms (Friston, 2011b). Effective connectivity can be achieved without the need for extensive rewiring or physiological changes, and is thus a suitable candidate for transient-assembly of distributed systems.

precision-weighting story in PP.

Firstly, it has been claimed that the mechanisms behind precision-weighting involve altering the post-synaptic gain on prediction-error units through key neuromodulators such as dopamine (Friston et al., 2014). They may also provide a way to reconcile the competing effects of signal suppression and signal enhancement (Clark, 2016b). Some have even gone as far as singling out specific types of neural cells (i.e. pyramidal cells) on the basis of salient characteristics that suggest key roles in neuromodulation (Phillips, Clark, and Silverstein, 2015).[3]

Secondly, as well as providing a way of balancing the influence between top-down and bottom-up signals, it has been argued that the mechanisms behind precision-weighting could provide a means of altering the brain's effective connectivity (cf. Clark, 2013b). The potential for these neuromodulatory mechanisms to exert wider influence by changing the brain's effective connectivity is of crucial importance for an embodied PP account of decision-making.

As we saw in chapter 3, recent work in decision theory has argued in favour of a distributed systems approach for encoding decision related measures of value. Therefore, the brain needs to have a way of integrating these disparate sources, without having to integrate their signals into one central region that can deliberate over a unified, abstract representation. Neuromodulators seem to be an effective means for achieving this. Dayan (2012, p. 241) provides a comprehensive review of the relevant properties that neuromodulators have in terms of organisation and effects (a selection of these are presented):

1. Neuromodulatory systems can report selective information.

[3] Others have argued that these types of cells could form the basis of cortical microcircuits (i.e a local neural population that represents something like a basic wiring diagram), which play a fundamental role in predictive coding (Bastos et al., 2012).

2. Reporting can take place over multiple timescales (including very quick timescales).

3. Different receptor types can respond selectively to separate characteristics of the signal, and can be localised on anatomically different pathways.

4. Interactions among different neuromodulators are very widespread.

5. Neuromodulatory signals can be turned to different uses.

6. Neuromodulators can influence the course of activity by regulating which of a number of gross pathways determine the activity of neurons.

7. Neuromodulators affect plasticity over many time scales.

8. Neuromodulators are involved in the regulation of energy utilisation in the brain and body.

9. Individual differences in neuromodulatory receptors or transporters have observable effects on decision-making behaviour.

1, 6 and 7 provide indirect support for the claim that precision-weighting in PP acts as a gating mechanism, modulating the influence that error signals have on higher-levels. As Dayan (ibid., p. 251) states:

> “Neuromodulators both broadcast and narrowcast key information about the current character of the organism and its environment, and exert dramatic effects on processing by changing the dynamical properties of neurons.”

Furthermore, 2, 3 and 7 corroborate the claims regarding hierarchical organisation in the brain. The remaining points will be discussed in later sections.

In addition to these claims, Ouden et al. (2010) found evidence that striatal prediction errors play a modulatory role on the large-scale coupling between distinct visuomotor regions, and Cocchi et al. (2013) explored the context-dependent, transient changes in patterns of cooperation and competition between control systems, which result from higher-level cognitive control. If correct, this may allow PP, with its emphasis on bi-directional influences, to accommodate the ubiquitous effects of attentional modulation that Cisek and Pastor-Bernier discussed (chapter 3), but which remained underdeveloped in their account.

Furthermore, Feldman and Friston (2010) explore the neuromodulatory role of the cholinergic system as one possible source of precision-weighting, and more recently, Friston et al. (2012) investigated the neuromodulatory role of the dopaminergic system, with a specific focus on decision-making and reinforcement learning. They argue that dopamine controls the precision of incoming sensory inputs by balancing the respective weight of top-down and bottom-up signals during active inference. This balancing means, crucially, that the predictions that drive action, also determine the context in which the movements are made. It also provides further support for the blurring of perception and action in PP, and the important role that neuromodulation plays in uncovering this relationship.

Given the above, it is clear that the influence that neuromodulation has on the functional attributes of local regions should not be downplayed. However, the effect of neuromodulation adds both an astonishing depth to the possible dynamic functions that local networks can perform, but at the same time renders the pursuit of a simple wiring diagram (or connectome) for the brain as a somewhat naive pursuit on its own (Anderson, 2014). One concern is that the effects of certain neurotransmitters can propagate widely through extra-synaptic channels, influencing regions that are distant from the initial point of release (Agnati et al., 2010; Park and Friston, 2013). Although this means that neuromodulatory mechanisms provide a way for the brain to influence distal regions, and thus assemble distributed systems that can flexibly adapt to shifting task-demands, it also means that determining which population a particular system is communicating with becomes particularly challenging. Postsynaptic cells are not able to distinguish between various types of presynaptic activation of inhibition, despite the potentially different effects that may be induced by the transmitted information (Dayan, 2012). This is why a relatively fixed anatomical structure makes sense, as different receptor sites (responsive to different neuromodulators) can be restricted to certain regions of the brain.

However, although it complicates matters, this need not preclude the possibility of functional diversity in local regions. Consider the following example from Price and Friston (2005, p. 268) regarding the structure-function mapping for the finger:

> "At one level, the forefinger can be attributed multiple and diverse functions including 'piano playing,' 'typing,' 'scratching,' 'pinching,' 'feeding.' At a second level, these functions could be classified so as to distinguish them from those the forefinger cannot perform—such as 'digestion,' 'thinking,' or 'walking.' [...] At a third level of description, however, the forefinger can only do one thing—'bend' and 'straighten.' Its role in other tasks is therefore entirely dependent on what the other fingers and thumbs are doing and what environmental context they are in."

In an analogous manner, the anatomical and physiological structure of the brain constrains neural communication, whilst neuromodulation alters which functions certain populations can participate in. We will return to this point shortly.

Strong evidence exists that neuromodulators are able to rapidly alter effective connectivity, to allow the brain to flexibly and transiently recruit networks of distributed systems in response to changing task demands from the environment. In the case of decision-making, this becomes increasingly important when we consider that the utility of certain actions will necessarily be tied to the current state of the organism (e.g. food is more valuable to a hungry animal than water is). The brain needs to be able to adjust the weighting of certain key regions involved in decision-making in response not only to incoming interoceptive information, but also with information corresponding to higher-level knowledge (e.g. area x has a high-chance of encountering a predator based on prior experience). Some have argued that these transiently assembled networks are the general rule for the brain (Anderson, 2014). Acknowledging this requires revisiting some of the cognitivist assumptions that have dominated the cognitive sciences, and subsequently moving towards an embodied account of cognition and behaviour.

5.2 The Interactive Brain

> "[...] exactly what sort of metaphysical stance would lead one to suppose that something as versatile as a knife blade has anything like a "fundamental function"? It has some fundamental physical characteristics that make it useful in a variety of circumstances. Knowing what those characteristics are is surely useful, but to search for the functional essence of a knife is to be in the grip of a deep philosophical, ontological error."
> (ibid., p. xix)

It may appear as though the above quote is setting up a straw-man. Surely no one would disagree that a knife has a number of different functions that are dependent partly on the usage and intentions of its wielder, and partly constrained by the physical characteristics of the object itself (e.g. used for cutting or as a makeshift screwdriver)? Perhaps not for a knife, but the crucial point in Michael Anderson's recent book *After Phrenology* is that this error has been made repeatedly in the cognitive sciences. As a result of failing to rid itself of several implicit assumptions regarding the functional architecture of the mind, the cognitive sciences have continued to emphasise functional decomposition and localisation as guiding methodological principles. He argues that these assumptions are impediments to the progress of the cognitive sciences, and need to be reconsidered. In their place, Anderson defends a theory he calls ‘neural reuse’. This theory treats the brain as a dynamical system, which has evolved for the purpose of controlling an organism’s interactions in its environment.

This theory echoes the view outlined in the previous section, whereby the brain responds to the task demands on the agent by fluidly altering its effective connectivity in the short-term, and adjusting its structural connectivity in the longer-term.

5.2.1 Neural Reuse

That nature finds new uses for old tricks should be beyond dispute, and the brain is certainly no exception. The structures of the brain, Anderson (2014) argues, evolved within certain efficiency constraints, and where possible redeployed existing capacities, rather than developing new structures *de novo.* Though initially established for one purpose, the theory of *neural reuse* states that existing circuits can be exapted (reused) to acquire new uses, often *without loss* of their original function and without the need for unusual circumstances such as injury. As we will see, this leads to a number of challenges for traditional cognitive psychology, which understands the functional architecture of the brain as composed of interacting modules, and attempts to individuate these components through the processes of *functional decomposition* and *localisation.*

Functional decomposition has a long history in philosophy of mind, but contemporary approaches in the cognitive sciences are largely dominated by the relatively more recent assumptions of faculty psychology—the view that the mind is composed of interacting modules (or faculties) that are recruited to perform certain tasks (Fodor, 1983). Each module plays a specific functional role, and the task of the cognitive sciences is to determine what this role is and which states of the brain are responsible for implementing it. For example, one may start by identifying some phenomenon of interest (e.g. some overt behaviour), determine which states of the brain correlate with the behaviour through the use of neuroimaging (or perhaps lesion studies if concerned with the loss of some behaviour), and then attempt to map a hypothesised function onto these structures in line with any additional theoretical constraints. These constraints may be based on the assumption that cognition is adaptive in some manner, as assumed by rational analyses (Anderson, 1991) and evolutionary psychology (Barkow, Cosmides, and Tooby, 1995), or on considerations of whether the functional architecture is neurally-plausible (Eliasmith, 2013).

Both Poldrack (2010) and Anderson (2014) worry that cognitive psychology (and to a lesser extent cognitive neuroscience) has been dominated by a limited taxonomy of cognitive and mental functions; a taxonomy that they claim has been inherited from faculty psychology, and is a hindrance to uncovering the real functional organisation of the brain. To rectify this, Anderson starts from the premise that brains initially evolved for the purpose of controlling action in our earlier environments, and that we should expect to see traces of the repurposing of these pre-existing neural structures in our phylogenetic history. More recent capacities (e.g. abstract reasoning or mathematics) would have had to find their neural niche within the constraints imposed by this control architecture. His view parts ways with the evolutionary psychologist in the expectation of what these traces will look like:

> "[...] whenever possible neural, behavioral, and environmental resources should have been reused and redeployed in support of any newly emerging cognitive capacities. Functionally autonomous and dedicated neural modules just do not seem to make good design sense given the importance of efficient use of available resources and of ongoing interactions in shaping function." (Anderson, 2014, p. 7)

Anderson's approach acknowledges the fact, well established in neuroscience, that regions of the brain are functionally differentiated. However, he notes that functional *differentiation* is a conceptually distinct claim from that of functional *specialisation*. This latter view states that each region of the brain expresses a specialised function, which implements a single cognitive operation. However, Anderson's principle of neural reuse, as we will see shortly, is committed to *functional differentiation* without *functional specialisation*—we can differentiate one region from another, but not attribute to it a single cognitive operation.

So what are the specific claims made by this principle, and what is the evidence for each of them? Given the limitations of individual neuroimaging studies, and the murky window into the mind that they provide, researchers have turned to interpreting these studies collectively, using computational methods to analyse and determine patterns in the wealth of data collected in recent decades. Anderson points to three studies, which in turn support a number of predictions of neural reuse. Each of the studies involves some type of meta-analysis, performed using a database such as BrainMap.org. This database publishes functional and structural neuroimaging experiments, with coordinate-based results, which allow users to perform statistical analyses over a wide-range of experimental studies, rather than being limited to one. Arguably, this provides a more robust conclusion—subject to the application of appropriate statistical techniques—than the report of a single neuroimaging study.

The predictions that Anderson focuses on are:

1. Individual brain regions should support numerous cognitive functions across diverse tasks (e.g. classification or working memory).

2. Functional differences should be reflected less in what neural regions are implicated, and more in the different patterns of interaction between similar elements.

3. Newly evolved cognitive functions or behaviours (e.g. language) should be supported by a greater number of structures.

In support of the first claim, Anderson and Pessoa (2011) performed a meta-analysis of the functional diversity of 78 anatomical regions of the brain by determining whether (and how often) each was active in 1,138 experimental tasks across 11 task categories (e.g. emotion, reasoning, working memory etc.). As Figure 5.1 shows, different regions of the brain display a greater functional diversity than others, with subcortical regions having the lowest overall average functional diversity. To demonstrate this, the authors used a measure of diversity variability (DV) that was based on standard deviation:

$$DV = \sqrt{\frac{\sum_{i=1}^{k}(Cat_i - mean)^2}{k}} \tag{5.1}$$

Here Cat_i refers to the proportion of activations in each task category, *mean* refers to the average proportion, which is always 0.091 with 11 categories, and k equals the number of categories. Diversity was normalised so that the values range from 0 (all activations in one category) to 1 (activations spread equally across all 11 categories).[4] The overwhelming finding from their study is that functional diversity appears to be a genuine feature of local brain organisation, with the overall average diversity of cortical regions placed at 0.71 and subcortical regions at 0.63.

Anderson and Penner-Wilger (2012) performed a similar kind of meta-analysis in support of the second claim, but this time were interested in measuring the *functional connectivity* of key regions of the brain, based on an analysis of functional coactivation. A functional coactivation analysis determines how often multiple, spatially separated regions of the brain *coactivate* under certain task conditions (i.e their func-

[4] See (Anderson and Pessoa, 2011) for further details, including an index of the regions and task categories explored.

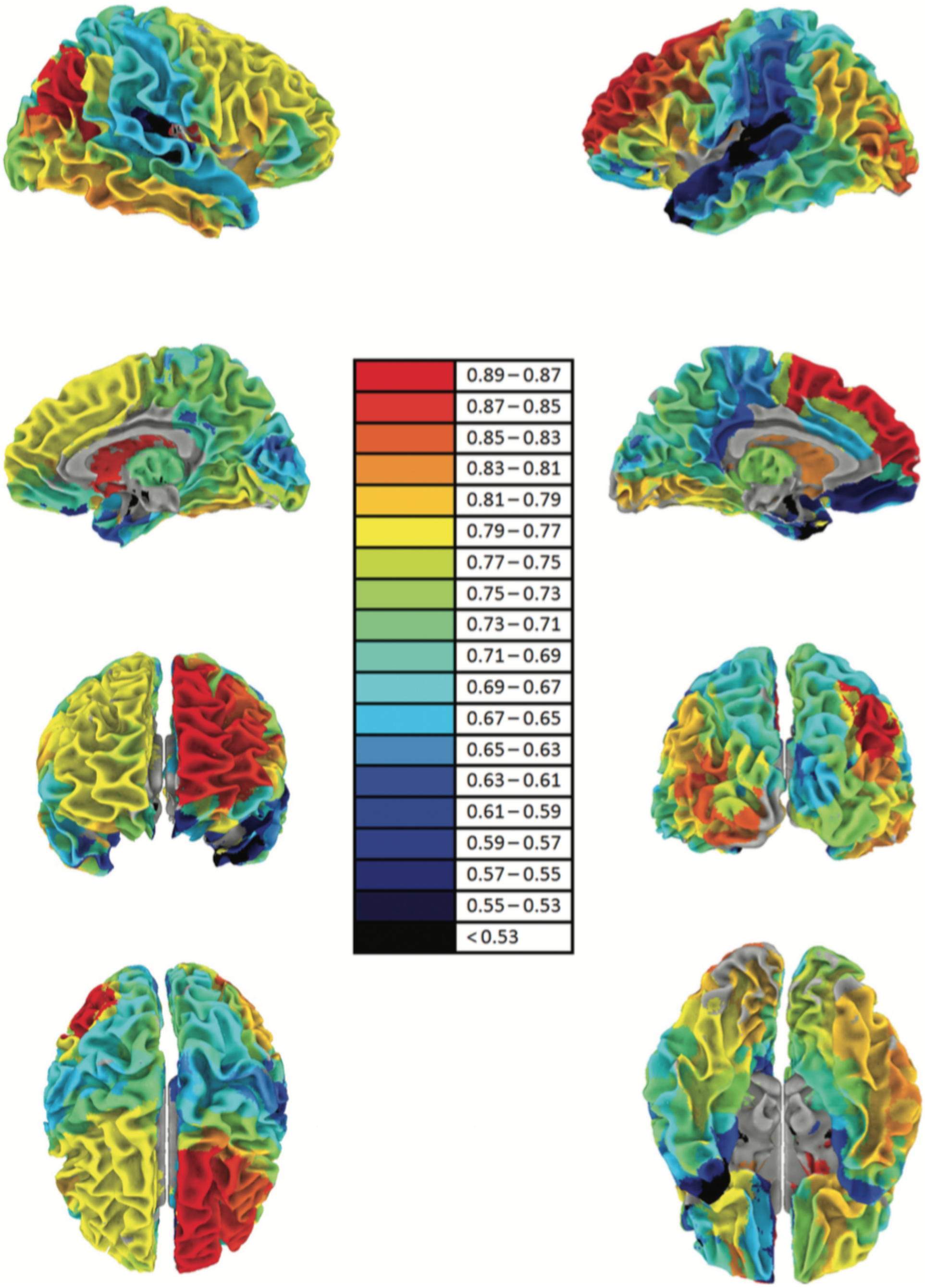


Figure 5.1: Task diversity of brain regions (grey indicates no information). The normalised values range from 0 (all activations in one category) to 1 (activations spread equally across all 11 categories). Reprinted from (Anderson, 2014, p. 11).

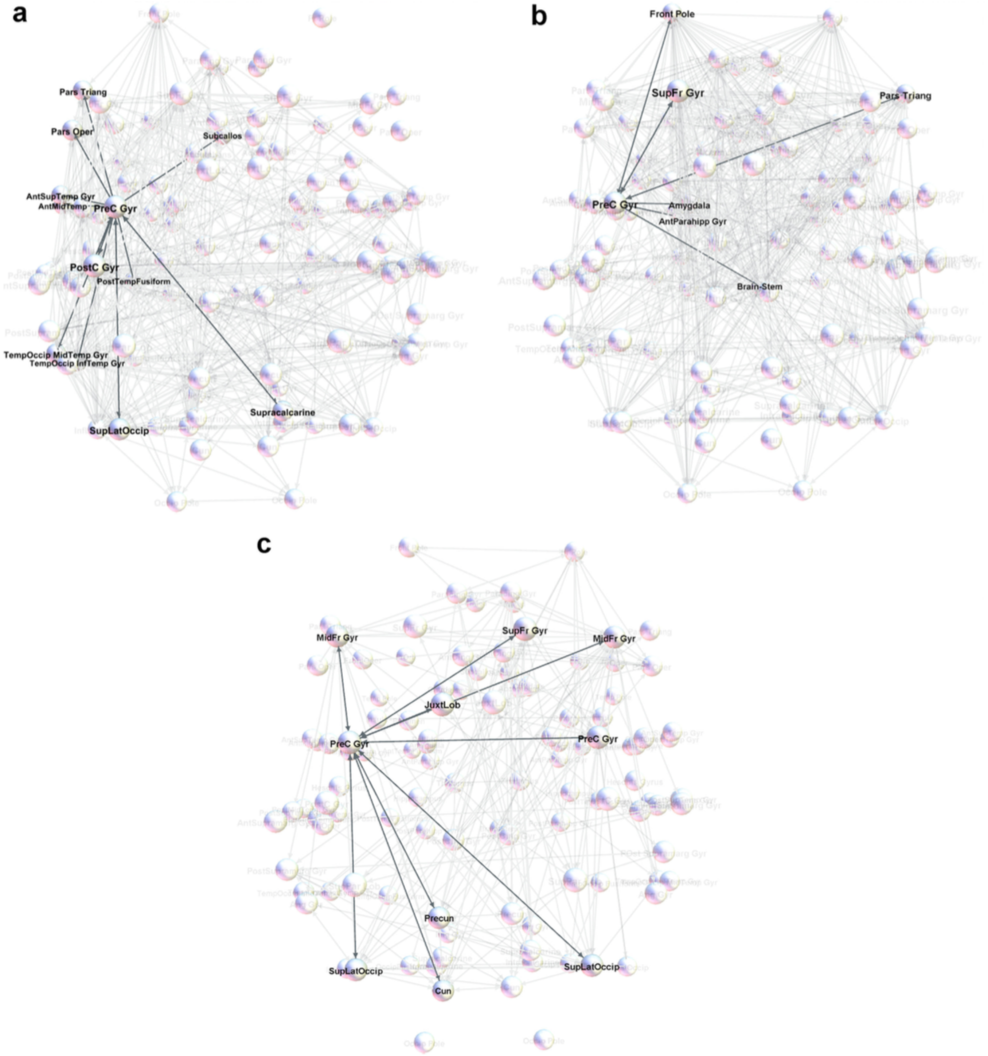


Figure 5.2: Functional connectivity graphs for left precentral gyrus under three different taks: (a) semantic, (b) emotion, (c) attention. Reprinted from (Anderson and Penner-Wilger, 2012, p. 45).

tional connectivity). Anderson and Penner-Wilger (ibid.) claim that if the regions are simultaneously active more often than on their own, during some task, then this indicates that there is a functional connection between the regions.

Developing on the first prediction, they found that as well as a particular region being active across diverse tasks, the functional connectivity of this region with neighbouring regions was also likely to be varied across tasks. For example, Figure 5.2 depicts a number of graphs showing the functional connectivity of Left Precentral Gyrus under semantic, emotional and attentional conditions. As the graphs demonstrate, the edges which are most active in each task (thick lines) are connected to different nodes (representing other neighbouring regions). As we will see shortly, this means that trying to map a function onto a particular structure of the brain becomes increasingly difficult. Not only do regions of the brain have a highly diverse functional profile, but what functional role a region is currently performing is determined less by an intrinsic property of the region, and rather by which local network it is functionally connected to. In short, the function of a region cannot be determined solely on the basis of a localised analysis, as it neglects the constitutive role that neighbouring regions play in determining its functional role.

Regarding the final claim, Anderson (2014) argues that newly evolved behaviours should be supported by a larger number of structures, on the assumption that the later a function emerges in evolution or development, the more potentially useful existing elements there will be to exploit. Taking language as a prototypical instance of a recently developed function, Anderson (2008) explored a wide-range of fMRI studies and found that language functions are on average more widely scattered in the cortex than both attention and visual perception tasks. Given this, it is unsurprising that distributed regions of the cerebral cortex implicated in the semantic processing of language (collectively known as the 'semantic system') have proven so difficult to map, and why the semantic selectivity of these regions is still unknown (cf. Huth et al., 2016). However, this need not lead us to deny the *relative* specialisation of some regions. Those regions at the most peripheral parts of the nervous system are likely to have a more limited range of possible states due to the decreased number of connections to neighbouring regions.

Given these studies, what is the positive proposal of Anderson's framework? To begin, he argues that a fundamental property of the brain is its ability to *self-organise*, by locating and assembling the appropriate coalition of neural circuits that will allow the organism to deal most effectively with the changing demands of an uncertain environment. Where a local neural circuit can be redeployed to fulfil a particular role, Anderson argues, a number of mechanisms exist to search for, and in turn recruit, a relevant subsystem of the brain. These mechanisms exist to change the *effective connectivity* of the networks in the brain by transiently assembling the set of local circuits that have the appropriate functional biases (i.e. the possible functional roles instantiated by the local network when effectively connected to one or more of its possible neighbours) to collectively respond to the task at hand.

To highlight the importance of these local, effectively connected neural networks Anderson utilises the acronym TALoNS, which stands for Transiently Assembled Local Neural Subsystems. He states:

> "TALoNS are the temporary, reproducibly-assembled functional parts (large and small-scale networks and other elements) of the brain. TALoNS have intrinsic causal properties or dispositions determined by their internal structure and effective connectivity, but their functional selectivity emerges from the way these dispositions are constrained by the other functional structures with which they interact." (Anderson, in press, p.10)

TALoNS, and the underlying mechanisms that form them, provide us with important clues for understanding what the brain is responding to in the environment, and why an embodied perspective is most suited for the task. Before we turn to these latter points, it is worth exploring some proposals for the underlying mechanisms.

5.2.2 TALoNS: Some Proposed Mechanisms

Increasing popular awareness of large-scale projects such as the Human Connectome Project[5] has led to a misconception that understanding the brain is simply a matter of determining its 'wiring diagram' (i.e. obtaining a map of the structural and functional details of all the neural connections in the human brain). This would require a translation of how neurons decode, transform and re-encode signals, by investigating the spiking patterns of different regions (Eliasmith and Anderson, 2003). It is assumed that once achieved we will understand how the brain works. But will we?

As a tentative proposal for the mechanisms that are responsible for the brain's ability to recruit the relevant TALoNS, Anderson (2014) points to several "extraconnectomic contributors". Each of these possible contributors, should their empirical validity be cemented, would likely fulfil an integral cognitive role in the processing of information in Anderson's framework, and indeed in PP, which also makes regular reference to their importance in precision-weighting. Importantly, their dynamic nature challenges the cognitivist assumption that the mechanisms underlying cognition must be encapsulated not only between perception and action, but also within the boundaries of the brain.

Volume Transmission

Volume transmission (VT) is a type of signal diffusion that takes place within the brain's extracellular fluid. It refers to the diffusion of neurotransmitters that cause

[5] http://www.humanconnectomeproject.org/

activation of extrasynaptic receptors, which are remote from the initial point of release from the neurotransmitter system (Agnati et al., 2010). As such VT is intertwined with other bodily systems such as the endocrine system, and enteric nervous system.

VT is contrasted with wired transmission (WT), where the communication channel has well-delimited physical boundaries (i.e. axons, synapses and gap junctions). Furthermore, unlike the relatively rapid and precise signalling of synaptic transmission, volume transmission is considerably *slower*, and is thus more suited to modulatory or tuning functions. As such, it may provide a separate mechanism, operating at a different spatio-temproal scale to WT, which could play the necessary modulatory role required to transiently assemble local neural subsystems. As PP makes reference to a multi-level, hierarchically-organised architecture that is structured according to a spatiotemporal scale, further work on the dynamic integration of these mechanisms is important.

Due to the difficulty of modelling a brain the size of a human's, Bargmann (2012) studied detailed neural circuits in crustaceans, *C. elegans*, and *Drosophila*, revealing the ability of neuromodulators, in combination with sensory context, to reconfigure information processing by changing the composition and effective connectivity of functional circuits. Bargmann argues that these studies support the claim that information flow through local neural circuits is partially determined by neuromodulatory states—an important component of volume transmission.[6]

[6] Agnati et al. (2010) provides a more detailed overview of some of the mechanisms believed to be involved in these processes.

Neuron-Glia Interactions

Another important mechanism involved in VT is neuron-glia interaction. Glial cells (or neuroglia) are non-neuronal cells, which have long been considered to play an ancillary role in supporting neurons. For example, they surround neurons to provide structural support, and play a role in supplying nutrients and oxygen to neurons. However, the role of these purportedly "housekeeping cells" is being questioned. Referring to glial cells as the "other brain", Fields (2009) argues that they may play an important role in VT, as well as possibly holding the key to answering some difficult questions surrounding the medical treatment of neurological disorders such as dementia and schizophrenia. As all glial communication is extrasynaptic and chemical in nature, Anderson (2014, p. 78) argues that it provides, "[...] an independent, complementary [chemical] network for information flow in the brain. Glia are also thought to regulate the formation of synapses, modulate learning mechanisms such as long-term potentiation, and regulate synaptic transmission because they both manage the clearance of neurotransmitters from the synaptic cleft and also release their own neuromodulatory substances. None of this crucial interaction is captured by connectomics."

Weak Endogenous Electrical Field

Though not explicitly identified by Anderson, we can also add a further potential candidate to the list—the weak endogenous electrical field. Qiu, Shivacharan, and Zhang (2015) observed a group of neural waves that share the same speed as standard synaptic transmission ($\sim$0.1 m/s), and which persisted after the relevant synapses and gap junctions were blocked. The authors argue that the only remaining explanation is an endogenous electrical field effect. As it is traditionally assumed that the brain's endogenous electrical fields are too weak to propagate wave transmission, their study seems to challenge this notion, instead supporting the claim that neural signals can propagate by means other than synaptic transmission, gap junctions, or diffusion (i.e. a non-synaptic governing mechanism). The implication of such a finding is, they claim, that such directed electrical fields can be used to interact with other cognitive processes, which may help regulate a variety of processes in the brain. As with the previous two elements, this work demonstrates a need for going beyond the connectome.

It is important to emphasise that each of these elements is only partially understood at present, and their joint contribution to the global dynamics of the brain even less so. However, in spite of this, that they are contributing to the dynamic interactions of the brain (i.e. as search mechanisms or neuromodulators) is fairly well supported. Furthermore, acknowledging that VT operates at a different spatio-temporal scale to WT is important in recognising the multi-level, recurrent nature of neural processing.[7] Given that Anderson wishes to argue that the brain evolved primarily to control the *situated* action of an organism, this multi-level, reccurent nature of the brain will be fundamental, given that actions also unfold in the world across a range of spatio-temporal scales (see chapter 6).

However, each of these elements only provides a potential mechanism that may or may not function as a vehicle itself, or as part of a larger vehicle of communication in the brain. Some, including Anderson, have recently argued that trying to determine the answer to these sorts of problems requires moving away from viewing the brain's architecture as composed of interacting modules or regions, and instead viewing it from a network perspective (Pessoa, 2014; Sporns, 2011).

[7] Unlike in standard usage, where it means 'occurring repeatedly', in neuroanatomy, the term 'recurrent' refers to the direction of a nerve's signal 'turning back in an opposite direction'.

5.3 From Regions to Networks: Pluripotency and Degeneracy

The shift in emphasis—from regions to networks—may not seem particularly radical at first. Some may argue that a similar shift in emphasis has already occurred in the move from classical computational theories of cognition to connectionist networks. The latter emphasised parallel processing of information, and took the vehicles of computation to be distributed across a network. However, these similarities are insufficient to draw a parallel with the notion of TALoNS discussed in the previous section. The reason for this is that TALoNS are *transiently assembled*, and indicative of potentially one among many of the possible functions that a local neural region can instantiate when functionally connected to the relevant neural partners.

Recall that neural reuse is committed to the view that the function of a region is determined by its functional and effective connectivity to neighbouring regions—until then the region is merely disposed to perform one of potentially many functions. If local neural circuits support a number of tasks across different domains, they must also retain a more complex response profile (i.e. a probabilistic representation specifying the parameters for the range of conditions under which a neuron (or neural ensemble) responds). This is incompatible with a strict specialisation view defended by nativism or modularity. However, incompatibility with a position does not necessarily entail the adoption of the diametrically opposed perspective, which in this instance would be something like the brain as an undifferentiated, homogeneous mass of tissue.

Instead, neural reuse is compatible with the idea that during ontogenetic development local regions will come to possess a range of distinctive response profiles. Their profiles could be determined by local cortical biases (e.g proximity to peripheral regions with highly constrained response profiles), as well as factors such as learning and experience, which are themselves shaped by internal factors (e.g. interaction and recurrent co-activation with other regions) and external factors (e.g. socio-cultural constraints). Though it may be valid to construe this as a form of specialisation, Anderson (2014) argues it is a form that is far removed from the sort offered by traditional cognitive psychology and philosophy of mind, whereby certain regions come to specialise in tasks such as "face perception", or detect *organism-independent* properties of the world. The primary issue here is to determine how we should attempt to map cognitive (or psychological) functions onto neural structures such as TALoNS. If a region possesses multiple, dispositional functions, then isolating a single region as the object of interest is unlikely to provide any useful constraints for a function to structure mapping because of its functional diversity. We can refer to this feature of a region's functional capacity with the label 'pluripotency'.

Pluripotency refers to a structure-function mapping relation where a particular structure performs multiple functions—a one-to-many relation (Price and Friston, 2005). Determining the functions that a structure realises is typically investigated by neuropsychology, occasionally using transient lesion techniques such as transmagnetic stimulation (TMS)[8]. However, this structure-function mapping relation can be reversed, and investigated using functional neuroimaging techniques that determine which structures are sufficient for a given function. Such techniques often uncover an alternative one-to-many relation between a function and structures. We can call this relation 'degeneracy'.

Degeneracy refers to the capacity for different regions to carry out the same function (e.g. when a brain area is damaged or disabled (Price and Friston, 2002)). The fact that both of these properties are evident in the brain presents a methodological

[8] TMS is a noninvasive procedure that uses magnetic fields to stimulate neurons in the brain. During stimulation, normal ongoing brain activity is disrupted by the magnetic current, and as such TMS creates a transient period of brain disruption known as a 'virtual lesion'.

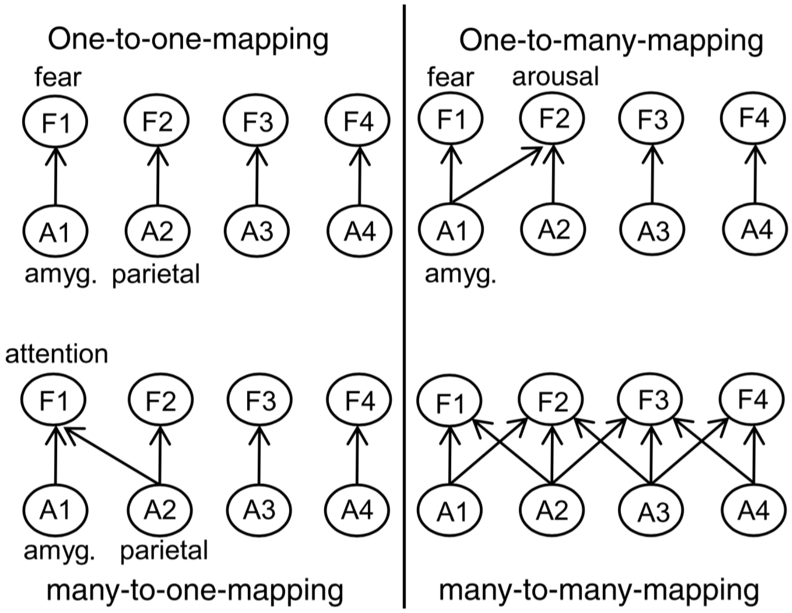


Figure 5.3: A schematic of the possible structure-function mappings in the brain. Because the brain exhibits both degeneracy and pluripotency (Price and Friston, 2002; Price and Friston, 2005), the mapping is *many* to *many*. This requires a drastic reinterpretation of cognitive frameworks, moving from regional specialisation to functional differentiation, where the basic units of study are networks. Abbreviations: A1, . . . , A4: areas 1 to 4; amyg: amygdala; F1, . . . , F4: functions 1 to 4. Reprinted from (Pessoa, 2013, p. 194).

challenge to classical cognitive psychology, as well as the pursuit of a well-defined taxonomy of cognitive states. This is because, taken together, pluripotency and degeneracy represent a possible *many-to-many* relation between structures in the brain and cognitive functions, as shown in Figure 5.3.

One may argue that the possibility of any function-structure mapping function depends on first determining an appropriate level of description. For example, given an abstract enough level of description, one may claim that an anatomical region such as the amygdala can be associated with emotional processing, rather than a particular emotional response (e.g. fear or arousal). Of course, for this to be useful, the level of abstractness is going to have to be relatively constrained to rule out simply labelling a region with the function 'cognitive processing'. Instead of trying to deal with these tricky conceptual worries, Pessoa (2014) argues that traditional anatomical regions are simply the wrong unit of description to explain how the brain's structures are linked to functions.

Returning to the notion of functional connectivity introduced in section 5.2.1, when the functional connectivity between two regions is high, the degree to which we can isolate them from one another becomes increasingly challenging. The regions become increasingly coupled such that they stop acting as isolated components. When this happens, we are forced to consider the interacting regions as a single system, which is *non-decomposable*. At the other end of the continuum is something like a module, which is fully *decomposable* and operates according to its own intrinsic properties. In between is a continuum of possible organisations that collectively represents a possibility space for the brain's architectural organisation (ibid.). Where particular regions of the brain fall on this continuum is partly an empirical question, which requires further investigation.

Anderson (2014) provides compelling reasons to believe that the majority of the brain is going to be organised in a non-decomposable manner due to the transient flexibility of neural regions, and the ubiquity of neural reuse throughout the brain. If transiently-assembled, non-decomposable systems, exhibiting both pluripotency and degeneracy are indeed as ubiquitous as the above accounts suggest, then we seem to require an alternative way of individuating structures in the brain. Anderson (2014), Pessoa (2014), and Sporns (2011), have suggested that we move away from localisable brain regions as the object of interest in function to structure mappings, and instead take networks to be the relevant objects of interest. Each network is anchored in some set of regions, but is not localisable to any particular region with a distinct functional profile. Furthermore, multiple networks may overlap, such that a single region is employed to fulfil different roles given the network it is currently a part of—echoing the findings of Anderson discussed in section 5.2.1.

The network approach is indicative of a recent trend in the natural and social sciences, which reflects a shift in how we understand the behaviour of complex systems. To understand these systems, as was ilustrated in the case of Rayleigh-Bénard convection in chapter 1, we need to have knowledge of how the lower-level components interact, as well as the emergent properties that may result from these interactions. Knowledge of the properties of the components is insufficient on its own.

Complex systems display characteristic, ordered patterns of collective behaviour (hence the use of the terms collective variable or order parameter). By adopting a network approach, scientists can gain important insight into the means by which the lower-level components of a system self-organise into ordered patterns. This is because unlike the consideration of individual components in isolation, a network approach is by definition interested in the webs of connectivity that structure the components of the network under investigation. Developments in computational modelling and statistical techniques (e.g. graph theory), have empowered researchers to discover new forms of connectivity in the nested, hierarchical-structure of networks, and uncover new methods of understanding the emergence of more structured behaviours. As Sporns (2011, p. 2) states:

> "In multiscale systems, levels do not operate in isolation—instead, patterns at each level critically depend on processes unfolding on both lower and higher levels. The brain is a case in point. We cannot fully understand brain function unless we approach the brain on multiple scales, by identifying the networks that bind cells into coherent populations, organize cell groups into functional brain regions, integrate regions into systems, and link brain and body in a complete organism."

As the tentative mechanism proposals by Anderson and others intimates, the multiple networks of the brain are deeply interconnected, but not strictly isolated to the brain. Rather, these networks depend intimately on their dynamic coupling to the body, and the ongoing interaction that the organism as a whole has with the environment through continual cycles of action and perception. By contributing to the behaviour of the organism, these brain-body networks partially structure the incoming sensory information (any action leads to new perceptions whether external or internal), and in turn modulate the internal dynamics of the system. In this sense, the brain-body system can be seen as dynamic and self-organising.

Unfortunately, advocates of the network approach are careful to point out that the challenges posed by the many-to-many mapping between regions and functions do not simply disappear when we move to a network perspective (Pessoa, 2014). So what do they propose instead?

5.4 Towards a New Taxonomy

> "[...] the project [of revising the taxonomy of the cognitive sciences] is manifestly not aimed at the wholesale expression of the theorems of

> psychology in the low-level language of neuroscience. No one in this conversation cares to reduce pain to c-fiber firing, or cognitive processes to electrochemical ones [...] the identification of a set of brain-friendly psychological primitives could make the possibility of psychoneural reduction more plausible, but whether and how such reduction might occur—and whether and how such categories would facilitate it—is a largely orthogonal debate." (Anderson, 2015, p. 70)

One of the aims for a truly naturalistic science of mind and behaviour is the pursuit of the brain's native taxonomy—a description of the architecture by which it interprets and acts in the world. For those who dismiss calls for the autonomy of psychology from more fundamental disciplines such as neuroscience (e.g. Fodor, 1983), this taxonomy should ideally be applicable to researchers across the cognitive sciences. However, calls for revision are often met with resistance, often on the mistaken assumption that it involves some sort of eliminative reduction of cognitive psychology to neuroscience. The above quote should remove any worries that this is what is being argued for, though we follow Anderson in acknowledging that this is a question that deserves a proper treatment in its own right. Nevertheless, given the claims in the previous two sections, it is unlikely that the current taxonomy of cognitive psychology will escape unscathed.

Exposing anomalous data is often helpful in challenging orthodoxy. For instance, there is evidence that the visual word form area responds not only when words are viewed, but also when they are a) heard and read in Braille, and b) responds to other kinds of visual objects as well (Price and Friston, 2005). This provides initial reasons for casting aspersions on the empirical adequacy of traditional cognitive psychology. However, even when combined with the earlier theory of neural reuse, it is unlikely to be sufficient for replacing the orthodox picture. What is needed is a positive proposal to supplant the current framework.

5.4.1 NRP Factors

Anderson's positive proposal begins with a concern about the prospects of cognitive psychology. What is measured in neuroimaging studies (especially non-invasive human studies) is often a *mixed-signal*, consisting of the activity of neural ensembles. When this is combined with the ubiquitous functional diversity of neural regions outlined in the earlier studies, it leads to a serious challenge for cognitive neuroscience and psychology. To highlight why this is problematic, consider the problem of trying to determine the variety of sources that contribute to an audio recording from a busy public space (e.g. a train station). If you were to listen to a single-track recording of the many mixed signals present in the environment, you may be able to discern some characteristics of the individual sources, but there would also be a significant amount of ambiguity that would prevent you from decomposing the signal into its well-delineated parts. For example, think of the ambiguity that may result from a recording taken during rain or wind. Additionally, think of what the measure of something more abstract like the value of an economic good represents. Even the value of something relatively simple (e.g. a pen) can be considered a product of multiple interacting factors such as economic supply and demand, perceived worth (i.e. a Mont Blanc pen versus a Bic Biro), and the value of the materials used in production. These challenges are analogous to the sorts of problems faced in interpreting the measurements taken from neuroimaging studies. In the current context, it is akin to asking what are the relevant psychological factors that contributed to the recorded signal?

Although the field of neuroscience has developed a number of impressive technologies and algorithms for disentangling these mixed-signals (i.e. independent component analysis), Anderson (2014, p. 129) points out that this does not guarantee that the *physically* unmixed signals are not in fact *psychological* mixtures. Put simply, there is no guarantee that the taxonomic categories picked out by cognitive neuroscience and psychology are neatly realised in the functional traits of these signals. Anderson's claim is that psychological states such as emotions or concepts, and processes such as attention or reasoning, involve mixtures of the same domain-general ingredients. To uncover these ingredients, his proposal is to move to a multidimensional perspective, which weights the functional characteristics of neural networks in a probabilistic manner, according to a set of neuroscientifically relevant psychological (NRP) factors. This will alter the previous question subtly, such that the neuroscientifically relevant psychological factors (or ingredients), which contribute to a particular recorded neural pattern, must be given in terms of probabilistic weightings.

To illustrate what is meant by this proposal, consider the charts in Figure 5.4. These charts depict a number of machine learning classifiers (or categories), which resulted from training a network on a large set of fMRI results. The classifiers were trained to predict the outcomes of subjects' responses across 8 different tasks (e.g. risk taking, working memory, reading aloud) on the basis of the neuroimaging data. The classifiers were then simplified into a reduced set of dimensions, according to their predictive accuracy, and the resulting weightings shown in Figure 5.4 were taken to represent the degree to which the various dimensions exemplify the original tasks (cf. Poldrack, Halchenko, and Hanson, 2009, for further details). Anderson believes that these sort of dimensions could provide the relevant candidates for a set of primitive NRP factors, but at present require more comprehensive analysis. His colloquial reference to these dimensions (or NRP factors) as constituting a region's "personality" or "functional fingerprint" suggests that the right way to think of them is akin to how we may describe a friend's personality. For example, he suggests, you may know someone whom you would describe as considered, loyal and introverted, but not very funny or motivated. So too should we consider an agent's behaviour as partly *two-y* or *six-y*, but not very *five-y*. But what does it mean to be five-y?

As Figure 5.5 demonstrates, these dimensions are not as simple as the labels by

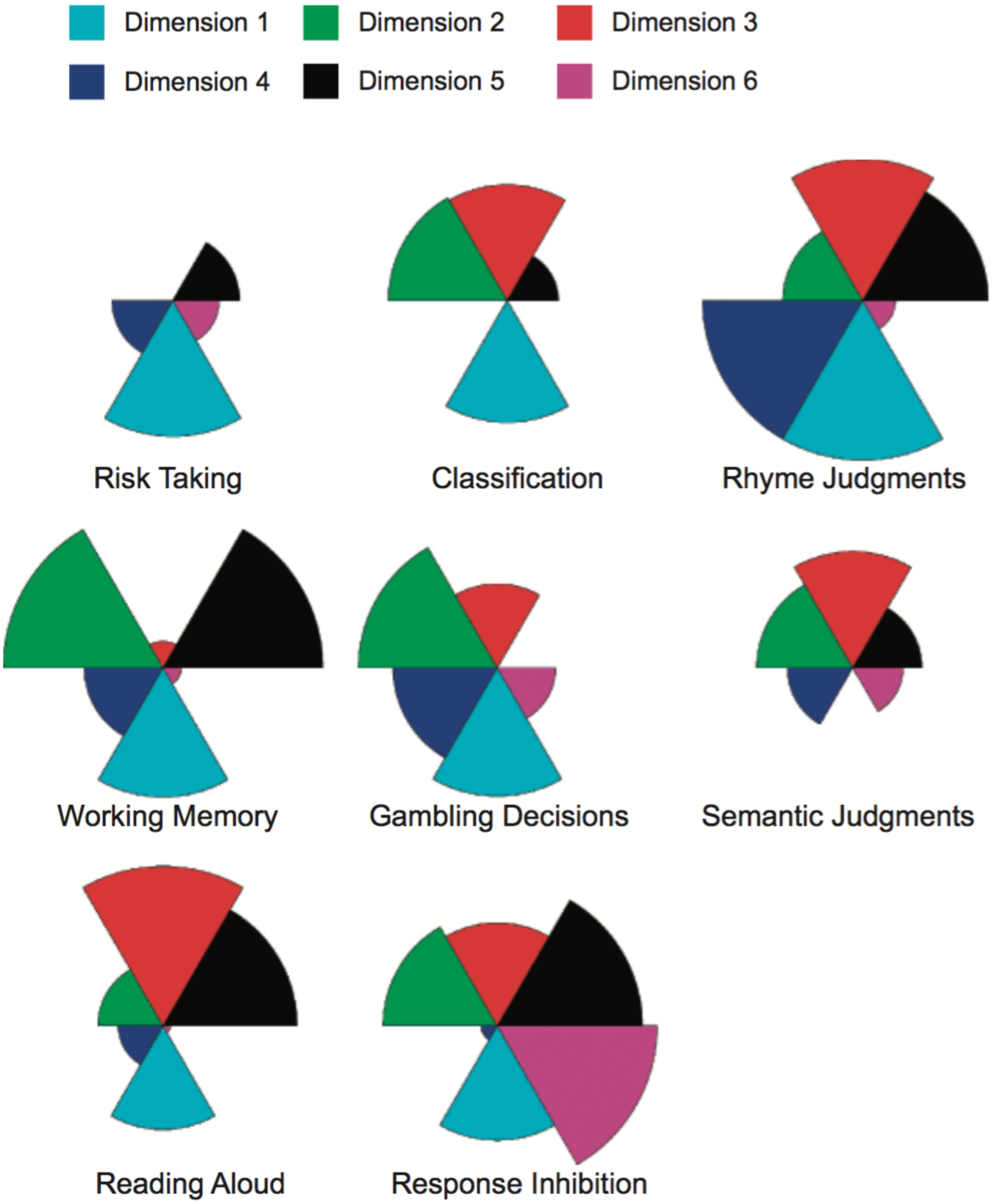


Figure 5.4: Reprinted from (Poldrack, Halchenko, and Hanson, 2009, p. 1369).

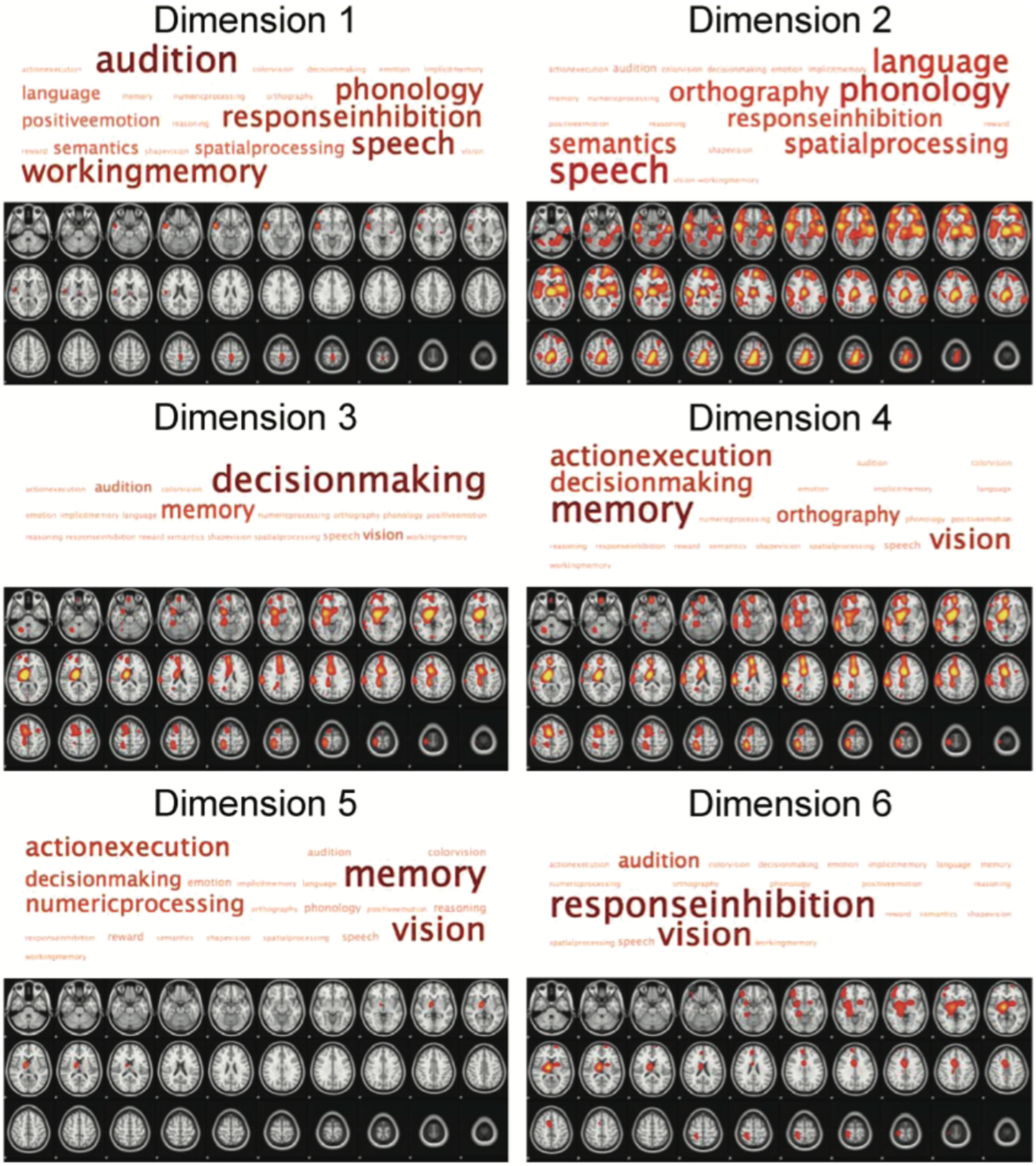

Dimension 1
actionexecution audition colorvision decisionmaking emotion implicitmemory
language memory numericprocessing orthography phonology
positiveemotion reasoning responseinhibition
reward semantics shapevision spatialprocessing speech vision
workingmemory
Dimension 2
actionexecution audition colorvision decisionmaking emotion implicitmemory language
memory numericprocessing orthography phonology
positiveemotion reasoning responseinhibition reward
semantics shapevision spatialprocessing
speech vision workingmemory
Dimension 3
actionexecution audition colorvision decisionmaking
emotion implicitmemory language memory numericprocessing orthography phonology positiveemotion
reasoning responseinhibition reward semantics shapevision spatialprocessing speech vision workingmemory
Dimension 4
actionexecution audition colorvision
decisionmaking emotion implicitmemory language
memory numericprocessing orthography phonology positiveemotion
reasoning responseinhibition reward semantics shapevision spatialprocessing speech vision
workingmemory
Dimension 5
actionexecution audition colorvision
decisionmaking emotion implicitmemory language memory
numericprocessing orthography phonology positiveemotion reasoning
responseinhibition reward semantics shapevision spatialprocessing speech vision
workingmemory
Dimension 6
actionexecution audition colorvision decisionmaking emotion implicitmemory language memory
numericprocessing orthography phonology positiveemotion reasoning
responseinhibition reward semantics shapevision
spatialprocessing speech vision workingmemory

Figure 5.5: Reprinted from (Poldrack, Halchenko, and Hanson, 2009, p. 1370).

which we refer to a friend's personality. Each dimension represents a different set of physical realisers in the brain, and a further weighted set of cognitive processes. For example, dimension 5 relates most strongly to memory, vision, action execution, decision-making and numeric processing, but only to a limited number of brain regions. By contrast, dimension 2 seems largely related to language in general, and is widely distributed throughout the brain. This fits with the research on the semantic system discussed earlier (Huth et al., 2016), and would be expected given the evolutionary picture that neural reuse is committed to, whereby more recently emerging capacities or behaviours should be supported by a greater number of structures.

That we are yet to identify what these NRP factors are precisely may present a concern for some. However, Anderson (2014, p. 134) claims that these findings, although challenging given the difficulty involved in understanding what the dimensions means, are nevertheless "very, very promising". They are promising because they point towards a new taxonomy for the cognitive sciences—a taxonomy that is naturalistic insofar as it is grounded in a neuroscientifically informed framework, and one which may help to unify the cognitive sciences. However, it is promising only inasmuch as it rejects the pursuit of neural specialisation. *Relative* specialisation can emerge if the loading on one factor significantly outweighs that of the other possible loadings, but specialisation considered by itself should not be a guiding methodological principle. As Anderson states:

> "I am suggesting that we adopt a framework according to which individual regions of the brain exhibit not functional specialisation (the implementation of a single mental operation) but rather relative functional differentiation—the development of regional *functional biases.* [...] In interpreting these factors we need to be open to *relational, interactive properties of situations.*" (ibid., emphasis added, p.151)

It is worth highlighting two aspects of the above quotation: a) the dispositional element involved in the use of the phrase 'functional biases', and b) the appeal to situational factors. The development of regional functional biases reiterates the claims made earlier regarding the notion of TALoNS. As many regions will exhibit pluripotency, trying to characterise their functional profile will necessarily require a dispositional perspective, such that all we can reasonably provide is a description of their functional biases (i.e those dimensions a region is most likely to implement).

Secondly, we should recall that these TALoNs are formed in order to shape an organism's behaviour in light of the relevant task demands, and to help determine its interactions with some aspect of the environment. The separate treatment of these aspects belies the fact that they are intertwined. The functional bias of a brain region is dispositional only inasmuch as the situation remains unspecified. When a situation demands the assembly of a group of neural circuits in response to a task demand, the token instance of any particular circuit will of course be set (albeit transiently).

With these final components of Anderson's framework in place, we begin to see why the principle of neural reuse is problematic for the cognitivist, and better suited to an embodied framework. Firstly, some of the cognitivist's commitments to claims such as modularity (Fodor, 1983) are simply incompatible with the principle of neural reuse. However, even if the cognitivist managed to defend these claims, it still seems as though an embodied perspective has a greater explanatory grip on the evidence that Anderson outlines. For example, consider the acquisition and development of certain functional biases that are shaped over ontogenetic timescales. In addition to more immediate changes in the current environment, the transient assembly of local networks must also be receptive to bodily-dependent adaptations, which themselves result from interactions with the environment. Consider again the case of learning to play the guitar. As you become more skilled, you are not merely acquiring theoretical knowledge about the instrument (e.g. chord structures), your body also adapts to the interaction. For example, the muscles in your palm and forearm that control your fingers strengthen, and callouses appear on your fingertips to dampen the feeling of pressing on the strings. Responding to these bodily developments is as important a consideration for your brain as any other stimuli. Moreover, they are a necessary part of what it means to become a more skilled musician, enabling more advanced interactions with the instrument. With regards to the brain, learning-dependent neural plasticity may afford it the ability to develop simple control loops for moving swiftly between chord sequences, which in turn demand less cognitive control than was initially required in the earliest stages of learning. By reshaping the cognitive architecture in such a manner—perhaps offloading some of the task demands onto the environment—the brain adapts to the situation and opens up new possibilities for learning (e.g. more creative interactions with the musical instrument). Anderson (2014) explores this idea in length, and argues that embodied cognition is best suited to explain the initially early empirical findings from areas such as mathematical cognition and linguistics.

What is important for the present discussion is the fact that these findings are more easily accommodated by an embodied framework. This could be acknowledged by adopting the principles defended by Clark (2008) (e.g. nontrivial causal spread, or the principle of ecological assembly, see chapter 1), or by pointing to the need for an action-oriented, embodied encoding scheme (e.g. moderate embodied theorists (Barsalou, 1999; Hommel et al., 2001)). Alternatively, because of the close connection with dynamical systems theory, and rejection of more traditional principles in cognitive psychology, the principle of neural reuse may lend support to theories of embodied cognition that adopt the *replacement* theme from chapter 1.[9]

[9] As was mentioned earlier in the thesis, we will not evaluate which of these options (or combination of options) is the most likely, given the empirical considerations.

Depending on the scope of their anti-representational commitments, in the case of decision-making and the neural encoding of value, these theories may align with the *comparison-only* position from the previous chapter. Not only is this a live option for embodied cognition in general, but due to the present uncertainty regarding the exact implementational details of the PP architecture, this must for the time being also remain a live possibility in the case of PP. In spite of this, and while we wait for further developments in computational modelling (Pezzulo et al., 2011), there are still many interesting questions to explore.

5.5 Effective Decisions

> "Until recently it has been widely assumed in the cognitive and neurosciences that, from a functional point of view, neurons can be adequately conceived of as simply adding up all of their excitatory and inhibtory inputs and transmitting an axonal spike if that integrated value exceeds a threshold." (Phillips, Clark, and Silverstein, 2015, p. 2)

The assumption alluded to in the above quotation has undoubtedly contributed to the defence of computational models in decision-making that rest on a model of deliberation and commitment (e.g. Padoa-Schioppa, 2011). Recall that in these types of models the brain accumulates sensory information in order to build a reconstructive model of the world, before deciding (perhaps by maximising expected utility) on one of the options, and then encoding this choice into a motor program, which commits the organism to the performance of some overt behavioural routine. As we hope to have shown over the last two chapters, increasing evidence from cognitive neuroscience and neurobiology supports an alternative picture whereby the co-ordination of multiple decision-making systems (some overlapping with sensorimotor regions) is achieved by widely distributed processes of contextual modulation, and transiently assembled networks.

If the accounts defended in this chapter are true, we argue that a number of claims—all of which can be elucidated and accommodated by an embodied version of PP—follow:

1. Decision-making is facilitated by distributed, competing systems in the brain that overlap with sensorimotor regions.
2. The distributed systems are transiently assembled into functional networks in order to respond to the shifting demands of the situated organism.
3. If we wish to understand how these systems contribute to effective decision-making, we need to acknowledge the explanatory importance of additional constraints that go beyond the brain.

We acknowledge that the truth of these claims rest on the empirical adequacy of much of the research outlined in the previous chapters. However, as is to be expected there is of course room to manoeuvre regarding specific details. More interesting than discussing these specifics is the requirement that we look beyond the brain for explanatory considerations pertaining to the mechanisms that underlie effective decision-making.

We have already acknowledged that an evolutionary approach constrains certain theoretical and conceptual commitments (e.g. cognitive processes are constrained by phenotypic attributes), but are there further components that are key to understanding an evolutionary approach to effective decision-making?

In an edited collection by Hammerstein and Stevens (2012), contributors focused on four key components for an evolutionary approach to decision-making:

1. Understanding the origins of decision-making mechanisms

2. Exploring why these mechanisms are robust

3. Accounting for variation between and within individuals

4. Investigating the pressures of social life on decision making

Although we can't hope to provide comprehensive arguments and reviews for each of these claims, we can discuss their relation to the research in this thesis.

Starting with (1), we can acknowledge the limitation of only highlighting the mechanisms in an adult human brain. Among other limitations, this would only provide an answer to the 'how' questions, without addressing the ultimate 'why' questions. Badcock, Ploeger, and Allen (2016) raises a similar worry in response to Anderson's neural reuse hypothesis, and argues (contrary to Anderson) that evolutionary psychology, with its focus on massive modularity, has the potential to be highly complementary to neural reuse. He claims that evolutionary psychology is appropriately situated to address the 'why' question due to its focus on the adaptive significance of cognition and behaviour—albeit often framed in the language of intentional psychology. By contrast, Anderson's account can help to show how these adaptive structures are physically realised. Even if a collaboration such as this requires careful conceptual consideration, in order to account for the aforementioned worries regarding taxonomic classification, it is likely that the division of labour could be beneficial.

(2) places an interesting emphasis on PP's account of precision-weighting and the careful balance between learning and action. Given the underlying imperative to minimise prediction error that drives much of the PP machinery, not only does this component emphasise a need to explain how the PP mechanisms contribute to an organism's fitness over various timescales, but also how they allow for the creativity and risk-seeking behaviours that seem to be so ubiquitous in biological organisms. We will explore a tentative solution to this in the next chapter.

Initially, it may appear as though (3) is of little concern for PP due to its status as a functional-level account. However, this would be to downplay the explanatory scope of the framework, and its claim to offering a unified account of cognition and behaviour, thanks in large part to the support from the free-energy principle (Friston, 2010, 2013). Further integration with both comparative psychology and behavioural ecology for cross-species comparison, as well as psychopathology for intra-species comparisons could be enormously beneficial for the framework (see (Barrett, 2015) for some initial examples pertaining to the former, and see (Bruineberg and Rietveld, 2014; Seth, 2014) for examples pertaining to the latter).

As is the case with (2), further exploration of the themes raised by (4) will undoubtedly expose numerous instances of how our embodied interactions with the world (here understood in a more inclusive sociocultural sense) have been importantly shaped by our evolution (here understood in a more inclusive sense that extends to the evolution of our sociocultural niche), and how this effects the sorts of decisions we routinely make. Of particular interest is whether the idea of a temporarily assembled network of local systems, which makes sense of the nested hierarchies found in the brain, can extend outwards to the nested hierarchies found in social structures. Does this idea gain any traction when transposed to the case of social organisation, where the context of a situation (e.g. party versus meeting) modulates the behaviours of the individuals that comprise the situation? As with (2) we will explore a tentative proposal concerning this point in the next chapter.

Chapter 6

Scaling Up?

Let's recap on what has been discussed so far. In chapters 1 and 2 we introduced embodied cognition and PP, in order to provide the necessary terminology for later chapters. In chapter 3 we argued that decision-making should be approached from the perspective of embodied cognition, using the idea of dynamically specifying, and selecting between, multiple action opportunities that are represented probabilistically in the brain. In chapter 4 we showed why an embodied account of PP was well-equipped to handle this notion of decision-making, and why it was opposed to a neurocentric conception of the mind. In the previous chapter we turned to explore some of the additional mechanisms that are required for understanding how the predictive, interactive brain flexibly adapts to changes in the environment, in order to support adaptive choice behaviour.

As promising as the PP framework appears, it has not yet been able to offer a truly scalable learning system that could successfully and efficiently learn to interact with the complex, real-world environments that characterise our world. Nevertheless, advocates of embodied PP can remain optimistic—given how the framework offers a neuro-computationally plausible account of cognitive processing—but should also acknowledge the theoretical and conceptual challenges that remain.

One challenge is particular pressing. Research on embodied decision-making has focused primarily on exploring the neural mechanisms in simple, *visually-guided* motor tasks, such as grasping an object or pressing a button—so called habitual decisions. Though this may be sufficient for explaining a wide variety of simple behaviors across a number of different species, humans (and some non-human animals) appear to possess far more complex decision-making capacities. Therefore, it is possible that the embodied decisions approach will be unable to account for the rich, and seemingly heterogeneous practices that traditional decision theory tends to concern itself with.

For example, when you decide to buy a house, or choose where to go on holiday, it is not immediately obvious how a notion of embodied decisions could be of any use. Buying a house or going on holiday are both activities that require long-term planning, and the prolonged maintenance of a desired goal-state in order to coordinate and constrain relevant behaviours (e.g. acquiring a mortgage and communicating with solicitors). It is not immediately clear how the predictive brain handles the representation of distal goal-states by making solely embodied decisions of the kind hitherto discussed. For example, what arrays of motor commands would be in competition within the sensorimotor system during long-term decision processes of this kind? Before turning to a speculative proposal in response to this, it is worth drawing a few distinctions and addressing some philosophical concerns.

6.1 Deliberative and Habitual Decision-Making

> “Behavioral and neuroscientific data on reward-based decision making increasingly point to a fundamental distinction between habitual and goal-directed [deliberative] action selection. Habits, in this context, are actions arising from direct situation-response associations. Goal-directed

> [deliberative] action, in contrast, involves prospective planning: selection among actions based on a forecast of their potential outcomes." (Botvinick and Toussaint, 2012, p. 485)

Within the decision-making literature, a distinction is often made between *deliberative* and *habitual* forms of decision-making, with competing model-based and model-free accounts put forward that try to capture the associated phenomena (Daw et al., 2011; Doll, Simon, and Daw, 2012; Lee, Shimojo, and O'Doherty, 2014).[1] In the case of deliberative accounts, model-based methods deploy structured, internal models of the respective domains, in order to decide between the various options based on their expected values. These accounts are increasingly studied in neuroeconomics (Glimcher and Fehr, 2014b), and are considered flexible enough to apply to a wide-range of circumstances, due to the abstract nature of the models utilised. In contrast, habitual decisions rely on previously-learned "cached" or "heuristic" strategies, rather than building a representation of the options to deliberate over (they are frequently used in reinforcement learning). Although less flexible than model-based accounts, the benefit of model-free methods, as Clark acknowledges, is that they "implement "policies" that associate actions directly with rewards, and that typically exploit simple cues and regularities while nonetheless delivering fluent, often rapid, responses" (Clark, 2013b, p. 5). In the case of deliberative versus habitual decision-making, more focus has been given to the latter. Most evidence in support of the former is limited to localisation claims, based on analysis of lesion studies implicating regions such as prefrontal cortex in the well-known work of Antonio Damasio (1994). Given the discussion of Anderson's work in the previous chapter, we have reason to

[1] In the quote at the start of this section Botvinick opts for the term 'goal-directed' in place of 'deliberative'. As will be discussed, we favour the latter due to the fact that we see the distinction as a matter of degree, and therefore acknowledge that some forms of habitual decision-making may nevertheless be goal-directed.

doubt some aspects of these studies (i.e. the taxonomic classifications).

A particularly pressing matter pertains to the notion of goal-directedness, which is sometimes used interchangeably with the term 'deliberative' (Botvinick and Toussaint, 2012). We believe that habitual decisions are also in some sense goal-directed, but before addressing this point, it is useful to focus on an intuitive example in order to raise a related philosophical issue.

Consider the decision of whether to go for a run. The initially distal goal-state of running is satisfied once you begin your workout. However, there is a more fine-grained series of causal events that exists between the time when you purportedly "decide" to 'go for a run' and the satisfaction condition of 'having gone for a run'. We wish to argue that the decision to 'go for a run' should be equated with the full series of fine-grained causal events—beginning with the mental representation of the goal-state considered, and ending with the overt performance of the necessary behaviour.[2] As such, the decision of whether to 'go for a run' is temporally extended over time, and as we will see, is partially constituted by events that extend beyond the brain and body.[3] This analysis fits with the characterisation given by Lepora and Pezzulo (2015) for their embodied choice model. When comparing this model with two alternative models (based on the drift-diffusion paradigm), they argued that by incorporating ongoing action into the deliberative process, the natural deadline for the termination condition of a decision was the completion of the relevant situated

[2] For simplicity we put no requirements on how far you travel, or how fast you run in order for the statement, "I went for a run" to obtain. This vagueness is likely to be a characteristic of many behaviours, and we believe that a certain flexibility is necessary to account for differences in an individual's own satisfaction conditions (e.g. less than 1km may not suffice for a professional long-distance runner.

[3] Obviously some decisions will be extended over shorter or longer period of times dependent on the framing of the decision (e.g. deciding between two sandwiches at a shop versus deciding on a new career path).

action. This stands in contrast to traditional models that view the termination condition for some decision (i.e. the commitment) to be some threshold—perhaps reached on the basis of accumulating evidence encoded in some neural region—and action to be the mere means for reporting the decision outcome. However, our evaluation of embodied decisions should make it clear that we favour a view of decision-making whereby deliberation and commitment are not construed in such cognitivist terms. Defending this statement requires a number of claims to be explored.

The first is that goal-states do not exist independently of the agent who is representing them; that is, they have no mind-independent objectivity (Gallese and Metzinger, 2003). Only *goal representations* have a physical existence, realised by particular patterns of neural activity.[4] Secondly, although we speak of goal *representations*, as we use the term, they differ from traditional notions of representation in a number of ways: a) they have no truth-conditions, only conditions for satisfaction that are directed towards the deployment of certain actions that minimise prediction error through active inference, b) they are strictly grounded in facts about the agent's embodiment, and although possibly multimodal at some high-level of abstraction, are not amodal in the sense used by the cognitivist (Burr and Jones, 2016).

The motivation behind (a) follows from the truth of the first claim, defended by Gallese and Metzinger (2003, p. 371), that "no such things as goals exist in the objective order of things", therefore, "a goal representation *cannot* be true or false." However, in PP, goal representations (in the form of higher-level predictions) are required by active inference, and thus have satisfaction (or fulfilment) conditions based on the imperative to minimise sensory prediction error. This leads to consideration of (b), and to the question of whether the existence of goal representations, such as the one posited in the running example, require more than can be provided by an

[4] We believe this to hold also in the case of intersubjective goal-states, where multiple agents are pursuing "shared goals". However, we do not consider this point any further at present.

embodied account of PP.

For example, the distal goal-state to ‘go for a run’ appears to be abstract (i.e. possibly encoded in something like a language of thought), despite being decomposable into more fine-grained sub-events (e.g. put on trainers; warm-up muscles; fill water bottle; lock door on leaving house; spend 20 minutes attempting to get your GPS watch to detect your location). Furthermore, each of these multi-functional events can be considered independent of the specific goal-state—I may fill my water-bottle because I am thirsty and require a drink; I will lock my door whenever I leave my house (irrespective of whether I am going for a run). This fact regarding the multi-functionality of sub-events doesn’t appear to change even when the series of sub-events is so frequently performed that I rarely deviate from the order of performance. Alternatively, another decision (e.g. whether to buy a house) may be performed so infrequently, and contain such a diversity in terms of sub-events, that I will have very little idea of the string of events in advance.

With the aforementioned in mind, we can pose several questions: 1) does a complete account of decision-making require use of both habitual and deliberative strategies, 2) if so how does the brain choose between them, and 3) does an affirmative answer to the first question require augmenting an embodied account of PP with more traditional cognitivist principles? In this chapter we will argue that the difference between habitual and deliberative strategies is a matter of degree, but still requires the positing of an arbitration mechanism. We will also argue that an embodied account of PP is sufficient to accommodate a complete account of decision-making. In fact, the particular solution to the issue of goal representations we defend is unique to the embodied account.

6.1.1 Combining Approaches

To begin, we will consider the first possibility that the agent is able to make use of both habitual and deliberative strategies, and that this corresponds to the use of model-free and model-based approaches respectively.

On the one hand, model-free approaches seem incompatible with PP due to its strict adherence to the existence of hierarchical generative *models* throughout the cortical hierarchy (Hohwy, 2016). Yet on the other, we have seen how these methods can implement policies that associate actions directly with rewards (chapter 4). In addition, model-based strategies are compatible with PP's adherence to generative models, but may require proponents of embodied PP to develop alternative computational methods to account for the seemingly abstract nature of goal representations.

To try to resolve the first conflict, Clark (2013b, 2016b) has appealed to precision-weighting mechanisms (chapters 2 and 5) to provide a way for the agent to switch flexibly between these two strategies on the basis of expected precision and accuracy. Commenting on the model-free strategy, he claims:

> "[...] the use (when ecologically apt) of simple cues and quick-and-dirty heuristics is not just compatible with prediction-based probabilistic processing: it may also be actively controlled by it." (Clark, 2013b, p. 8)

Here, Clark appeals to work in reinforcement learning (e.g. Daw, Niv, and Dayan, 2005; Gläscher et al., 2010), which shows how model-free strategies can be developed that embody implicit values associated with certain action sequences (policies) through trial and error. This can be achieved without the need to retain an explicit value or construct a detailed representation, as the policies will have been reinforced over developmental learning because of the high probability of leading to (or being-correlated with) rewarding states. These "cached" policies can then be redeployed at a later stage by the agent if they are estimated to be more reliable than the alternatives. In this manner, policies related to habitual decisions can be subsumed within the generative models of the brain, by virtue of higher-level predictions that assign a high probability to the policy on the basis of PEM (Pezzulo et al., 2016). Such work bears close resemblance to work by neo-empiricists, such as Prinz and Barsalou (2000), who argue that context-sensitivity in cognition requires a collaboration between dynamic approaches (akin to model-free methods) and representational approaches that are grounded in an agent's embodiment (akin to model-based methods).

This allows PP to make use of model-free and model-based methods respectively, but doing so requires positing a mechanism that is able to switch flexibly between them as required by the environmental demands. Clark's (2016) solution to this is to again appeal to precision-weighting mechanisms as a way of modulating the effective connectivity of the brain's networks in response to the myriad biasing inputs that collectively determine an agent's needs (e.g. affective information, sensory information, prior knowledge). In addition to appealing to precision-weighting mechanisms, he also acknowledges (in line with earlier work[5]) a recent argument by Pezzulo, Rigoli, and Chersi (2013) regarding the development of a neural control mechanism that switches between the separate systems. A number of studies, including the aforementioned work by Pezzulo, Rigoli, and Chersi, have argued that the brain decides between these heuristic (model-free) strategies and more deliberative forms of model-based reasoning by employing some form of arbitration mechanism (neural

[5] Clark (1997a, p. 136) argued that in addition to neural structures that respond to external stimuli, we need to acknowledge the existence of so called "neural control structures", which are "any neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures, or processes—that is to say, any items or processes whose role is to control the inner economy rather than to track external states of affairs or to directly control bodily activity."

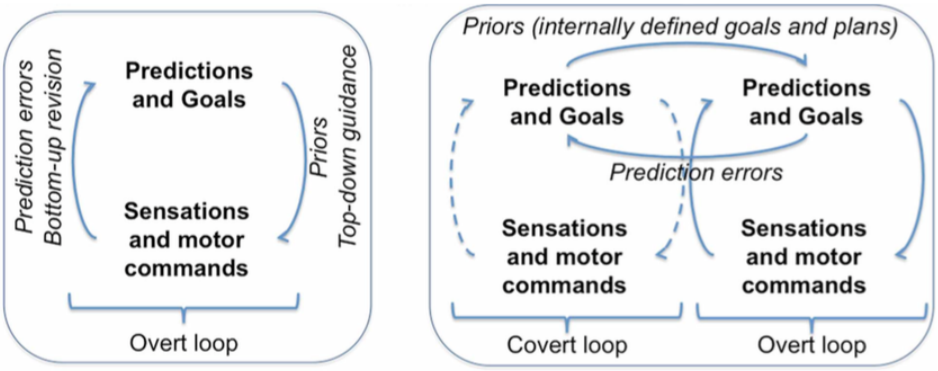


Figure 6.1: Left: simplified view of online active inference. Right: Offline "optimising" loops coordinate with online control of action and overlap with relevant sensorimotor circuits, but are also detachable from overt motor control. Figure reprinted from (Pezzulo, 2012).

controller), which predicts the respective reliability of various policies (Daw, Niv, and Dayan, 2005; Dayan, 2012; Lee, Shimojo, and O'Doherty, 2014). These positions acknowledge the importance of combining *both* habitual and deliberative mechanisms for determining instrumental choice behaviour—neither is sufficient on its own.

The stronger claim that instrumental behaviour depends on both mechanisms, leads to the claim that deliberative forms of reasoning entail another kind of cost to the agent—mental effort and delay. Pezzulo et al. (2016) argue that different types of policies can be distinguished according to whether they are associated with extrinsic value (i.e. the expected physical reward for completing the action) or epistemic value (i.e the additional information gain or resolution of uncertainty). For simplicity we will refer to these types of policies as *extrinsic policies* and *epistemic policies* respectively.

Pezzulo (2008, 2011) has argued at length that the importance of deliberative forms of reasoning is best seen in their ability to allow agents to plan for future actions by emulating potential actions utilising sensorimotor representations that have been optimised through successive interactions with the world. He argues that the former anticipatory capacity depends on habitual forms of choice behaviour, and may have evolved as successive elaborations on earlier sensorimotor circuits. As more deliberative forms of reasoning emerged, organisms gained the ability to internally optimise extrinsic policies by simulating their sensorimotor consequences before performing the overt behaviour. This idea is compatible with the PP framework, and is depicted schematically in Figure 6.1.[6]

Policies that can be used for optimising offline action plans are also important for understanding choice behavior in exploration-exploitation dilemmas, where the agent faces a decision between exploiting some previously learned strategy, or risking exploration and increased uncertainty. As Pezzulo et al. (2016, p. 324) state:

> "[...] epistemic value is key in so-called "costly" choices, when an accurate estimation of the context is necessary to secure a reward and a wrong choice implies a "cost" such as long delay in reward consumption."

In cases like this, there may be a payoff for considering actions with high epistemic value in order to ascertain whether there are other options that have not yet been considered. Such cases represent a sort of best-guess for the agent, based on prior knowledge of how similar situations have played out in the past. However, sometimes the situation will be too complicated to deliberate over in this manner. This idea has important connections to understanding choice behaviour in ecologically-valid situations, where seemingly irrational behaviour may be explained by appealing to

[6] Although tangential to this thesis, such a view is closely connected with work on the origins of mental imagery and its connection to motor control (cf. Jeannerod, 2006), where it is argued that the ability to simulate motor behaviour offline may be a pre-requisite for mental imagery. This is because true mental activity should be produced endogenously, and not as a direct response to perceptual stimuli (Grush, 2004).

the proper complexity of the task environment, and acknowledging the cost of more deliberative forms of reasoning (Fawcett et al., 2014).

It seems, therefore, that there is a strong case to be made for the interaction of both habitual and deliberative mechanisms in the predictive brain. We do not wish to reject this idea in its entirety, nor disagree with the empirical findings, but we do wish to propose an alternative conceptual interpretation focused on the underlying physical substrates that give rise to the distinction in the first place. Moreover, we wish to argue that the distinction between habitual and deliberative forms of decision-making is a difference in degree, rather than a difference in kind. The motivation for this echoes the earlier motivation for both the account of embodied decisions defended by Cisek, and the neural reuse hypothesis defended by Anderson. That is, the brain is a product of evolution, and is thus subject to descent with modification and natural selection. We will argue that accepting this requires a greater consideration of the scope for habitual decision-making when properly situated in the world, and that more deliberative forms of decision-making (where required) will be embodied in nature.

Cisek and Pastor-Bernier (2014, p. 10) seem to acknowledge this when they state that “phylogenetic continuity motivates us to consider how abstract decisions such as economic choice evolved within a system originally adapted for realtime embodied choices, and how the architectures subserving these abilities may be related.” And yet, this very statement is preceded by the following:

> “Obviously, humans are capable of making decisions that have nothing to do with action, and understanding such abilities is of great scientific and clinical interest. In fact, it is quite possible that the distinction between different kinds of decisions, such as abstract versus embodied decisions, is paralleled by a distinction between different neural structures and circuits that subserve these scenarios.” (ibid.)

The remainder of this chapter is aimed at reconsidering the distinction expressed above, by attempting to weaken the strict separation between a) the types of embodied decisions explored in this thesis hitherto, and b) the abstract (disembodied) economic decisions that are alluded to in the above quote.

6.2 Hierarchical Cognitive Control

To understand how the brain could have evolved more sophisticated mechanisms for choice behaviour we need to understand how the more complex goal-directed choices, characteristic of so-called deliberative decisions, can be decomposed. Developing on the earlier proposal of choice behaviour in PP and active inference, Pezzulo (2012) and Pezzulo, Rigoli, and Friston (2015) have begun to explore the role of associative learning in active inference, and how predictive mechanisms initially concerned with online control could be detached for offline cognitive control of more abstract, long-term consequences of behaviour. Pezzulo claims:

> "As the sensorimotor control system of early organisms evolved (to face increasingly harder individual and social problems), it gradually began predicting increasingly long-term and abstract consequences of behaviour." (Pezzulo, 2012, p. 1)

This idea is related to recent developments in computational neuroscience (Klaes et al., 2011; Pezzulo et al., 2014), which have begun to explore how hierarchical models of cognitive control can be understood as successive elaborations of earlier sensorimotor control mechanisms. *Cognitive control* is the ability to internally guide behaviour in concert with goals, plans and wider contextual knowledge. It requires the simultaneous management of multiple, hierarchically nested goal representations, across different spatiotemporal scales, in order to constrain action selection. As such we can refer to it as *hierarchical cognitive control*, in order to acknowledge the nested structure of goal representations.

This definition of hierarchical cognitive control draws our attention to several components:

1. The hierarchical organisation of goal-directed choice and behaviour.

2. The simultaneous co-ordination of multiple goal representations (across different spatiotemporal scales).

3. The constraining nature of this process.

This is important for the development of PP and embodied decisions, as long-term planning, unlike the continuous, situated interaction inherent in many of the earlier examples, requires the agent to maintain a commitment to an extended goal-state, which may not afford an immediate action opportunity (e.g deciding to buy a house). Let's see how this could be achieved within the embodied PP framework.

To begin, Botvinick (2008) discusses how the hierarchical structure of more abstract goal-states can be understood as a successive elaboration of lower-level action representations. For example, think of the process of making a cup of coffee, and the number of successive steps that are required. The process can be decomposed into separable control sequences (policies) (e.g. go to kitchen, get objects from cupboard, heat up water, prepare coffee), and each of these sequences in turn will unfold into further nested sequences of actions (e.g. getting objects from the cupboard will involve opening and reaching actions, and possibly relocating occluding objects). Though we can see the decision to make a coffee as a successive decomposition, this presumes we have first learned the molecular structure involved in the necessary actions—including very fine-grained movements (e.g. gripping objects). Rather than a successive unpacking, the process of learning over both evolutionary and developmental timescales requires a successive *elaboration* on previously learned molecular control sequences, which progressively tunes the relevant sensorimotor circuits to become associated with some represented goal-state (e.g. grasping for a desired object).

Related to this work, Pezzulo (2011) has argued that more complex cognitive architectures (assumed to be required for long-term planning) could have emerged as developments on control mechanisms for earlier situated action, and importantly retain the embodied aspects of the earlier systems (contra the classical sandwich model), given that they rely on the emulation of the same underlying sensorimotor circuits. Though admittedly speculative at this early stage, he argues that the capacity for cognitive control (i.e. the ability to internally guide behaviour in concert with goals, plans and wider contextual knowledge) is an elaboration on the earlier anticipatory architecture of sensorimotor control apparent in many living organisms. In short, as agents began to face more complex problems, they faced increasing evolutionary pressure to predict longer-term consequences of their actions, and at some point began mentally simulating these consequences in covert loops, without the need for activating overt behaviour (Grush, 2004; Jeannerod, 2006) (see figure 6.1). This raises the possibility that scaling up the notion of embodied decisions may be possible, and the PP framework may have significant explanatory interest to those developing computational models of decision-making. It also leads to the possibility that the distinction between habitual and deliberative strategies are a matter of degree, rather than distinct modes. Even though many goal-representations will have distal satisfaction conditions, they will nevertheless be grounded in the sensorimotor mechanisms that gave rise to them in the first place, and in some cases (as demonstrated by the embodied decisions work) still play a fundamental role in choice behaviour.

It could be argued that this 'evolutionary continuity' perspective commits its advocates to holding the related view that the main adaptive problem for cognition is not the reduction of uncertainty between some abstract representation of how the world is, but the the identification of adaptive actions. Moore (2012, p. 1) takes this line, claiming that it accords with the idea that every organism has what he calls "sunk capital" in preferred ways to interact with the world, based upon its evolved neural architecture and physiological traits. He argues, "[the organism's] challenge is to use that capital to operate adaptively. Modelling the world in any disinterested manner is a luxury; quickly identifying adaptive ways to go on is a necessity." This brings us to the first proposal for how PP can scale-up the notion of embodied decisions to accommodate more deliberative forms of decision-making. We should begin by attempting to see how a deliberative choice is in fact decomposable into a successive series of hierarchically nested actions, and then try to identify the local networks that are responsible for learning the associations, and deploying the sensorimotor commands in the first place. This would be the preferred approach for some in the enactivist tradition, such as Barrett (2011, p. 16), who claims "[a]s clunky and unparsimonious as it may seem, it is possible that long chains of associations are exactly the way in which many skills are learned, and complex behaviors are brought about." It is also an approach that views cognition as a wholly action-oriented adaptation.

6.2.1 Action-Oriented Hierarchies

A growing number of researchers in the cognitive sciences have begun exploring what Engel et al. (2013) call the 'Pragmatic Turn' (cf. Engel, Friston, and Kragic, 2016, for a collection of recent papers). This is the view that the brain is primarily action-oriented, and its purpose is to coordinate and regulate the organism's ongoing interactions with changes in its environment. The ACH can be seen as an example of this approach, as it is committed to the idea that our diverse repertoire of choice behaviours is supported by a dynamic process of distributed probabilistic competition between representations of action opportunities. Recall, this process coordinates multiple neural regions in order to process separate streams of sensorimotor information in line with higher-level goals, which are in turn biased (and determined) by interoceptive information communicating the current and future needs of the agent.

As Anderson (2014) has convincingly argued, the brain, having been established by natural selection and descent with modification, is able to support this flexibility due to a functional architecture that is characterised by interactivity and functional differentiation (also see Bickhard, 2015, for another view committed to interactivity in neural architecture). PP unifies this work in a neurocomputationally plausible framework that has the conceptual and theoretical tools to be able to explain how the brain is able to support these processes by using key neuromodulatory mechanisms.

A potential challenge to this action-oriented, pragmatic turn is to point to the empirical evidence in support of sensory mappings (e.g. in visual cortex), which seem to represent objective properties of the world, independent of any particular action. Putting aside the earlier claims made by those in the active vision framework, which already count against this challenge, we can acknowledge that the transductions at the sensory peripheries exist and are vital for adaptive functions. However, we can acknowledge this without the need to make the stronger claim that the mappings or encodings therefore *represent* the objects that caused them. Instead, we can see the encodings as more closely integrated with the global functioning of the brain, and as setting up indications of what the organism could do on the basis of higher-level beliefs (priors).

We have already seen evidence in favour of this view from Hosoya, Baccus, and Meister (2005), who demonstrated how the receptive fields of retinal ganglion cells dynamically alter when the organism moves to a new environment (chapter 2). It was argued that these changes, when understood from the perspective of PP, are

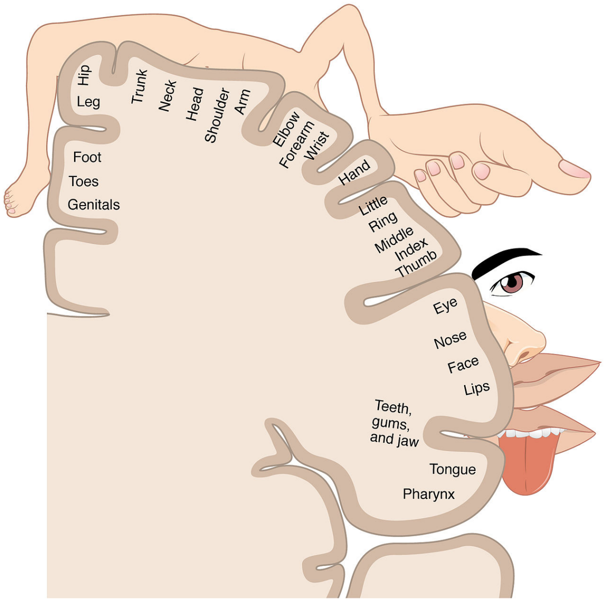


Figure 6.2: Homunculus representing the mapping between somatosensory cortex and body parts.

Image reprinted from Wikimedia Commons under Creative Commons Licence 3.0: `https://en.wikipedia.org/wiki/Cortical_homunculus#/media/File:1421_Sensory_Homunculus.jpg` [Accessed: 12/09/16, Author: OpenStax College]

adaptive because they contribute to the efficient coding of the shifting statistical information coming from the environment. An action-oriented account would argue that this makes sense when one considers how different environments will potentially require different actions, and thus different sensorimotor dynamics. Although this work undermines the serial, encapsulated nature of cognitivism, a re-interpretation of neural encodings—influenced by global network dynamics—is not sufficient on its own to support the view that the brain is primarily action-oriented. However, in addition to this work, Graziano (2016) reviews a large body of empirical data originating with a study reported in (Graziano et al., 2002) that points to another unexpected finding regarding the functional organisation of the brain—this time in motor cortex.

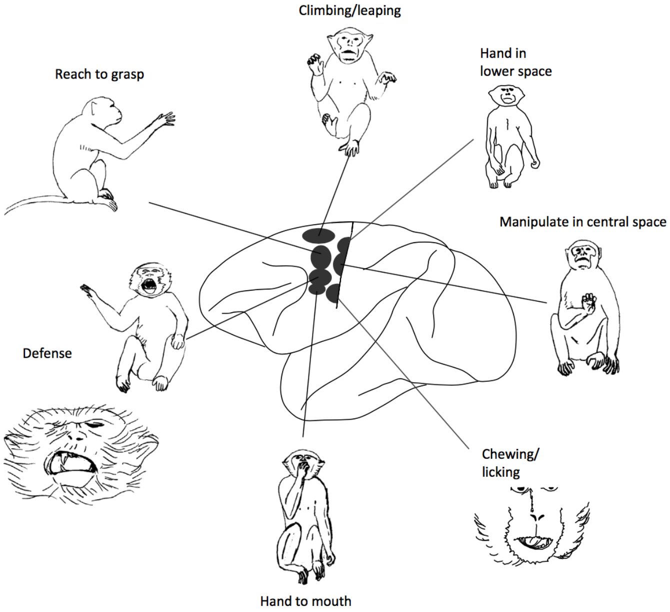


Figure 6.3: Example of an ethological action map in macaque motor cortex representing organism-relevant behaviours. Figure reprinted from (Graziano, 2016).

Motor cortex has traditionally been understood as containing a somatotopic representation of the body (in primary somatosensory cortex), and is often depicted by way of a homunculus (see figure 6.2). However, this map is not as neatly delineated as the homunculus depiction would suggest, and in fact the map contains substantial overlaps between regions and corresponding bodily parts. Graziano's proposal is to reinterpret the functional organisation of this region in terms of what he calls an 'ethological action map'. Simply put, motor cortex is composed of zones which are related to a different "ethologically relevant type of action". In the case of macaques[7],

[7] Graziano is keen to point out that although the map depicts motor cortex in macaques, the empirical findings for ethological action maps are significantly more robust, involving different species and multiple methods. He states, "The action map has now been studied in rats, mice, prosimians, monkeys, humans, squirrels, and cats, using a great range of methods including electrical and optogenetic stimulation, chemical manipulation, lesions, single neuron recording, functional imaging, anatomical tract tracing, behavioral analysis, and computational modeling." (Graziano,

figure 6.3 depicts examples of some ethologically relevant actions. Using cortical stimulation to target different zones (also see previous footnote), Graziano and his colleagues found that the activation of a particular region (over behaviourally-relevant timescales) was akin to pressing a button that reliably activates the entire network to collaboratively produce a particular action. This produced a coordinated behaviour that is both species-typical, and which likely evolved to serve some adaptive function. Echoing the sentiments of previous network approaches, Graziano therefore dismisses the idea that there is a one-to-one mapping between a cortical region and a muscle or body part. Instead he argues that we should adopt a more global systems-level approach, and acknowledge the multi-functional role that ethnologically-relevant action representations can play within a more interactive brain. This work also appears to strongly complement the frequent appeal to policies (control sequences) in PP.

Both of the above findings from retinal cells and motor cortex seem obviously well-suited to their respective proposals, given their anatomical connections and location. However, what about more distal regions such as prefrontal cortex; how is a region such as this considered action-oriented? One emerging area of interest comes from the study of mixed-selectivity neurons. Using single-cell recordings, Rigotti et al. (2013) demonstrate that neurons in prefrontal cortex (PFC) demonstrate *mixed-selectivity*: that is they respond non-linearly to a wide variety of inputs, and thus have high-dimensional receptive fields. These findings also challenge the traditional idea that the brain can be understood using simple methods of functional decomposition. Rigotti et al. (ibid.) propose that this high-dimensionality is key to the acquisition of more advanced cognitive capacities (e.g. cognitive control). A densely populated set of *mixed-selectivity* neurons, provides the brain with a way to flexibly adapt to changes in the environment, without the need for extensive rewiring. Rather,

2016, p. 121)

dense hubs could play a fundamental role in the co-ordination of distributed neural activity. However, high-dimensionality in mixed-selectivity neurons does not need to be interpreted as indicative of abstract, amodal representations *a la* cognitivism. Rather, it is perfectly consistent to refer to these regions as multimodal, serving the co-ordination of widely distributed regions that are themselves action-oriented. From an evolutionary perspective, the existence of these regions also makes sense when we consider the flexibility that they bestow upon an agent in responding to a changing environment, which is efficiently achieved using limited neural real-estate. Novel computational methods are being developed that explicitly make use of such mixed-selectivity neurons to demonstrate how networks composed of them enable far greater complexity than traditional methods, with limited neural resources (Enel et al., 2016). Additionally, the existence of these types of neurons helps us understand the wide-spread effects of lesions in prefrontal cortex, given the myriad roles they serve in co-ordinating disparate regions of the brain.

Taken together, the aforementioned research appears to indicate the need for a hierarchical, action-oriented perspective on the brain's function (see Engel, Friston, and Kragic, 2016, for a range of additional arguments and evidence); one which an embodied account of PP is well-equipped to handle. And yet, in spite of this, the picture remains incomplete.

Consider the earlier example of deciding to go for a run, and the subsequent formation of a goal-representation that is posited to co-ordinate the subsequent series of actions that lead to the fulfilment of this goal. Recall, that this goal-representation was underspecified, and could be satisfied in a number of ways. If we restrict ourselves to appealing solely to neural mechanisms, it seems unlikely that we will be able to explain how any particular set of actions is selected, except in the most restricted of cases. This point is made by Basso (2013), who asks how PP accommodates long-term planning. He states:

> “[...] the future goal state created in the beginning is accurate only in some particular circumstances (i.e., when both the task and algorithm are well-defined). In most cases, people are used to facing *underspecified* tasks in which a future goal state cannot be employed to derive the intermediate states” Basso (ibid., p.1, emphasis added)

PPs proposal that the brain selects policies that drive action based on predicted success (i.e. expected probability and precision) is more intuitive in habitual cases of choice behaviour, where the associated control sequences are simple actions such as grabbing (e.g. ethologically relevant types of action). But in more deliberative forms (e.g. whether to go for a run), as we saw earlier, there are vastly more ways in which the goal representation can be satisfied. Echoing the above quotation, we can refer to this as the *underspecification challenge*. Appealing to expected probability and precision-weighting mechanisms is an important first step, as it helps us understand why certain policies are favoured over others—they have a higher posterior probability of minimising prediction error. However, this answer is incomplete, and so we now turn to look beyond the brain for additional constraints.

6.3 Constraining and Coordinating Decisions

The previous sections lend support to the more moderate forms of embodied cognition discussed in chapter 1, which argue that neural encodings represent sensorimotor activity, rather than some organism-independent reality. However, many forms of embodied cognition (e.g. enactive, embedded, and extended cognition) go further in acknowledging the key role that organism-environment interactions play in shaping or constituting cognition. In this section we explore how an action-oriented view of hierarchical cognitive control, can be bolstered by appealing to wider influences and constraints placed on the organism. We will explore four types of constraints:

physiological constraints, temporal constraints, affective constraints, and sociocultural constraints. It should be noted that each of these sections reflect somewhat speculative sketches for the development of an embodied account of PP, and further work is required to fully defend the claims that are made. Nevertheless, the following sections provide further reason to support the claim that decision-making is an embodied process, and PP offers a suitable framework to further investigate and develop the following ideas. They also provide additional responses to help us overcome the underspecification challenge, which will be revisited in section 6.4.

6.3.1 Physiological Constraints

The body undoubtedly plays a fundamental role in shaping cognition. It can provide a grounding relation for conceptual content as in cases of moderate embodied cognition (e.g Barsalou, 2008), or, as we will see in this section, it can provide a more reliable basis for active inference. This latter claim can be decomposed into two focal points concerning the explanatory role that the body plays in understanding decision-making: sensorimotor constraints and efficiency constraints. We will begin by looking at the first.

We have already seen how, in PP, predictions arise from generative models in the brain. These models are encoded as probability density functions, which are structured according to an increasing level of spatiotemporal scale. The predictions at the lowest levels correspond to the activity of sensory receptors encoding input at small and fast spatiotemporal scales, whereas the higher-level models provide more general contextual information concerning larger and slower structures in the environment. The theoretical and empirical support for this picture has already been documented in work by Friston et al. (2010), who argue that the formal similarities of their hierarchical models to the hierarchical structure of the motor system lends them biological plausibility (Kanai et al., 2015), as well as offering a wide explanatory scope (Friston, 2010). Additionally, Hohwy (2013) has argued that the hierarchical structure leads to a highly restricted set of possible parameters that exist at the lowest-levels of the control hierarchy, because of the limited ways that certain parts of the body could be configured. These parameters further restrict the set of possible actions, and may allow for automated or simple reflexive patterns in specific circumstances.

For example, consider the case of a fine-grained goal representation such as 'grasp the apple'. The object being grasped may be replaced with any number of appropriately sized objects, although only some of these objects will have any significance for the agent. However, the behaviour of grasping, will be highly restricted, based on the physiological characteristics of the organism (e.g. size of hands). As such, the policies that are associated with this action will rarely deviate from a set of highly-restricted control parameters, unless the agent undergoes some physiological change (e.g. loss of a hand, or muscular atrophy). In this manner, we can see how those goal-representations associated with habitual decisions, perhaps involving ethologically relevant types of behaviours, are less susceptible to the underspecification challenge (see above).

Far from being a hindrance to an agent, these restricted features can have adaptive value, allowing the agent to more easily detect and learn about the relevant features that emerge in the course of interacting with the world. This will in turn help to optimise the selection of possible actions in decision-making, as the specification of the relevant parameters can be reliably constrained by relevant factors of their embodiment. For instance, as the eyes saccade from left to right, the visual scene will shift from right to left in a predictable manner, relative to the speed and direction of saccadic motion. An active perceiver can exploit regular relations between sensory input and motion of this kind in order to detect objective structural and causal features of the environment. As we first saw in chapter 1, these predictable relationships between bodily movement and sensory input are known as sensorimotor contingencies (SMCs) (O'Regan and Noe, 2001). Despite being a commonly referenced notion in the embodied cognition literature, Hohwy (2014, 2016) has claimed that this view need not entail that the mind is embodied in any important manner. Instead, he argues, concepts fundamental to embodied cognition can be subsumed within an internalist PP. He states, when discussing the reliably occurring relationships between movement and expected sensory input: "It is crucial to acknowledge that accommodating embodied cognition in this way happens within the strictures of the self-evidencing brain." (Hohwy, 2014, p. 17)

We have elsewhere argued, contrary to Hohwy, that this aspect of embodied cognition is implied by active inference (Burr and Jones, 2016), and does not suggest an internalist reading. In addition, Seth (2014) has also proposed utilising SMCs to extend the PP framework to account for phenomena such as perceptual presence, and its absence in synaesthesia. To see why this is the case, we should first note that reliable law-like regularities do not merely exist between the world and individual (active) senses. There are also law-like regularities that can be detected between senses, and are fundamental to an organism's development.

For example, Dahl et al. (2013) explore the case of the development of a wariness of heights. This appears to be absent in human infants with little or no crawling experience, but becomes exceptionally strong (sometimes debilitating) over the lifespan of an individual. To explain this, Dahl et al. demonstrate that when an infant is carried there is no real correlation between the infant's proprioception and vision, but this changes when an infant begins to crawl. At this point, the infant is able to keep their head oriented towards a particular point, and begin to learn about their body by experiencing the reliable and consistent correlation between proprioceptive signals—importantly including the motor commands—and the optic flow. As Soliman and Glenberg (2014, p. 209) highlight, it is this correlation that becomes the "basis for a stable world". This basis originates from (and is constituted by) sensorimotor interactions between the body and the world. Moreover, disrupting this correlation, can provide useful information that the world is changing, while the body remains stationary. For example, consider the case of sitting on a stationary train at the platform while watching another train move, and the temporary feeling of uncertainty that is experienced prior to your realisation that there is no correlated feeling of acceleration. Returning to the crawling infant, an analogous disruption can be caused by placing the infant near a visual cliff (e.g. a cliff that is covered with a perspex sheet to give the illusion of an actual cliff-edge). Doing so causes the infant distress, but only when they have learned to crawl and have thus experienced the correlation between different sense modalities. (Dahl et al., 2013)

Other correlations also exist between different senses (e.g. detecting the location of a sound-source by moving your head to alter the temporal asynchrony originating from the ears, and simultaneously centring the visual field in order to detect the cause). Prior to learning these sensorimotor contingencies, it may be that there is a latent imperative to partake in what Hohwy (2013) calls “itinerant wandering”, but is also known in some cases as ‘motor babbling’ (e.g. infant behaviour). This seemingly random wandering can unfold while the agent determines which action is most likely to minimise prediction error most effectively. Though this unguided exploratory behaviour can result in local (temporary) increases of prediction-error, in cases of high uncertainty the undirected initiation of movement can be helpful in exposing further (potentially more valuable) options. This is a useful strategy for agents to adopt when exploiting the current environment is no longer viable, but where the possibility of exploration is over-determined by too large a number of possible options. Although there may be uncertainty in the environment, there will always remain a high-degree of reliability that emanates from sensorimotor interactions.

In (Burr and Jones, 2016) we argued that overlooking this reliability meant that the significance of the body in cognitive processes would be diminished if one attempts to subsume embodied cognition within the structures of an internalist PP. An organism's phenotype determines what is valuable (in the autopoietic sense discussed in chapter 4) for the organism at birth. The task of the organism is then to find the most efficient ways to acquire valuable states—according to PP, this is governed by the imperative to minimise prediction error. Therefore, by encoding information that pertains to body-world interactions (i.e. action opportunities), an organism can exploit this sensorimotor knowledge to reliably minimise prediction error, over less reliable organism-independent representations.

An apparent problem at this stage is that there seem to also be law-like regularities in the world that do not immediately pertain to an agent's interactions (e.g. the regular rising and setting of the sun). Given this, it may seem like an agent with the necessary cognitive capacities should also encode representations of this interaction-independent causal structure. However, there are differences between our access to environmental and embodied regularities worth noting. Firstly, it is possible to decouple oneself from environmental regularities in a way that one cannot from bodily ones. Secondly, in the case of sunrises, the sensory input will vary depending on contextual features such as the direction one is facing, whereas sensorimotor contingencies are relatively invariant across contexts. We argue, that sensorimotor interactions are more reliable because, unlike other statistical regularities in the environment, the agent can exploit them through action-oriented representations, which, as some have argued, could be adapted and reproduced over phylogenetic timescales (for some theoretical arguments in support of this claim, see Clark, 2013c; Friston, 2010, 2013). To paraphrase, while the statistical regularities in the environment would have to be internalised through interactions and learning, it is likely that the statistical regularities pertaining to the ways in which our bodies interact with the environment have been stable enough over evolutionary time-scales so as to be genetically determined. It isn't necessary to learn about most important sensorimotor relationships because they can be built in to an organisms morphology and neural architecture, thereby setting the priors in advance (this is admittedly a speculative claim). Furthermore, as we have just seen, the controllability of these interactions by the agent during ontogenetic development is likely to contribute significantly to the shaping of the representations.

Therefore, we would expect that an agent is more likely to exploit the sorts of reliable organism-environment interactions that are contingent upon its phenotype over less reliable (more uncertain) organism-independent worldly structures. Interacting vicariously with the environment via sensorimotor contingencies affords the agent a more reliable manner in which to minimise uncertainty. Just as scientists test hypotheses by conducting experiments using well-calibrated lab equipment, perceivers must likewise test their predictions by using their bodies to interact with their environment.

Hohwy (2013, p. 224) argues that we are able to cope with noisy signals from the environment because the world is a uniform kind of place that kindly affords reliable statistical inference. However, as previously discussed, this reliability does not arise merely because the world is uniformly reliable. It arises precisely because certain parts of the environment, namely our bodies, behave in a more reliably predictable manner than the rest of the environment beyond them. The world would be a far less kind place if it werent for the fact that our bodies are part of it and that their predictable behaviour is, in some sense, under our own control.

We are beginning to see how the body can offer adaptive constraints that afford the agent a reliable basis for interacting with (and importantly learning) about the world. These interactions are dynamic, but we have not yet considered the additional temporal constraints that add important evolutionary and adaptive pressures. This will be important for understanding how a distal goal-state, which initially appears underspecified, is in fact more constrained than we may imagine. The ongoing pursuit of prediction-error minimisation demands learning the most efficient ways to achieve these more distal goals, and environmental pressures play an integral part in this process.

6.3.2 Responding to Urgency

PEM must be responsive to causes in the environment across a number of spatiotemporal scales. For example, perhaps a perturbing influence happens regularly at the order of milliseconds, but is also nested within a further perturbing influence that occurs on the timescale of minutes. The hierarchical structure of the brain is well-suited to accommodate these changes, but it is also well-suited to regulate additional factors such as the biomechanical costs involved with certain actions, which themselves may differ across spatiotemporal scales. Anyone who has done long-distance cardiovascular activities (e.g. running or cycling) and suffered with the difficulty of inadequate pacing over extended timescales will attest to the importance of being receptive to the body's changing demands across multiple timescales. The neural mechanisms, which are responsible for encoding the relevant expectations associated with biomechanical costs, must themselves be governed by efficiency constraints in order to use energy effectively—the brain requires energy, just like the rest of the body.

Some have argued that the efficient coding of information (e.g. predictive coding) may be responsible for the existence of suboptimal choice behaviour (Summerfield and Tsetsos, 2015), as the agent will not be receptive to all the relevant information it could be. However, efficient coding schemes may also lead to more robust decision-making abilities, which despite departing from optimality in many situations, may nevertheless maximise information-processing for a limited capacity system situated in an uncertain, changing world.

Recent work by Cos, Duque, and Cisek (2014) provides an interesting development to this idea. They argue that human subjects make a rapid prediction of biomechanical costs when deciding between actions. For instance, when deciding between actions that yield the same reward, humans show a preference to the action that requires the least effort, and are remarkably accurate at evaluating the effort of potential reaching actions as determined by the biomechanical properties of the arm. Cos et al. argue that their study (a reach decision task) supports the view that a prediction of the effort associated with respective movements is computed very quickly, and furthermore, that measurements of cortico-spinal activity initially reflects a competition between candidate actions, which later change to reflect the processes of preparing to implement the winning action choice. Although there may be a possible disagreement concerning the exact manner in which cost functions are encoded (see chapter 4), note how the representation is encoded as an expectation of the cost associated with motor control, rather than as a cost involving some economic good. Studies like this provide further reasons for taking the work of Lepora and Pezzulo (2015) seriously, due to the close connection with the aforementioned commitment effects (see chapter 3).

Learning about the average biomechanical costs associated with performing certain actions could be a useful first-step in the formation of simple heuristics that stand in lieu of rational deliberation, and may also explain the presence of purportedly maladaptive decisions (e.g. sunk-cost fallacy). In short, some tasks may simply require more effort to formulate a deliberative plan, which outweighs the risk associated with simply choosing incorrectly (e.g. choosing between a pair of socks). In situations like this the risk may be minimal, and may lead to the misapplication of a strategy that is maladaptive in the current environment. Being receptive to these changes in context is therefore of the upmost importance, as the value of many actions will vary contextually, dependent on factors such as fatigue, injury and environmental resistance (e.g. hill-climbing). However, an alternative strategy is to simply allow the constraints of the body and environment to stand-in as a constituent part of the decision-making process. This is where dynamic, responsive feedback from the body, as input back into an ongoing decision is so important, and where work in situated cognition can provide constructive assistance. As Lepora and Pezzulo note:

> "In situated cognition theories, the current movement trajectory can be considered an external memory of the ongoing decision that both biases and facilitates the underlying choice computations by offloading them onto the environment." (ibid., p. 16)

This also reflects work in embodied cognition such as Clark's *principle of ecological assembly* (first seen in chapter 1), which states that an agent will "recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort." (Clark, 2008, p. 13) An interactive, predictive brain, which can flexibly alter its effective connectivity on the fly, is well-suited to such a distributed form of adaptive decision-making. Sometimes, the best decision is to offload part of the choice onto the typically reliable dynamics of the body. Moreover, in cases where this strategy leads to undesired commitment effects, there may also be an opportunity for researchers to learn about the cognitive architecture of the agent in question.

Accommodating urgency exposes another important connection between PP and embodied decisions. Given the level of urgency of an agent's higher-level goal states, the gain of incoming sensory information should be adjusted accordingly. Higher-level goals should therefore encode more abstract expectations regarding the optimal amount of time taken to deliberate in any given decision. Cisek and Pastor-Bernier point to the importance of an urgency signal in their work:

> "[...] in dynamically changing situations the brain is motivated to process

> sensory information quickly and to combine it with an urgency signal that gradually increases over time. We call this the 'urgency-gating model'."
> (Cisek and Pastor-Bernier, 2014, p. 7)

When the urgency of a decision is low, only an option with strong evidence will win the probabilistic competition. However, as the urgency to act increases, the competition between the options can increase, such that a small shift may be sufficient to alter the distribution. Cisek and Pastor-Bernier highlight a number of neuroimaging studies that support the existence of such an urgency signal, and argue that evidence accumulation may not be the only cause of the build-up of neural activity seen during decision-making experiments.[8] By emphasising the importance of precision-weighting as a neuromodulatory mechanism for altering the brain's effective connectivity, PP may be able to further develop this line of thought in a more unified framework, which demonstrates the closely intertwined nature of perception, action, emotion, learning and decision-making. Doing so will also provide an additional, explanatorily relevant factor that could be fruitful in understanding how more deliberative forms of choice behaviour unfold according to various embodied constraints.

6.3.3 Stabilising Disorder: A coordinating role for emotions

Emotions play a key role in our ability to make decisions, and the impacts of this role are becoming increasingly well understood. This is in large part thanks to work in behavioural economics and cognitive neuroscience (see Lerner et al., 2015; Phelps, Lempert, and Sokol-Hessner, 2014, for reviews). One significant area of development comes from work in cognitive neuroscience that demonstrates how cortical and

[8] Connecting this with work on the dark-room problem in PP (Friston, Thornton, and Clark, 2012) could be of mutual benefit for the two areas.

sub-cortical regions mutually influence each other, reflecting coordinated patterns of activity. As Pessoa (2009, 2014) has argued, emotional processes can directly influence (i.e. enhance or impair) key regions associated with cognitive control, and have a direct influence on behavioural performance. Further work needs to address these important processes, in order to overcome what Pessoa (2014, p. 410) calls "cortical myopia", or a "cortico-centric" perspective. He defines this as follows:

> "A cortico-centric framework is one in which the "newer" cortex controls subcortical regions, which are typically assumed to be relatively unchanged throughout evolution. In this view, cortical expansion is thus a matter of cortical regions being set up so as to control "lower" centers. In sharp contrast, if both cortex and subcortex change, they may change in a coordinated fashion. In this case, the resulting circuitry is one in which cortex and subcortex are mutually *embedded*." (ibid., p. 413)

Add to this, work by Friston (1997), who demonstrated how activity in a given region was functionally dependent on the activity of another region (e.g. whether inferotemporal cortex is considered "face selective" depends on activity in posterior parietal cortex), and we are again reminded of the earlier point regarding functional plasticity in the brain's networks. Does this principle help us understand the role of emotional processing in decision-making, or does it simply add further complexities to the picture? To attempt to answer this question, we can focus on two areas: a difference between *immediate* and *expected* emotions, and emotions as modulatory mechanisms.

In emotional theory, a common distinction is made between immediate and expected emotions. The former pertains to the emotions that are felt during the process of decision-making, whereas the latter pertain to the emotional states that are expected to obtain given the outcome of a decision. It should be obvious how the differences between these states can affect decision-making respectively. For example, if you are angry while making a decision, you may be more likely to make a rash decision without properly considering the various options, whereas if you are currently experiencing fear and expect that performing some action will result in a more positive state, you may rank this option as most desirable—potentially overlooking whether it has a higher probability than alternative options. Traditionally, decision theory has focused on the latter emotional state, but the importance of immediate emotions for decision-making cannot be overlooked. Loewenstein and Lerner (2003) review a wealth of behavioural studies to demonstrate the importance of both kinds of states—the details of the studies need not concern us. What is important is that if an embodied account of PP is to accommodate a complete account of decision-making, it should be able to provide an explanation for how these states respectively influence our choice behaviour. This is where the aforementioned work on the mutual influence between cortical and sub-cortical regions, and the widespread effects of neural modulation is so important.

Consider again the ideas discussed in the previous chapters regarding the empirical evidence for a distributed, mutli-systems perspective on the neural mechanisms that underlie our decision-making capacities. In addition to the anti-cognitivist picture that is supported by the ACH and embodied PP (i.e. a probabilistic competition between action-oriented representations in key sensorimotor regions), there is also evidence that cortical and sub-cortical regions mutually interact to influence cognitive control and emotional processing. Why would the brain have evolved to employ so many regions in a distributed fashion, and how does this bear on the distinction between immediate and expected emotions?

Two responses can be given to the first part of this question. Firstly, neural reuse is an efficient method of minimising the costly production of building new neural regions. If redundancy and pluripotency (e.g. mixed-selectivity) can support robust, adaptive interactions in a less demanding manner, this principle of efficient (predictive) coding seems sensible. Secondly, we can ask how many kinds of values an agent must be receptive to? Consider the case of possible starvation. In these instances, few other action opportunities will take priority (except maybe protecting oneself from a threat). Therefore, actions that lead to obtaining food will exceed, and possibly suppress entirely, any other non-food options. However, once food is obtained, attention can be directed towards an alternative need (perhaps encoded in a separate region). Although it seems like the cognitivist picture, which emphasises deliberation over multiple representations in an executive region (using a common currency), is compatible with this idea, the empirical evidence seems to point towards a different view.

As has already been discussed, key regions like the amygdala (Pessoa, 2010), the AIC (Gu et al., 2013; Seth, 2013) and the thalamus (Barrett and Simmons, 2015; Kanai et al., 2015) have been shown to play fundamental roles in the integration and modulation of affective information, and in turn a key role in active inference. Cisek and Pastor-Bernier (2014) argue that this information is *inseparable* from decision-making, acting as a biasing input that is a proper part of the probabilistic competition that underlies our decision-making capacities. If we are to take this view seriously, the view of a detached cognitive module, which sits upstream of sensory processing and integrates and deliberates over abstract representations needs to be thrown away. In its place we need to consider decision-making in the brain in a more dynamical manner, with less strict boundaries between perception, cognition, action and emotion (chapter 4). An embodied account of PP, developing on the work of embodied decisions, is well-suited to fill this role.

So what does this mean for the distinction between immediate and expected emotions? As Bickhard has argued, in an interactive brain we can understand higher-level processes such as cognitive control and emotional processing as a contextualising basis for lower-level processes:

> “[...] (local) temporally slow processes set parameters for—thus modulate—the dynamics of faster processes, and large spatial scale processes can induce weak coupling among smaller scale processes, thus inducing and modulating attractor landscapes in the dynamics of those faster, smaller scale processes.” (Bickhard, 2015, p. 233)

In PP, these slower processes will take place at the higher-levels in the hierarchy, and cascade down to influence the lower-levels, which will in turn propagate back up through the hierarchy communicating error signals. This bi-directional influence dovetails nicely with the aforementioned mutually-influencing communication between different regions in the brain—communication that is assisted (and influenced) by key neuromodulatory mechanisms (see chapter 5).

We propose, tentatively, that the difference between immediate emotions and expected emotions can fit naturally into this scheme by appealing to the multi-level processing of PP. Immediate emotions would thus be considered as a contextualising influence that is realised by the (slower) higher-level processes, which in turn coordinate the lower-levels, and determine what is considered valuable in any given instance. These higher-level emotions will impact how incoming information is perceived and evaluated, and which regions are assembled in order to assist in action selection. Expected emotions would correspond to the goal-state associated with multi-modal predictions, which would likely include predicted interoceptive states, as well as other possible perceptual predictions commonly associated with the relevant emotional episode. Support for this view can be found in the work of (Lewis and Todd, 2005).

Lewis and Todd (ibid., p. 211) view the synchronization of neural structures as a “rapid self-organizing process that consolidates activity across all levels of the nervous system", and posit that emotions act as a coordinating influence within this process. The emotional episode is posited to explain how an agent selectively attends to certain perceptual states, and is not perturbed by alternative goal obstructions (i.e. alternative action opportunities). An emotional episode thus acts as an important coordinating process within the aforementioned synchronisation.[9]

The dynamic approach to emotions is important, as it means that an emotion is an evolving process, rather than simply an end-point. As such, they are well-suited to perform long-term, action-guiding roles, as according to Lewis and Todd, they are directly concerned with "improving our relations with the world through some action or change of action."[10] For example, under Lewis and Todd's account, an emotional episode can direct attention away from obstructions that prevent the agent from obtaining some goal. They take this to be a fundamental factor that (partially) defines an emotional episode, and means that an emotion can assist an agent in overcoming and responding to goal obstruction, perhaps explaining why emotional episodes persist over time. This is important for understanding how distal goal-states, which require commitment to a series of coordinated actions, can be satisfied. Furthermore, it is during these stages that we can most clearly understand the difference (and relation) between an immediate emotion and an expected emotion.

Under this interpretation, an agent's immediate emotion, is partially responsible for co-ordinating actions by directing an agent towards some salient goal-state,

[9] Recall that an emotional episode is differentiated from core affect. The former is associated with an intentional object, which is considered to elicit the agent's attention along with a corresponding cognitive appraisal of the stimulus (chapter 4). However, even when an emotional episode is not present, there is always some felt core affect in the background, potentially ready to evolve into an intentional emotional episode.

[10] Lewis and Todd (2005) also provide a dynamical model of how an emotion evolves from a pre-reflective state to an attention-grabbing conscious state. Although interesting, we need not worry about this aspect of their account for present concerns.

but also impacts the evaluation of other goal-states that are associated with a secondary (expected) emotion. Alternate goal-states (and associated emotions) being considered, will be evaluated (in part) relative to the current affective state of the agent, and not independently of the current emotional episode. Recall that even if an emotional episode is not conscious, there is always a core affective state that is felt by the agent (see chapter 4). This means that the same option will be weighted differently according to the current emotional episode (or core affect), and the value of some action opportunity will thus be dependant on the contextual factors involved (cf. Loewenstein and Lerner, 2003).[11]

These constraints help the interactive, predictive brain to assemble the appropriate regions of the brain, best suited to respond to the current challenges it faces, based on current expectations defined by its ongoing activity. Action opportunities are thus selectively attended to, based partly on the needs of the organism as determined by affective information, and may be modulated by ongoing dynamics that can be associated with action-guiding emotional episodes. Similar to alternative constraints, emotional episodes can be seen as playing a coordinating role, the absence of which would likely result in unmanageable disorder. Although this picture is not sufficient to account for how all distal goal-states are obtained through co-ordinated action selection, it appears to be an important contributing factor—as suggested by the work on embodied decisions.

[11] This principle is similar to the idea of narcissitic sensory systems explored by Akins (1996, p. 345), who argues that rather than representing veridical information from the world, sensory systems (e.g. pain system) work by encoding information in a manner that she characterises with the phrase "But how does this all relate to ME?". Under this view, the brain is not attempting to reconstruct an organism-independent representation of the world, but is responding to incoming sensory information on the basis of ongoing endogenous dynamics. This may not be a "rational" way of responding to the world, but it is perhaps optimal given the limitations of the organism.

6.3.4 Scaffolding Cognition

Finally, we turn to the social influences that are commonly associated with emotional understanding and cognitive processing in general.

It is strange to consider how willing we are to recognise our physical limitations, when we contrast this with the seemingly contrary attitude we hold towards our mental abilities. When it comes to our physical limitations we embrace their existence, as is evidenced by the myriad physical constructions we assemble in society to augment our natural abilities (e.g. bridges to help us cross rivers, transport that enables us to move faster than we are able to otherwise, and medicine that allows us to live longer). However, even if we acknowledge the existence of mental aids such as calculators or external memory devices, we seem to be reluctant to see them as anything other than convenient *alternatives*, which resemble the computationally-equivalent inner workings of our minds, rather than the true augmentations that they are. An embodied account of PP embraces our cognitive limitations, and looks to explore how culture and the external world have been shaped to enable us to smoothly interact with the world in the most efficient manner, as determined by PEM. The need to move beyond a neurocentric perspective is nicely expressed in a quotation from anthropologist Clifford Geertz:

> "Man's nervous system does not merely enable him to acquire culture, it positively demands that he do so if it is going to function at all. Rather than culture acting only to supplement, develop, and extend organically based capacities logically and genetically prior to it, it would seem to be an ingredient to those capacities themselves. A cultureless human being would probably turn out to be not an intrinsically talented, though unfulfilled ape, but a wholly mindless and consequently unworkable monstrosity. (Geertz, 1973, pp. 67-68, quoted in (Lende and Downey, 2012))

Of interest to this project is recent work by Lende and Downey (2012) on *The Encultured Brain*, which explores recent interdisciplinary work in the fields of neuroscience and anthropology. The relevance of this interdisciplinary work (known as *neuroanthropology*) to an embodied account of PP is captured in the following:

> "A central principle of neuroanthropology is that it is a mistake to designate a single cause or to apportion credit for specialized skills (individual or species-wide) to one factor for what is actually a complex set of processes."(ibid., p. 24)

Like embodied PP, neuroanthropology realises that exploring the brain alone is insufficient to explain the myriad skilful interactions that define adaptive life, and instead requires turning to the notion of *enculturation*. Enculturation can be defined as the idea that socio-culturally shaped cognitive processes only emerge from the interaction of an organism situated in a particular environmental niche. Neuroanthropology claims that many neurological capacities, such as language or skills, simply do not appear without the immersion of an organism within a particular culture (i.e. enculturation). In fact, Lende and Downey (ibid., p. 47) even state that "embodiment constitutes one of the broadest frontiers for future neuroanthropological exploration", and thus neuroanthropology is interested in "brains in the wild", to appropriate a phrase from (Hutchins, 1995). This requires understanding not only how brains support skillful activity, but also how interactions with the environment in turn re-wired our brains. Initial evidence for this is found in the following studies: differences of neural structure and function between East Asian and Western cultures that may account for differences in notions of self (Park and Huang, 2010); cross-cultural differences in the ability of subjects to accurately judge the relative and absolute size of objects (Chiao and Harada, 2008), as well as evidence for differences in their spatial representation of time (Boroditsky and Gaby, 2010).

Of particular interest for the current thesis, the notion of enculturation is often appealed to by those most accurately described as enactivists. For example, De Jaegher and Di Paolo (2007) appeal to a notion of *participatory sense-making* to account for how social meaning can be generated and transformed through the interactions of a group of individuals collectively participating in collaborative activities. The notion of *participatory sense-making* is an extension of the enactivist notion of *sense-making*, which is the process that describes how an autopoietic system creates meaning through its lived experiences (Thompson, 2007) (see chapter 4). For the enactivist, meaning does not exist independently of a system, but is defined by the selective interactions that are specific to certain phenotypes. This is the idea behind Francisco Varela's statement that "living is sense-making", which is captured in the following example from (Thompson, 2004, p. 386):

> "That sucrose is a nutrient isn't intrinsic to the structure of the sucrose molecule; it's a relational feature, linked to the bacterium's metabolism. Sucrose has significance or value as food, but only in the milieu that the organism itself brings into existence. [...] Living isn't simply a cognitive process; it's also an emotive process of sense-making, of bringing signification and value into existence. In this way the world becomes a place of valence, of attraction and repulsion, approach or escape."

The fact that sucrose has value to a bacterium is partly constituted by the autopoietic interactions of the bacterium. In a similar manner, meaning is constituted by the participatory activities of a social group interacting with one another. However, although enactivism is commonly associated with an anti-representational approach to cognition, the notion that meaning is created in social activities need not imply an anti-representational view.

For example, Steels (2003) describes some fascinating experiments using robots whose behaviour is controlled by simple control mechanisms that are themselves non-representational. Each robot had simple phototaxic sensory systems that enabled it to navigate towards a light-source, while avoiding obstacles. The light source was connected to a charging station that re-charged the robot's batteries. Importantly, there was no centralised module that controlled the behaviours, rather the light-seeking/obstacle avoiding behaviours emerged from simple mechanisms that were active in parallel and interacted with the environment in real time. However, in this initial set-up, there was no motivation for the robot to leave the charging station, so Steels introduced a competitor in the form of a black box near the charging station that diminished the available energy source unless regularly pushed away by the robot. Secondly, multiple robots were placed in the same environment. These additional influences resulted in fascinating emergent behaviours, that could not be explained by positing some inner representational control system, as the robots had been explicitly constructed without them. For example, cooperative strategies relating to the robot's distribution between workload (i.e. black box pushing) and recharging developed, with multiple stable strategies emerging. In some cases, one robot performed twice as much work as another, whereas in other cases the workload was balanced.

In spite of this, Steels still refrains from advocating an anti-representational account. Instead, he believes that representations emerge from the need for agents to co-ordinate increasingly more complex behaviour. Representations begin as material structures that can later be internalised by sufficiently complex cognitive systems. He gives the following example to highlight this emergence:

> "Consider a grass lawn in the form of a square between two buildings on a university campus. The buildings are on diagonally opposite sides. There is a path around the square but people who need to go from one building to another naturally take the shortest path, which cuts right

> across the lawn. Even though the gardener has planted a nice smooth grass lawn (and perhaps put up a little sign saying ‘Do not step on the grass’), a natural path arises sooner or later, as the grass starts to fade away in the places where people step on it. The gardener can try to fight this, but is probably better off creating a real path by clearly marking the naturally emerging path with some sort of material structure and by using gravel on the path so that the grass will not grow. Now everybody, even someone who has never been on campus, will recognize instantly that this is the logical path to take.” (ibid., p. 2390)

Representations emerge from interactions with an environment and become meaningful because of their role in organising some activity (e.g. walking between buildings). These material representations may become internalised at some point to form the basis of inner thought, but if they do they are still understood as “organizers of activity rather than abstract models of some aspect of reality” (ibid.). Such a view requires moving beyond the brain, to understand cognition as a distributed activity, constituted by a brain, a body and a world interacting with one another. Even some of our most complex cognitive capacities, such as mathematical cognition have recently been argued as examples of enculturation (Menary, 2015).

As our cognitive capacities have become increasingly advanced, we can now see how our ability to shape our environment can be understood as a way of simplifying the requirement to minimise prediction error, by making our environments more predictable. Hutchins (2014) offers a nice example of restructuring our material environment through certain behaviours that can be understood as cases of dimensionality reduction. For example, he offers the case of queueing as an instance of enabling a more straightforward perceptual experience. This is because the experience of a one-dimensional line, is more predictable than the experience of a two-dimensional crowd, and in turn the experience of a queue has a lower entropy (and thus a lower source of surprise) than the experience of a crowd. He states, “[t]his increase in predictability and structure is a property of the distributed system, not of any individual mind.” (ibid., p. 40)

There are outstanding philosophical issues with the above claims. For example, it is unclear whether enculturation commits us to defending an account of extended cognition (Clark, 2008), scaffolded cognition (Sterelny, 2010) or distributed cognition (Hutchins, 1995). Despite being worthwhile questions, yet again they are too tangential for the current discussion. More pertinent is explaining how the above notion of enculturation can help with our current project of attempting to answer the underspecification challenge raised at the start of this chapter.

To begin, by connecting with work in neuroanthropology and enculturation, PP may find a complementary approach that provides a way of answering the ultimate ‘why’ questions behind how the brain co-evolved alongside our body and external environment to supplement the ‘how’ questions that the PP framework seems well-suited to explain. This is an important distinction to make, and was raised at the end of the previous chapter in connection with the key components needed for an evolutionary approach to decision-making. These components, outlined in (Hammerstein and Stevens, 2012), were:

1. Understanding the origins of decision mechanisms

2. Exploring why these mechanisms are robust

3. Accounting for variation between and within individuals

4. Investigating the pressures of social life on decision making

Neuroanthropology’s connection with each, though especially (1), should be obvious. However, each of the four constraints discussed above has the potential to elucidate one of these four components. For example, with (2) there is an obvious connection between what we termed the reliability of physiological constraints and their “robustness”. Certain areas of evolutionary theory such as comparative anatomy, which explore important concepts such as homologous and analogous structures, can undoubtedly provide evidence for why certain physiological constraints are more robust in particular evolutionary niches, and how the exploitation of related SMCs can assist with adaptive choice behaviour. Furthermore, where variation exists (3), we can also gain insight into differences in choice behaviour. And of course, neuroanthropology is well-suited to investigating the pressures of social life on decision making (4). PP should not be concerned with this division of labour, but rather embrace the additional theoretical constraints that can supplement the answers it provides to the ‘how’ questions.

This complementary approach has the potential to help us understand how the brain, working in collaboration with the body and the (sociocultural) world, has evolved to facilitate more effective means of decision-making. Identifying constraints is an important part of scientific discovery, and is vital when attempting to bridge the sometimes large conceptual gaps that exist between disciplines that nevertheless share a common interdisciplinary goal. We believe an action-oriented approach to embodied predictive processing offers the most fruitful interdisciplinary framework for the cognitive sciences. However, as was the case with Anderson’s positive proposal in the previous chapter, by adopting this framework we appear to be required to rethink the status of some of our concepts. We end this chapter with a speculative suggestion for how to respond to the underspecification challenge, which requires rethinking the nature of decision-making.

6.4 Back to Scaling Up

We began this chapter by considering the relation between different forms of decision-making (i.e. habitual and deliberative). We then explored a recent trend in the cognitive sciences to a more action-oriented framework, and developed on this idea by exploring several constraints that can inform our understanding of how decisions are made. The purpose of this was to see whether the notion of embodied decisions can scale up, by appealing to several constraints that may further our understanding of the role of embodied and embedded interactions in decision-making. The notion of the underspecification challenge in long-term decision-making was used to frame this discussion.

Herbert Simon famously claimed that choice behaviour should be understood as constrained by a pair of scissors, where the blades represent the limitations of the environment and the cognitive capacities of the agent in question (Simon, 1990). Although his ecological approach to (bounded) rationality was partially cognitivist in nature (i.e. favouring an abstract symbolic approach), the core truth of his statement remains valid.

PP has much to offer for the second of the blades, but if we acknowledge the lessons of neuroanthropology and enculturation and seek a complementary approach, it could also place important theoretical constraints on the first. For those familiar with the literature, it may seem strange to emphasise this, given that so much time has already been devoted to this task in the area of bounded rationality. It may also seem strange that so little of this thesis has explicitly dealt with this literature, aside from the scattered remarks in chapter 3. This is not an accident. There are many conceptual differences that exist between the two frameworks, and attempting to reconcile these difference before turning to the main focus of the thesis would simply have been too great a task. Nevertheless, it seems apt to pay lip service to bounded rationality, especially given the following suggestion for how to reinterpret decision-making in light of the work discussed in this thesis.

The previous sections have demonstrated how isolating the brain from the body and the world impedes our ability to understand how effective decision-making and the satisfaction of distal goal-states is possible. It has also been suggested that the decision to perform some action is a product of a probabilistic competition between multiple action-oriented representations encoded by the brain, which are further constrained and coordinated by the dynamics of the body and the world. PP offers a powerful framework to situate this reconceived notion of decisions as a more dynamic process of selectively attending to relevant action opportunities, which unfold as a result of worldly interactions. In this way, our decisions are enacted in a world that consists of co-developed, emergent material structures, which have been shaped through successive interactions, in order to facilitate effective choice behaviour and minimise prediction error. Does this idea help us scale-up habitual cases of decision-making to account for more deliberative forms?

We stated earlier that the distinction between these two forms should be approached as a matter of degree, following Clark's (2013) suggestion that we could understand this as a balance between the extent to which prior predictions or error signals drive action (section 6.1.1). The more that prior predictions influence active inference, the more we can associate the corresponding decision as deliberative in nature (i.e. relying more on stored knowledge). By contrast, if the organism expects high precision for the current sensory input (i.e. error signals) then we can treat the corresponding choice behaviour as more situated, and guided by the environment (i.e. utilising sensorimotor control loops or heuristics). Obviously we should treat this balance in a dynamic, fluid manner, and we can appeal to work by Anderson (2014) and others to reinforce this flexibility of the brain.

To connect this idea with decision-making, it is first important to note that this balancing occurs across multiple, hierarchically nested levels. This is especially important in the case of long-term decision-making. As we saw earlier in this chapter, long-term decisions will likely have distal goal states as their satisfaction conditions (e.g. whether to go on holiday). Each of the necessary sub-events that are required to obtain this goal state (e.g. book holiday online, arrange temporary visa, pack bags etc.), are likely to occur over an extended timescale. However, many of these components can be thought of as necessary components for the fulfilment of the overall decision, and thus the agent requires some means of coordinating them. This is where the above constraints can provide useful guidance, for each of these more fine-grained behaviours can be treated as a decision in their own right. Just as we saw with hierarchical cognitive control, the successive decomposition of each policy, brings us one step closer to the sorts of embodied decisions that were evident in the experiments of Cisek and Kalaska (2010) and Lepora and Pezzulo (2015). By appealing to the brain alone to explain deliberative decision-making, we may be adopting an unnecessarily restrictive and somewhat myopic perspective.

Instead, we claim that long-term decision-making may be best approached as a progressive series of coordinated embodied decisions, partially constrained by an agent's socio-cultural niche, and partially constrained by the successive series of embodied decisions, which themselves constrain future decisions by way of commitment effects. As an example, rather than seeing the behaviour of an agent booking a holiday online as a decision between whether to go to Tokyo or Lima, we could instead view their behaviour as the first in a series of successive decisions. For example, the decision to go on holiday is not made when you click the 'Book Now' button online, despite producing a large commitment effect in the sense of a financial cost. Nevertheless, as many potential holiday-goers will acknowledge, booking a holiday is no guarantee that you will end up going. Although most people follow through with their initial decision to book a holiday, due perhaps to a strong desire to have a rest and not waste their money, the initial decision in the above example is to ‘book a holiday’, not a decision to ‘go on holiday’. The distinction may appear trivial, but the manner in which we represent decisions has arguably led some in the decision sciences (e.g. neuroeconomics) to overlook the more dynamic, embodied aspects of how we actually make decisions, and instead place too much of an emphasis on the brain alone.

We should also note that in the occurrence of events, which restrict an agent from successfully obtaining their desired goal-state, we may wish to allow for the agent to claim that they nevertheless made the decision to ‘go on holiday’. In these cases, due to factors beyond their control, the agent was prevented from fulfilling their decision. Such cases do not threaten the claim that deliberative decisions often happen over extended timescales, and do not require us to posit a simple commitment mechanism that exists in some central executive region of the brain, constructing and deliberating over some set of abstract options. Instead a deliberative decision can be thought of as composed of a series of more fine-grained policies (selected through a distributed competition that here represents the decision process). These policies are coordinated and constrained by the higher-level expectations that they are the most probable sets of successive policies that will fulfil the desired goal-state. Each step in the series can be thought of as subsequently constrained by virtue of the agent’s prior (learned) belief that a significant cost (and corresponding expected feeling of regret) would be incurred were she not to go ahead with the subsequently implied actions (e.g. pack bags, head to airport etc.).

Similarly, in cases where social costs would be incurred (e.g. frustration caused from backing out of a verbally agreed arrangement), it is still possible to view the decision in an embodied manner. For example, where the choice is whether to utter the words that commit the individual to later perform some future action or not (e.g. getting married after proposing). Representations of social costs (with a key affective component) could possibly act as a further coordinating role on the unfolding of a series of policies, leading from the initial goal state through to the event that acts as the satisfaction condition. Here we see an obvious connection to work in neuroanthropology and enculturation, but also work in affective science and the role that emotions play in guiding social interactions.

Hierarchically-nested policies that reliably minimise prediction error, and are strongly associated with rewards, may over time form more abstract representations, which can be redeployed in habitual forms of decision-making.[12] These representations will still be embodied because of their grounding in the control sequences that gave rise to them initially, and will likely involve key sensorimotor regions to be redeployed or emulated (as in mental imagery). We tentatively propose that we treat decisions not as a deliberative process over a set of represented options, but as a dynamic, interactive process between brain, body and world, which is constrained by the many mechanisms pointed to in the previous sections.[13]

On this view, we are led to seek out expected action opportunities that satisfy a (possibly distal) goal-state, which is determined by the needs of our lived (and enculturated) body. Even if we are only able to account for a subset of what we wish to term ‘deliberative decisions’, this would still be a noteworthy achievement.[14]

[12] By *more* abstract, we mean something like multimodal, rather than amodal.

[13] We should note that this recommendation is made for the case of the cognitive sciences. Disciplines such as economics, which require a certain level of abstractness in order to deal with systems such as markets, may gain nothing from adding this additional complexity to their accounts.

[14] We acknowledge that this idea needs development, and is at present incomplete. There are many outstanding questions, but it is our understanding that part of the role of a doctoral thesis is to identify questions and areas for future research—this is certainly an example of such an area, and is one we intend to develop further.

Further Remarks

Over the course of this thesis we have defended an embodied account of predictive processing. Instead of adopting one specific aspect of embodied cognition (see chapter 1), the approach has favoured a more general, complementary approach. We hope this has the effect of demonstrating the wide explanatory reach of predictive processing, rather than appearing to simply ignore important debates. We acknowledge that some of these debates may eventually require a resolution (e.g. whether external artefacts are constituents of cognitive processing). Nevertheless, this thesis purposefully avoided engaging in them for a number of reasons.

Firstly, the main focus of the thesis was an exploration of decision-making, as understood from within the PP framework. Promising lines of experimental and theoretical evidence, which favour an embodied approach, were discussed, and it was shown how PP provides a suitable framework to develop this research. With this in place, we turned to the underspecification problem, and discussed how appealing to various constraints (physiological, temporal, affective and sociocultural) can help overcome the challenge. This investigation raised the possibility of a novel approach to decision-making, which seems to be well-suited to embodied PP. At present, this work is still in its infancy, and therefore it seems unwise to prematurely attempt to draw any strong conclusions regarding conceptual interpretations.

Secondly, and in order to facilitate this integration, it was necessary to demonstrate why PP is best construed as an embodied framework, rather than merely subsuming the notion within a more neurocentric framework. Therefore, focusing on the shared opposition to cognitivism and neurocentricism, was of greater importance than the potential disagreements between the various themes of embodied cognition.

As many questions have been left open, in this final chapter we would like to point to some connected questions/topics in the philosophy of science. The treatment of this literature is necessarily brief, but indicates some important discussions that we believe should be addressed when developing the framework further. Specifically, we discuss the interpretation of decision theory, and connect this with work in comparative psychology. We also highlight a debate in the philosophy of science that concerns possible meta-theoretical stances to explanation. Both of these points are offered as further avenues for investigation.

The Interpretation of Decision Theory

In chapter 3 we looked at the history behind the emergence of contemporary approaches to decision theory, and explored the origins of expected utility theory. We discussed how Bernoulli's (1738) suggestion that agent's maximise a utility function, was formalised in order to show how an agent's utility function could be derived from more basic preference relations. Unlike a utility function, these preference relations were in principle observable from the agent's behaviour. This raises an interesting question pertaining to the interpretation of decision theory that Okasha (2015) explores.

Okasha claims that there are two stances that can be taken towards the interpretation of decision theory: *mentalistic* and *behaviouristic*. The former states that credences and utilities are psychologically real, and is commonly adopted by philosophers. The latter, by contrast, takes them to be mere mathematical constructs (derivable from preference relations), and is often the perspective adopted by economists. However, in addition to this distinction, we can also ask what type of decision theory we are interested in. Recall that decision theory is an interdisciplinary framework, and is also divisible into descriptive and normative approaches. Okasha argues convincingly that if our aim is a normative account of decision theory, then the behaviouristic interpretation is indispensable, as the normative constraint of EU theory is on an agent's *preferences*, and not on their credences. He states, "it is quite wrong to view the normative content of the theory as saying that an agent should maximise expected utility relative to a psychologically real utility and credence function." (ibid., p. 17). As we are here interested in descriptive decision theory, we will accept this part of the argument and say no more on the matter.

With regards to descriptive purposes, Okasha claims that there is "no reason not to interpret credence and utility functions as psychologically real, at least to the extent that the theory fits the data." (ibid., p. 23) In connection with this, the behaviouristic interpretation is argued to be untenable on the grounds that modern science routinely goes beyond observable behaviour, and posits unobservable theoretical entities. This line of argument is also adopted by Dietrich and List (2016) who claim that rejecting the mentalistic interpretation of decision theory goes against commonly accepted naturalistic commitments to unobservable entities. They argue for this while also rejecting the claim that economics can be reduced to neuroscience, as is claimed by some neuroeconomists. The question that is of interest to us here is: does the validity of the embodied PP framework affect this claim, and should we reject the mentalistic interpretation of decision theory on the basis that embodied PP eschews the existence of abstract representations such as value or utility? To attempt an answer to this question, it is important that we follow the suggestion of Dietrich and List (ibid., p. 252), and first "distinguish clearly between the notions of mind and brain. The former is a higher-level, psychological notion, the latter a lower-level, physiological one."

Work in neuroeconomics is the best example to use in order to demonstrate this point. As is the case with many experiments in neuroscience, researchers often have to rely on behavioural or psychological data, in order to constrain their search for the neural correlates that act as measures of relevant variables. In the case of neuroeconomics, these measures are often taken to provide a neurobiological basis for the psychological construct of *utility*. Despite the many worries surrounding the isolation of some neurobiological measure, and the subsequent inference to some psychological construct such as utility (cf. Poldrack, 2006), the issue here goes beyond the mere worry of underdetermination, and even beyond the conceptual challenges raised by Anderson in chapter 5.[15] Rather, if the goal is economic insight, the concern is with *how informative* neurobiological measures can be. Few researchers working in the cognitive sciences would doubt that neurobiological measures play, or have the potential to play, some explanatory role. However, if we are interested in understanding the material substrate for psychological factors such as decision utility, is the brain the right system to focus on? Some in neuroeconomics, who argue that the brain literally computes subjective expected utility (Camerer, Loewenstein, and Prelec, 2005; Levy and Glimcher, 2012, e.g.), would seem to argue that it is. Unsurprisingly, we disagree with this position.

Consider the following:

> "[...] the operations within a mechanism are different from the phenomenon produced by the mechanism. Within a neuron, for example, neurotransmitters perform such operations as diffusing across a synapse and binding to a receptor; but the neuron itself generates action poten-

[15] It should be noted that the most recent edition of *Neuroeconomics*, edited by Glimcher and Fehr (2014b) contains a chapter dedicated to the challenges that arise from neuroimaging, which touches on the same worries expressed by Anderson.

> tials." (Bechtel, 2009, p. 560)

Here, Bechtel is drawing our attention to the profound conceptual differences that exist between variables such as the action potential of a neuron, and the operations of the mechanism that it is a part of. Why is this important? He continues:

> "The point of organizing component parts and operations into a mechanism is to accomplish something that cannot be performed by the individual components. Hence, assuming a homunculus with the same capacities as the agent in which it is posited to reside clearly produces no explanatory gain." (ibid., p. 561)

Bechtel has extensively defended a mechanistic account of explanation. He has also developed this account into a multi-level approach, which ties a mechanism's function to a) its component parts, b) the component operations or activities (as understood within a wider multi-level account), and c) their organization (Bechtel and Abrahamsen, 2005). It is in virtue of both the organisation and operation of the component parts, that a phenomena of interest is realised. Each level within the mechanism is identified with the realisation of some specific phenomena of interest, in virtue of the underlying component parts.[16]

Alternative accounts of mechanisms focus on other aspects. For example, Machamer, Darden, and Craver (2000) focus on the metaphysical nature of entities and activities, investigating how they interact to produce changes in a particular mechanism. Their characterisation includes the production of some change from the initial starting point through to its termination, which appears to acknowledge the dynamic nature of mechanistic production. However, Bechtel and Abrahamsen (2005) have

[16] 'Level' is employed here in a framework-relative manner, and should not be taken to identify some global level of analysis. See (Bechtel, 2012) for further details.

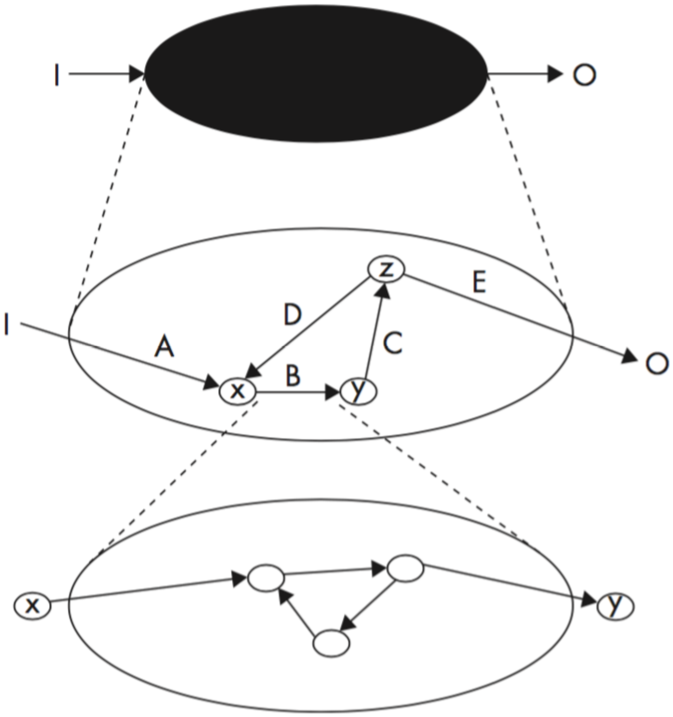


Figure 6.4: The mechanism of interest is shown in the top panel, and is responsible for performing some function. The explanation of how this function is achieved requires decomposition into its component parts and the operations they perform (indicated by uppercase letters). These operations produce changes in substrates (indicated by lowercase letters) of the system (middle panel). Explaining how one component (B) performs its operation requires treating it as a mechanism in its own right, and decomposing it per the steps above. This means acknowledging that the top-level mechanism will also be a component part in a wider, embedding mechanism. Reprinted from (Bechtel, 2012).

expressed concern with this conception, and argue that it focuses on narrowly individuated mechanisms. Instead they suggest that we should focus on a mechanism's overall contribution to the larger embedded mechanism of which it is also a part (see Figure 6.4). From this perspective an embedded mechanism is viewed as both continuously responsive to the changing conditions of the larger environment, but also effects change in a manner reminiscent of the notion of circular causality discussed in chapter 1.[17]

There are many aspects of Bechtel's account of multi-level, mechanistic explanation that are not only relevant to philosophy of science in general, but could also support a possible reconciliation for some of the conceptual challenges raised in this thesis. We shall not attempt this here, but do wish to highlight it as an area for further research, especially given the following statement that Bechtel makes:

> "Recently cognitive science has been confronted by challenges both from those advocating refocusing attention on the brain and those calling for attention to the embodied and situated aspects of cognition. The implication of the account of mechanistic explanation I have outlined is that these ought not to be viewed as challenges to cognitive science or as exclusive alternatives; both represent constructive avenues for advancing inquiry in cognitive science." (Bechtel, 2009, p. 563)

[17] It's important to note that there is an extensive debate between those who are committed to a mechanistic account of explanation, and those who employ dynamical systems theory to explain a system's behaviour. Discussing the intricacies of this debate is too tangential to the present aim (see Kaplan and Bechtel, 2011; Stepp, Chemero, and Turvey, 2011, for a representative sample). We restrict ourselves to discussion of Bechtel's multi-level account, as he has provided the most extensive discussion of the topic, and has importantly emphasised the dynamic nature of mechanisms, by arguing that the decomposition of a mechanism and modelling its dynamics can be viewed as complementary endeavours (Kaplan and Bechtel, 2011).

We began this thesis by asking which system we should be interested in when attempting to identify the mechanisms that underly some phenomena of interest. The defence of embodied cognition, throughout the course of this thesis, should hopefully make clear that the brain is an insufficient supervenience base if we are interested in phenomena such as decision-making. We have argued that decisions are behaviours made by situated agents, and that it is inappropriate to apply the terminology of decision theory to systems such as the brain (as is sometimes done in neuroeconomics). This applies equally for some psychological notions such as utility. Far from being a trivial semantic worry, a failure to identify the appropriate level of analysis can lead to significant methodological failures.

It makes little sense to isolate the components of the mechanisms identified by PP and attribute decision-making capabilities to them. Unless they are (a) organised in a particular fashion, (b) allowed to interact with other mechanisms (e.g. volume transmission), and (c) connected to the body plant, it is unlikely that we will be able to individuate their functional role at a level beyond that of cellular neuroscience (e.g. generating action potentials or transmitting chemical messengers). Decisions are dynamic behaviours taken by socially embedded agents, not the sorts of things that neurons (or neural populations) partake in. A closer examination of the mechanisms employed by PP, which utilises a multi-level mechanistic account, could help bring further rigour to the PP framework, and further solidify its embodied nature.

One may worry that the abstract functional-level characterisation of PP may present a challenge to this proposal. This is because functional level descriptions have previously been associated with a certain degree of autonomy, which separate the functional account from a description of its physical realisers. There are two responses to this worry. First of all, as we have seen in previous chapters, many have already started to provide accounts of the physical structures that could be responsible for realising the more abstract functions of PEM. Secondly, some of those who favour mechanistic explanation have also argued that functional analyses can be viewed as a *mechanistic sketches* (Piccinini and Craver, 2011). By this, they mean that a functional analysis can be treated as an initial model of some phenomena, with certain structural aspects omitted. Once these omitted aspects are filled in, a functional analysis becomes a mechanistic explanation. This proposal, with further development, could be viewed as a normative proposal for scientific discovery, and as a way of working towards unification in the cognitive sciences, by integrating multi-level mechanistic explanations.

This raises two points of discussion that are pertinent to the present enquiry. The first involves the interpretation of decision theory discussed above. The second concerns our place on a cognitive continuum with other non-human organisms.

Regarding the first, we can ask whether pursuing a unified, multi-level mechanistic account of PP threatens accounts of decision theory (whether descriptive or normative), which employ theoretical constructs such as utility or credences. We saw previously how PP eschews the use of utility functions, but is it possible to reject the pursuit of the *neural encoding* of utility functions, without rejecting the notion of utility entirely? Perhaps we could identify the notion of a utility function with a more emergent behaviour embodied in the interactions of an agent in an environment?

As Dietrich and List (2016) highlighted, there is a distinction between the physiological concept 'brain' and the higher-level concept 'mind'. If so, we could acknowledge the psychological validity of concepts such as utility, by associating them with mental states that have a wider supervenience base than merely the brain (e.g. embodied mind). Unfortunately, this suggestion is unlikely to work. As Okasha (2015, p. 15) notes, referring to utility and credence functions:

> "These entities have a specific mathematical structure, and in this respect are different from the internal states and processes that cognitive psychology usually traffics in. Even if one is happy to posit sub-personal internal

> states to explain behaviour, one might have qualms about positing internal states that satisfy certain specific measurability assumptions."

The point Okasha is making is that a real-valued utility function is measurable on a cardinal scale, and thus the underlying states that realises this function must lend themselves to this particular structure. He continues:

> "[...] if EU theory is construed descriptively, as a theory about peoples' actual preferences or choices, there seems no particular reason to interpret the theory behaviouristically rather than mentalistically [...] On the contrary, to the extent that the theory fits the data, there seems good reason to adopt a realistic attitude to the utilities and credences which the theory posits." (ibid., p. 17)

However, it appears as though proponents of PP dismiss the claim that neural structures should be descriptively modelled using utility functions, as they are not in fact the best fit for the data. Moreover, it is not clear how embodied states or behaviours could realise the specific mathematical structure required for grounding the notion of decision utility.

A related argument has also been made by Oppenheimer and Kelso (2015) who review a number of empirical findings that have questioned the descriptive validity of EUT as a model of human cognition and behaviour. Instead they call for a paradigm shift towards information processing models, which prioritise more basic cognitive building blocks, and see decision-making as recruiting distributed processes to guide action. This idea seems to be echoed in much of the work we have reviewed over the course of this thesis (e.g. Anderson, 2014; Cisek and Pastor-Bernier, 2014; Lepora and Pezzulo, 2015).

We will not worry about evaluating the competing accounts here. What is important, as Okasha notes, is that this is an empirical matter; one that PP will need to address moving forward. To argue favourably against the mentalistic interpretation of utility and credence functions requires positing an alternative, and this alternative should be able to make predictions that can be experimentally verified. The embodied decisions framework has begun this, but further work is needed. This brings us to the second concern.

A Cognitive Continuum

> "[...] much of what is true of us, even as cognitive agents, is true of us, because it is true of all vertebrates—or, at any rate, all primates. Moreover, whatever it is that is distinctive of us alone must be such that it could have been built on that common foundation with only rather modest physiological changes." (Haugeland, 2002, p. 27)

Over the course of the thesis, the claim that cognition evolved was repeatedly made, but some specific points need to be addressed. First, the discussion in this thesis has focused primarily on human cognition.[18] In short, the focus has been relatively anthropocentric in nature. This assumption may appear innocent enough, but can also lead to confusion when we attempt to transfer the account to other disciplines. For example, comparative psychology is concerned with the identification of which, if any, cognitive processes are shared by other species. However, this approach often assumes that we have first identified an appropriate notion of cognition in humans. If we haven't, then transposing this notion to non-human organisms will be problematic. Why?

[18] Although some of the empirical evidence pointed to throughout this thesis has come from neuroimaging studies on non-human organisms, the implicit assumption has been that this evidence is useful for uncovering the functional architecture of our own cognitive systems.

One response is that an inappropriate (often inflated) notion of human cognitive capacities leads to a sorting procedure whereby some creatures possess intelligent, flexible thought and behaviour, and others possess merely non-cognitive, instinctive responses to environmental stimuli. For some this sorting procedure may assist in the identification of evolutionary antecedents to the more complex processes that humans possess. Furthermore, it seems to fit with the claim made in earlier chapters that the brain evolved through descent with modification (Anderson, 2014). However, there are a number of worries that we can point to, which seem to stem from an unchecked anthropocentricism inherent in the early cognitivist approaches.

Anthropofabulation

> "[...] our own introspection about how our own minds work need not be an accurate guide to how they actually do work. Our decision making may be much simpler than our conscious self-monitoring suggests." (Barrett, 2011, p. 13)

The inference from introspection to a working hypothesis of a cognitive architecture can easily mislead. Like the rest of our cognitive capacities, our ability to introspect evolved, and has been shaped and developed over evolutionary and developmental timescales. Given this, it is fair to state that it has been selected due to some adaptive role that it played in our survival, and continues to play in the standard practice of cognitive psychology. In comparative psychology, however, it is connected to a different strategy known as 'double induction'. This process takes as its starting point the existence of an inferred psychological state in humans—not necessarily inferred solely on the basis of introspection—and subsequently infers similar states in other non-human animals based on their observable behaviour. This strategy is wrought with methodological and conceptual worries, and overlooks the magnificent variety and richness of problem-solving mechanisms that natural selection has endowed upon life. What has been adaptive for humans, may not have been so for other animals.

Attribution of mental states identified in humans to other non-human animals not only risks over-generalising, it is also ignores the possibility that our own cognitive capacities have been *overestimated*, and that non-human animals may provide conflicting evidence regarding our possibly rudimentary psychological taxonomies, rather than simply failing to meet some overblown human standard.

Over-confidence in our own ability to accurately infer the structure of our mental lives is, therefore, not only a conceptual worry for our own folk psychology. The confabulation of our own mental abilities, combined with the anthropomorphic bias, leads to us stacking the odds against non-human animals by utilising a potentially mistaken, anthropocentric yardstick with which to measure their psychological capacities. This leads to the observation of a number of methodological biases, which Buckner (2013, p. 861) captures in his slightly awkward term 'anthropofabulation'. He defines this as the "tendency to set the criteria for psychological capacities to an artificially-inflated sense of what humans can or routinely do." The biases that lead to anthropofabulation are a) taxonomic anthropocentricism (the *anthropo*-morphic aspect) and b) an exaggeration about typical human cognitive performance (the con-*fabulation* aspect).

The anthropomorphic part of the bias, in which it is argued that we should avoid the attribution of purported human psychological capacities to non-human organisms with insufficient evidence, was made famous by Conway Lloyd Morgan. To caution against such a methodological and taxonomic bias, he put forward the following claim, which has subsequently become known as *Morgan's Canon*:

> "[...] in no case may we interpret an action as the outcome of the exercise of higher psychological processes, if it can be fairly interpreted in terms

> of processes which stand lower in the scale of psychological evolution and development." (Morgan, 1894, p. 59)

Both Barrett (2011) and Buckner (2013) argue that too much research in comparative psychology has ignored Morgan's canon (and sometimes misinterpreted it), leading to the mistaken endowment of psychological capacities on non-human animals, which they don't possess, and more importantly don't need. Furthermore, as Barrett (2011, p. 4) worries, "it promotes the idea that other organisms are interesting only to the degree that their capacities and abilities match our own."

Spotting this bias in experimental work is often hard, and made worse when psychology papers are published with insufficient specification of the conditions necessary or sufficient to obtain the results claimed, in turn leading to the sorts of reproducibility crises that were highlighted in a recent Science paper published by the Open Science Collaboration (2015). If many experimental studies are failing to meet the requirement of reliable, statistically significant effects that can survive possible theory change (Hacking, 1983), then spotting the first of these biases may be difficult.

A particularly salient example is explored by Heyser and Chemero (2012), who focused on object exploration studies in mice. Increasingly popular since its introduction in 1988, the setup of an object exploration study allows experimenters to test various effects on memory, using the premise that long-term exposure to familiar environments results in habituation and decreased exploration—as novelty decreases, so too does exploration. Conversely, the introduction of novel objects to the environment should result in increased exploratory activity. The validity of this premise allows for experimental manipulation of independent variables that range from genetic factors, effects of drugs on learning, and the role of particular brain areas. Rather than being interested in the effect of causally intervening on the physiological variables of the mice, however, Heyser and Chemero studied a different effect that the environment had on the results: the choice of objects by the experimenter.

By considering the choice of objects in terms of what actions they afforded the mice, as opposed to the typical objective properties highlighted by the experimenter (e.g. colour, size etc.), Heyser and Chemero found that the habituation effect could be significantly modulated according to whether the object afforded the mice a touching or climbing relation—two *organism-relative* properties. If a novel object was introduced that only afforded a touching relation to the mice, the time taken to habituate was significantly shorter than if the object was one that the mice could climb on. In other words, mice explored objects for a duration of time correlated with the type of *action afforded* to the mouse, not whether it had some discernible property identified by the experimenter (e.g. colour, shape or size). This study demonstrates how a failure to recognise the importance of anthropocentric biases, can easily creep in and effect the reliability of the results uncovered by a study. In this instance, the assumption was that the novelty of the properties perceived by the mice, were the same ones identified by the experimenters. Chemero (2011) argues that a more embodied, action-oriented approach could help overcome some of these challenges.

Furthermore, in a review of 116 articles published in neuroscience articles alone, Chemero (ibid.) reports that 44% of the articles, “gave little or no information concerning the specific objects that were given for exploration”, and of the remaining 56%, 28% of these used objects that offered nonequivalent affordances, e.g. objects that were climbable and some that were non-climbable. Far from simply falling prey to an anthropocentric bias, by selecting objects based on the relation their properties have to humans, rather than the test subjects (i.e. mice), almost half failed to even meet the standard of reproducibility that is so fundamental to the scientific method.

Avoiding anthropofabulation does not mean pursuing independent psychologies for each species, so that we end up with a human psychology, a dolphin psychology, a giraffe psychology and so on. This would of course be undesirable, and the cost incurred would be never knowing which capacities we share with animals (and each other), as well as a potentially improved understanding of their function in our own lives. Instead, as Barrett (2015) argues, what we need is a "better kind of continuity", one which can help inform questions concerning the extent to which animals are similar to us, and indeed the extent to which we are similar to animals. The research explored in this thesis argues in favour of an embodied approach, but also emphasises novel approaches to constructing psychological taxonomies (see chapter 5), and a novel framework that offers a wide explanatory scope (Clark, 2016b; Friston, 2010). It seems each of these elements is appropriately informed by evolutionary theory, and thus may be able to progress towards an ontology for the cognitive sciences that is able to account for more than merely human behaviour.

Behaviour and Cognition

> "We have many vocabularies for describing nature when we regard it as mindless, and we have a mentalistic vocabulary for describing thought and intentional action; what we lack is a way of describing what is in between. This is particularly evident when we speak of the "intentions" and "desires" of simple animals; we have no better way to explain what they do." (Davidson, 1999, p. 11)

Achieving a greater degree of continuity between ourselves and other non-human animals requires of necessity, that we adopt an evolutionary perspective; one which acknowledges the integral role that our natural environment has had in shaping our cognitive development. Though we are likely to find similar neural structures in our evolutionary ancestry, and nearby cladistic neighbours, bodies have been around long before brains were on the scene.

This is an important consideration in embodied theories of cognition, which emphasise the fundamental role that our bodies have in constituting our cognitive processes over evolutionary and developmental timescales. For example, as Barrett (2011, p. 37) nicely points out:

> "[...] as Vygotsky conceived of it, a child's mental processes are not the source and cause of her behavior in the world; rather a child's behavior in the world is the source and cause of what eventually ends up in her head: the exact reverse of what most modern psychology would have us think."

We see a similar acknowledgement in neuroanthropology and enculturation (see chapter 6). Given that our bodies are also products of our evolutionary history, emphasis should be placed on the way that our bodies engage with their environment, and may have engaged with our earlier environments. This requires a greater understanding of not only our own behaviour, but the behaviour of other animals.

Statements such as the above, however, have led some to criticise embodied accounts as conflating the notion of cognition with behaviour. As we saw in chapter 1, by rejecting behaviourism, cognitive science made it acceptable to posit inner states that functioned as the *cause* of an organism's behaviour, but which were not identical with the behaviour itself. Aizawa (2015) argues that by conflating these two notions, the embodied cognition theorist has lost any appeal to explanatory force provided by positing inner cognitive states. As an example, he highlights the following quote by Chemero (2011, p. 212):

> "I take it that cognition is the ongoing, active maintenance of a robust animal-environment system, achieved by closely co-ordinated perception and action."

Aizawa argues that as a definition of cognition, the above fails to distinguish how cognition is different from behaviour. Moreover, as the behaviour of an organism is necessarily embodied, if cognition is simply behaviour redefined, this trivialises the notion of embodied cognition. Responding to this worry is important, especially given the emphasis on the interpretation of decision theory at the start of this chapter.

Aizawa (2015) appears to acknowledge that construing Chemero's position as the equation of cognition with behaviour is perhaps too unsympathetic a reading. He seems to pay lip service to a more charitable interpretation by situating the above quotation in its full context (see below). However, the supposed 'context' that Aizawa provides omits relevant aspects of the original paragraph. To see why this is important, here is the paragraph from (Chemero, 2011, p. 212) (in full), with emphasis added to the section that Aizawa (2015) actually quotes:

> "Adams and Aizawa (2008) argue that defenders of the sort of view of cognition that I am defending here need to give a definition of "cognition." In comments on a draft of this chapter, Ken Aizawa suggests that I am defining "cognition" as "intelligent behavior," which definition [sic] Aizawa points out is almost surely circular. I do not intend such a definition, and *I disagree that proponents of radical embodied cognitive science actually require a definition of "cognition." That aside, I will say a few things about what I mean by "cognition." I take it that cognition is the ongoing, active maintenance of a robust animal-environment system, achieved by closely coordinated perception and action. This understanding of the nature of cognition is intended to reflect claims by radical embodied cognitive scientists in philosophy, psychology, AI, and artificial life.* (See Maturana and Varela 1980; Reed 1996; Beer 2003; Thompson 2007.) Note, finally, that these brief remarks are not intended to supply a set of necessary and sufficient conditions, or criteria for what Adams and

> Aizawa call the “mark of the cognitive.” In chapters 6 and 7, I lay out a Gibsonian theory of perception, action and cognition. This also does not provide criteria for the “mark of the cognitive.” There is no such thing.”

Without speculating on Aizawa’s intention for selectively choosing to include only the middle portion of this paragraph, a few things can be said. Firstly, Chemero is explicit that the very thing that Aizawa is considering as a possible interpretation, and in turn criticising, is *not what he intends* (i.e defining cognition as behaviour). Far from undermining Chemero’s position, it undermines Aizawa’s own critique, when he later argues:

> “If we read Chemero as offering a stipulative definition, then his account is misleading, it marks no theoretical advance, and it trivializes the hypothesis that cognition is embodied.” (ibid., p. 764)

Sure, we could read Chemero in this way, but he explicitly states that it is the wrong interpretation, and thus fails to take his framework on its own terms. Aizawa’s omission of the latter part of the above paragraph is also interesting. Here, Chemero states “these brief remarks are not intended to supply a set of set of necessary and sufficient conditions, or criteria for what Adams and Aizawa call the “mark of the cognitive.” [...] There is no such thing.” Unfortunately, at this point, it appears that the two authors are simply talking past each other. Aizawa’s commitment to the cognitivist paradigm, appears to prevent him from construing the terms adopted by Chemero in the manner they were intended. This itself is understandable, and may represent a simple matter of incommensurability between competing paradigms—in this instance ‘cognitivism’ and ‘radical embodied cognitive science’. For example, Aizawa’s claim that Chemero’s terminological shift of the term cognition to mean something like behaviour (as understood qua cognitivism) would lead to a trivialisation of the notion of embodiment, goes as follows:

> "If one understands 'cognitive processes' as behavioral processes, then of course, "cognitive processes" are typically realized in the brain, body, and world. Behavioral processes are typically realized in the brain, body, and world. That is just the consensus twentieth-century view. It is quite far from offering a radical embodied cognitive science; by itself, it is completely pedestrian twentieth-century cognitive science. What would be radical would be the conclusion that cognition understood as a particular kind of computation over representations is embodied. What would be surprising would be to find that what has commonly been thought to occur only within the brain in fact occurs in an unexpectedly larger space." (ibid., p. 762)

The first claim about *what would be radical* acknowledges an embodied perspective attributable to the likes of Andy Clark. However, it is clear that this is not what Chemero (2011) intends given the great lengths he goes to, in order to defend a dynamical, anti-representationalist view. However, the second claim is something that Chemero (and others who defend the replacement theme (see chapter 1)) definitely offers an account of. By showing how, we can both reject Aizawa's criticism that cognition is simply behaviour in the traditional sense, and also see how the alternative proposals retain their explanatory force.

Although Chemero (ibid.) provides relevant examples, a more striking example, which is consistent with Chemero's view, is offered by Louise Barrett (2011) in a chapter of her book aptly titled 'The Implausible Nature of *Portia*'. *Portia* is a genus of jumping spiders referred to as salticids, which have an incredible ability to seemingly stalk their prey, by appearing to plan complex routes, and use clever methods of distraction and deception. Most importantly, *Portia* have brains no bigger than a pinhead, and thus appealing to complex inner neural representations, as Barrett puts it, seems "implausible". Although it may be simpler to ascribe intentional psychological states, and speak of Portia *as if* it were ‘planning’ and ‘deceiving’, the simplicity of their brains demands a more biologically plausible account—one which Barrett is happy to offer.

Consider the following. When stalking prey that may be situated in a hard to reach location, *Portia* spiders appear to ‘scan’ their surroundings, acting as if they were considering and planning alternative routes. However, planning is often considered to be an example of a ‘representation-hungry’ process, which implies that an agent must construct an internal representation of the goal-state, which can be consulted during the period in which the goal is out of sight (i.e. during planning and decision-making). How does the small brain of a *Portia* spider achieve this remarkable feat?

To answer this, it is important to consider how things appear from the perspective of the *Portia* spider, and to avoid anthropofabulating. Barrett puts it as follows:

> “During scanning, the spider gives every impression of weighing up the routes for their suitability, planning its way around obstacles, and then setting off once it has worked out a suitable route. But is this really what the spiders are doing? Just because it looks like planning, in ways that make sense to us, doesn’t mean that the spiders are necessarily operating in that way.” Barrett (ibid., p. 62)

To consider how the perspective of the spider appears, Barrett provides a careful examination of the physiology of the salticid’s eyes. Salticids have eight types of eyes evenly spread around the front part of their bodies. Two of these eyes are considered their principal eyes, which face forward and can detect fine detail and colour. The other six, so called “secondary” eyes detect movement in lesser detail. The construction of their primary eyes can be thought to function as a narrowly focused magnifying lens, or like a pair of binoculars. However, unlike in normal binoculars, where focus is attained by altering the refraction of light, salticid's eyes are not constructed in a manner that allows this. Instead, the salticid's eyes are composed of multiple layers that splits the light into four layers, each with different levels of focus. Additionally, the saltcid's eyes are active, allowing the spider to perform the aforementioned scanning behaviour. Barrett claims that this means the salticid's principal eyes function a lot like a torch in a darkened room does, focusing selectively on small regions of space.

> "In this way, the eye itself acts a filter that excludes irrelevant information, a task that would otherwise have to be achieved by neural processing; spiders can compensate for their small brains by having their eyes do most of the work." (ibid., p. 63)

Combined with the lower-resolution secondary eyes, which detect movement quickly, the *Portia* spider is well-adapted to hunting, and at the same time avoiding other predators. But what about the apparent planning behaviour?

Once the functions of the *Portia* spider's eyes have been acknowledged, the scanning behaviour can be revisited in the context of the spider hunting its prey. This is exactly what Tarsitano and Andrew (1999) did, in a number of ingenious experiments. The apparatuses shown in Figure 6.5 were set up in order to test the hunting behaviour of *Portia*. The spider was initially placed on the starting platform in the middle, with a prey spider fixed to the lure. The scanning behaviour was observed across a number of varied setups (i.e. some with multiple complete choices leading to the prey, and some with gaps in either the left or right ramps). In the case of the two complete routes (a), it was found that no preference was given to either route. However, in cases where one of the routes to the prey was prevented by a gap, the Portia spider demonstrated interesting "scanning" behaviour. The distribution of their scanning is not split equally across the routes, as in the first case, but is con-

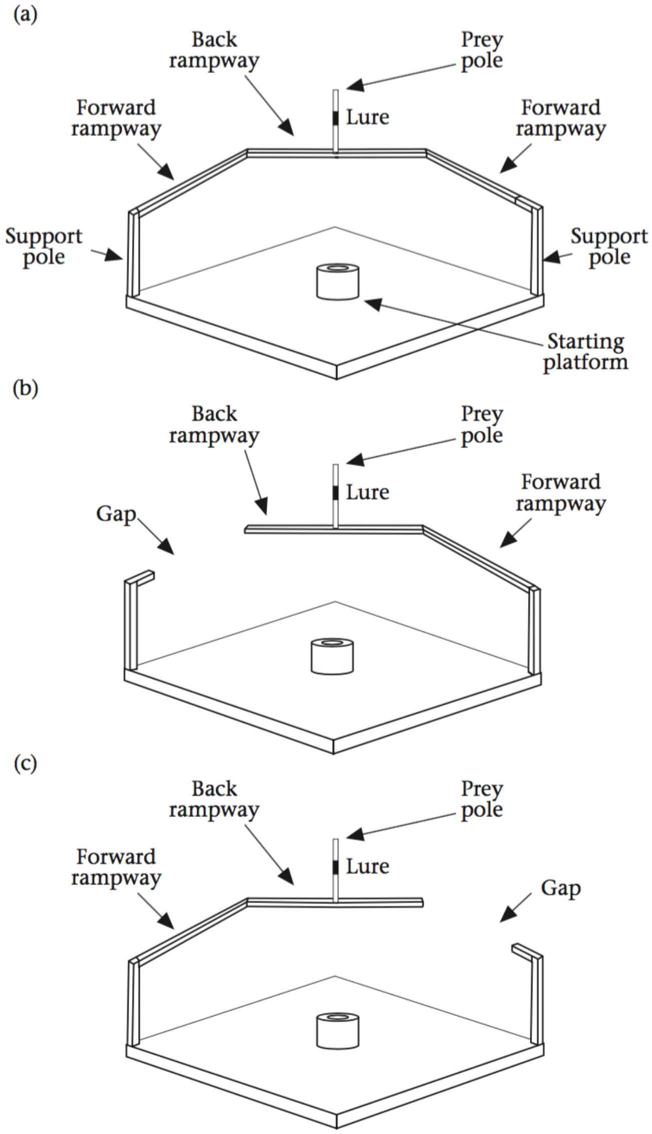


Figure 6.5: A number of detour apparatuses designed to observe the planning behaviour of *Portia*. (a) Two complete routes to prey. (b and c) one of the routes obstructed by a gap. Reprinted from (Tarsitano and Andrew, 1999, p. 201).

centrated initially on the gap in the wrong route. Over time, the scanning behaviour begins to fixate on the correct ramp that leads the prey, and eventually on the support pole of the respective ramp, shortly before movement occurs. By analysing the patterns of scanning behaviour, Tarsitano and Andrew (ibid.) found that the crucial factor appears to be whether the spider detects an unbroken horizontal line in its visual field. If the horizontal line is unbroken, the scanning behaviour proceeds away from the prey until the support pole is perceived. However, if the line is broken, or if the gap is perceived, the spider returns its scanning back to the prey, and switches directions until a complete path is detected. This pattern of behaviour thus seems to rely on a very simple feedback mechanism that involves two rules emphasised by Barrett (2011, p. 67):

> "If the end of a horizontal feature is detected, then change scanning direction," and "If the end of a horizontal feature is not detected, then continue to turn in the direction of the previous turn."

The point of this is to demonstrate that, given the basic neural system of the *Portia* spider, what initially appears to require an overly complex (and somewhat implausible) account can be successfully accommodated by providing a more in-depth understanding of the physiological characteristics of the organism's body and the properties of its environment. The scanning behaviour is not for the purpose of building a representation, but is much more simple and should be thought of as something akin to detecting. This is not to deny that inner cognitive processes are also necessary—Barrett acknowledges the importance of simple feedback mechanisms that are responsible for coordinating sensorimotor processes. Instead, we should treat it as a response to Aizawa that highlights how something initially thought to be a prototypical cognitive process (i.e. planning), may in fact be a product of simpler brain-body-world interactions, and thus not attributable solely to inner cognitive processes. The fact that these coupled components work so closely with one another to achieve the cognitive task of “planning” a route, is reason not to strictly demarcate the inner from the outer, and reserve the term ‘cognitive’ solely for the inner processes—if it even makes sense to try to delineate a strict boundary in the first place.

Perhaps this is a more charitable interpretation for why Chemero wishes to avoid providing a stipulative definition of cognition, and why he rejects the existence of the ‘mark of the cognitive’. In the case of the *Portia* spider and its hunting behaviour, it truly seems arbitrary to separate its neural processes from the closely coupled body-environment system. We are free to adopt the intentional idiom, and describe what the spider does as ‘scanning’ or ‘planning’, but we should be careful to acknowledge that it is only able to achieve this as a result of distributed mechanisms. And if the adamant cognitivist still wishes to cling to the more traditional definition of symbol-processing in spite of this, there is always the worry explored in Chapter 5 by Anderson and others. The pursuit of contentful inner symbols that represent anything like the sorts of objects required for a more traditional psychological gloss are unlikely to be found in the massively-recurrent networks of the interactive brain.

Therefore, returning again to the issue raised by Dietrich and List (2016) at the start of this chapter, we can retain mentalistic attributes, and intentional psychological states, because one of the fundamental lessons of embodied cognition (regardless of whether you’re a representationalist or not) is that the mind is not the brain. The failure of neuroeconomics is not the attempt to supplement explanations of economic behaviour by appealing to neuroscience (this is a praiseworthy endeavour). It is to incorrectly transpose intentional psychology into the ontological commitments of neuroscience. Likewise, PP can make do without value functions when describing the activities of the brain, but it may want to exercise caution when trying to scale-up to higher-levels of description. One suggestion is to adopt an explanatory pluralism, and to take as its primary focus, not merely predictive brains, but brain-body-world systems.

Explanatory Pluralism

> "Given computation determines perception and cognition, perception and cognition happen in the brain. The mind can then be understood in internalist, solipsistic terms, throwing away the body, the world and other people." (Hohwy, 2014, p. 7)

To end, we return once again to the matter of whether PP should be best viewed as an embodied framework. Rather than repeating the reasons for adopting the affirmative position, we shall consider the alternative, as expressed by Hohwy's brand of internalist PP. Hohwy is certainly no cognitivist in the sense that we first discussed back in chapter 1. Nevertheless, as the above quotation indicates, he does adopt a neurocentricism that is opposed to the embodied account we have been defending over the course of this thesis. In fact he delineates his version of neurocentricism for us very specifically:

> "[...] the mind begins where sensory input is delivered through exteroceptive, proprioceptive, and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord." (ibid., p. 18)

Insofar as cognitive scientists are interested in understanding the mind, the validity of a statement such as this entails a metatheoretical approach known as methodological solipsism (Fodor, 1980), and sometimes as methodological individualism (Chemero and Silberstein, 2008). More specifically, methodological solipsism is concerned with what is the object of analysis for scientific enquiry. In the case of the cognitive sciences this translates to questions such as: is the object of the cognitive sciences the brain, the organism, or the coupled brain-body-world system? Hohwy's adoption of this stance is evident in the fact that the neurocentric mechanisms of PEM are taken to subsume other domains of enquiry (e.g. embodied cognition), and provide the relevant explananda for understanding cognition and thus behaviour. Instead, the inferentially-secluded brain is able to do all of the explanatory work:

> "PEM says that prediction error minimization is the only principle for the activity of the brain [...] This is a very ambitious theory. If this is all the brain does, then perception, action, attention, and all other mental processes, must come down to prediction error minimization." (Hohwy, 2014, p. 2)

Methodological solipsism commits one to a particular epistemic goal of scientific explanation: aiming to uncover a single set of underlying, unifying principles to account for a diversity of phenomena and theories, which are themselves nothing more than approximations or derivations of the underlying theory (Dale, 2008). Advocates of this metatheoretical stance, should therefore aim to uncover these unificatory principles in their research. Although Hohwy is not alone in advocating the unifying power of PP (also see Clark, 2013c; Friston, 2010), his account is set apart by the appeal to an underlying computational principle (PEM) that explicitly rules out alternative research strategies such as embodied cognition, except inasmuch as its claims can be accommodated within the strictures of a neurocentric framework (Hohwy, 2014, p. 17). This is problematic, as we can see if we return to the decision tree from Chapter 1 and replace 'cognitivism' with neurocentric PP.

As is shown in Figure 6.6, given the ambitious claims of neurocentric PP, it certainly appears as though it is aiming to encroach on the subject matter of embodied cognition (i.e. cognition and behaviour). Furthermore, it should be evident by now

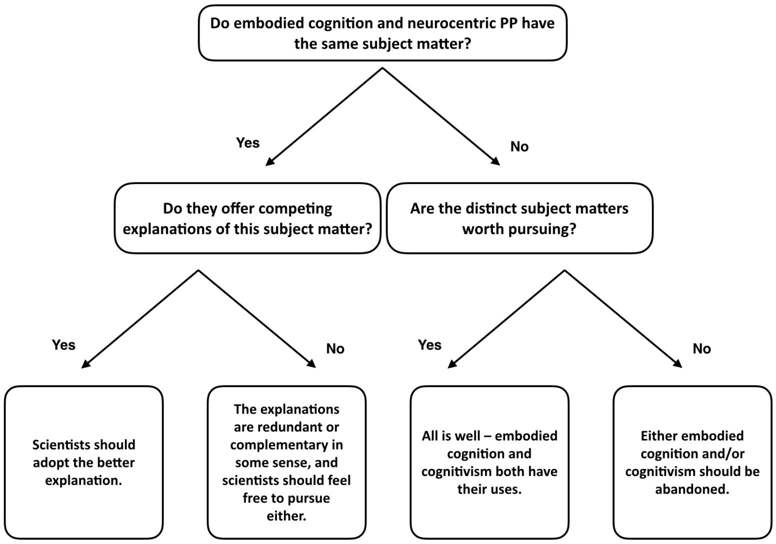

Do embodied cognition and neurocentric PP have the same subject matter?
Yes
No
Do they offer competing explanations of this subject matter?
Are the distinct subject matters worth pursuing?
Yes
No
Yes
No
Scientists should adopt the better explanation.
The explanations are redundant or complementary in some sense, and scientists should feel free to pursue either.
All is well – embodied cognition and cognitivism both have their uses.
Either embodied cognition and/or cognitivism should be abandoned.

Figure 6.6: The decision tree from Chapter 1 modified to contrast embodied cognition with neurocentric PP. Adapted from (Shapiro, 2011, p. 201)

that they offer competing explanations of the subject matter, insofar as they adopt different attitudes to the acceptance of methodological solipsism and the function of inner computational processes. Therefore, either the claims of embodied cognition are subsumed within the neurocentric PP framework, and consequently lose much of their explanatory worth and philosophical significance, or we acknowledge that the two methodological pursuits are in conflict, and "adopt the better explanation".

However as a metatheoretical approach, methodological solipsism is not the only option. In fact, it is increasingly common for philosophers to argue in favour of an explanatory pluralism (Chemero and Silberstein, 2008; Dale, 2008), at least as far as the cognitive sciences are concerned. For example, we saw at the start of this chapter how Bechtel and others have argued that some cognitive processes may be best explained mechanistically, and how mechanistic and dynamical approaches can be considered complementary (see also Clark, 1997b). Bechtel has even proposed a way of determining when a system will be amenable to a mechanistic explanation, and when it will not (Bechtel and Richardson, 2010). Again, this is not to deny that the debate between those who defend mechanistic explanation, and those who favour dynamic explanations is settled. However, as far as a strategy for discovery goes, it seems sensible to adopt an explanatory pluralism, as it forces us to consider how decision-making is not simply a product of our brains, but is rather constituted over time by the many ways that our brains, body and world interact. A staunchly internalist account of PP is unlikely to accept this, but we *predict* that an embodied PP will be able to embrace this meta-theoretical approach to develop and evolve into an even more powerful framework for the cognitive sciences—we hope that the corresponding *prediction error* is small.

Bibliography

Adams, Rick A, Stewart Shipp, and Karl Friston (2013). “Predictions not commands: active inference in the motor system”. In: *Brain Structure and Function* 218.3, pp. 611–643.

Agnati, Luigi F et al. (2010). “Understanding wiring and volume transmission”. In: *Brain Research Reviews* 64.1, pp. 137–159.

Aizawa, Ken (2015). “What is this cognition that is supposed to be embodied?” In: *Philosophical Psychology* 28.6, pp. 755–775.

Akins, Kathleen (1996). “Of Sensory Systems and the “Aboutness” of Mental States”. In: *The Journal of Philosophy* 93.7, pp. 337–372.

Allais, M. (1953). “Le Comportement de l’Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l’Ecole Americaine”. In: *Econometrica* 21.4, pp. 503–546.

Anderson, John R (1991). “Is human cognition adaptive?” In: *Behavioral and Brain Sciences* 14.03, pp. 471–485.

Anderson, M L and G Rosenberg (2008). “Content and action: The guidance theory of representation”. In: *Journal of Mind and Behavior* 29.1, pp. 55–86.

Anderson, Michael L (2008). “Circuit sharing and the implementation of intelligent systems”. In: *Connection Science* 20.4, pp. 239–251.

— (2014). *After Phrenology: Neural Reuse and the Interactive Brain.* Cambridge: MIT Press.

— (2015). “Mining the Brain for a New Taxonomy of the Mind”. In: *Philosophy Compass* 10.1, pp. 68–77.

Anderson, Michael L and Marcie Penner-Wilger (2012). “Neural reuse in the evolution and development of the brain: Evidence for developmental homology?” In: *Developmental Psychobiology* 55.1, pp. 42–51.

Anderson, Michael L and Luiz Pessoa (2011). “Quantifying the diversity of neural activations in individual brain regions”. In: *Proceedings of the 33rd annual conference of the cognitive science society*, pp. 2421–2426.

Ashraf, Nava, Colin F Camerer, and George Loewenstein (2005). “Adam Smith, behavioral economist”. In: *The Journal of Economic Perspectives* 19.3, pp. 131–145.

Badcock, Paul B, A Ploeger, and N B Allen (2016). “After phrenology: Time for a paradigm shift in cognitive science”. In: *Behavioral and Brain Sciences* 39, pp. 10–11.

Baddeley, A (1992). “Working memory”. In: *Science* 255.5044, pp. 556–559.

Ballard, D H (1991). “Animate vision”. In: *Artificial intelligence* 48.1, pp. 57–86.

Ballard, D H, M M Hayhoe, and P K Pook (1997). “Deictic codes for the embodiment of cognition”. In: *Behavioral and Brain Sciences* 20.4, pp. 723–767.

Bar, Moshe, ed. (2011a). *Predictions in the brain: Using our past to generate a future.* Oxford Scholarship Online.

— (2011b). “The Proactive Brain”. In: *Predictions in the brain: Using our past to generate a future.* Ed. by Moshe Bar. Oxford Scholarship Online. DOI: 10.1093/acprof:oso/9780195395518.001.0001.

Bargmann, Cornelia I (2012). “Beyond the connectome: How neuromodulators shape neural circuits”. In: *Bioessays* 34.6, pp. 458–465.

Barkow, Jerome H, Leda Cosmides, and John Tooby (1995). *The adapted mind: Evolutionary psychology and the generation of culture.* Oxford: Oxford University Press.

Barrett, L F and M Bar (2009). “See it with feeling: affective predictions during object perception”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1521, pp. 1325–1334.

Barrett, Lisa Feldman and W Kyle Simmons (2015). “Interoceptive predictions in the brain”. In: *Nature Reviews Neuroscience* 16.7, pp. 419–429.

Barrett, Louise (2011). *Beyond the Brain: How Body and Environment Shape Animal and Human Minds.* Princeton University Press.

— (2012). “Why Behaviorism Isn’t Satanism”. In: *The Oxford Handbook of Comparative Evolutionary Psychology.* Ed. by Todd Shackelford and Jennifer Vonk. Oxford: Oxford University Press.

— (2015). “A Better Kind of Continuity”. In: *The Southern Journal of Philosophy* 53.Spindel Supplement, pp. 28–49.

Barsalou, Lawrence (2008). “Grounded Cognition”. In: *Annual Review of Psychology* 59.1, pp. 617–645.

Barsalou, Lawrence W (1999). “Perceptions of perceptual symbols”. In: *Behavioral and Brain Sciences* 22.04, pp. 637–660.

Basso, D (2013). “Planning, prospective memory, and decision-making: three challenges for hierarchical predictive processing models”. In: *Frontiers in Psychology* 3.623, pp. 1–2.

Bastos, A M et al. (2012). “Canonical microcircuits for predictive coding”. In: *Neuron* 76.4, pp. 695–711.

Bechara, Antoine, Hanna Damasio, and Antonio R Damasio (2000). “Emotion, Decision Making and the Orbitofrontal Cortex”. In: *Cerebral Cortex* 10.3, pp. 295–307.

Bechtel, William (2009). “Constructing a philosophy of science of cognitive science”. In: *Topics in Cognitive Science* 1.3, pp. 548–569.

— (2012). “Reducing Psychology while Maintaining its Autonomy via Mechanistic Explanations”. In: *The Matter of the Mind.* Ed. by Maurice Schouten and Huib Looren de Jong. Philosophical Essays on Psychology, Neuroscience and Reduction. John Wiley & Sons, pp. 172–198.

Bechtel, William and Adele Abrahamsen (2005). “Explanation: A mechanist alternative”. In: *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, pp. 421–441.

Bechtel, William and Robert C Richardson (2010). *Discovering Complexity.* Decomposition and Localization as Strategies in Scientific Research. Cambridge: MIT Press.

Beer, Randy (2000). “Dynamical Approaches to Cognitive Science”. In: *Trends in Cognitive Sciences* 4.3, pp. 91–99.

Bernoulli, D (1738). “Exposition of a new theory on the measurement of risk.” In: *Econometrica [1954]* 22, pp. 23–26.

Bickhard, Mark H (2015). “Toward a Model of Functional Brain Processes I: Central Nervous System Functional Micro-architecture”. In: *Axiomathes* 25.3, pp. 217–238.

Binmore, Ken (2008). *Rational Decisions.* Princeton University Press.

Boroditsky, Lera and Alice Gaby (2010). “Remembrances of times east absolute spatial representations of time in an Australian aboriginal community”. In: *Psychological Science* 21.11, pp. 1635–1639.

Botvinick, Matthew (2008). “Hierarchical models of behavior and prefrontal function”. In: *Trends in Cognitive Sciences* 12.5, pp. 201–208.

Botvinick, Matthew and Marc Toussaint (2012). “Planning as inference”. In: *Trends in Cognitive Sciences* 16.10, pp. 485–488.

Brooks, Rodney A (1991). “Intelligence without representation”. In: *Artificial intelligence* 47.1, pp. 139–159.

Bruineberg, J and E Rietveld (2014). “Self-organization, free energy minimization, and optimal grip on a field of affordances”. In: *Frontiers in Human Neuroscience* 8.653, pp. 1–14.

Buckner, Cameron (2013). “Morgan’s Canon, meet Hume’s Dictum: avoiding anthropofabulation in cross-species comparisons”. In: *Biology and Philosophy* 28.5, pp. 853–871.

Burr, Christopher and Max Jones (2016). “The body as laboratory: Prediction-error minimization, embodiment, and representation”. In: *Philosophical Psychology* 29.4, pp. 586–600.

Calvo, Paco and Toni Gomila (2008). “Directions for an Embodied Cognitive Science: Toward an Integrated Approach”. In: *Handbook of Cognitive Science.* Ed. by Paco Calvo and Toni Gomila. An Embodied Approach. Elsevier, pp. 1–25.

Camerer, C, G Loewenstein, and D Prelec (2005). “Neuroeconomics: How neuroscience can inform economics”. In: *Journal of Economic Literature* 43.1, pp. 9–64.

Caplin, Andrew and Mark Dean (2008). “Dopamine, Reward Prediction Error, and Economics”. In: *The Quarterly Journal of Economics* 123.2, pp. 663–701.

Caplin, Andrew and Paul W Glimcher (2014). “Basic Methods from Neuroclassical Economics”. In: *Neuroeconomics.* Ed. by Paul W Glimcher and Ernst Fehr. 2nd ed. Decision Making and the Brain. Academic Press, pp. 3–17.

Casasanto, Daniel (2009). “Embodiment of abstract concepts: Good and bad in right- and left-handers.” In: *Journal of Experimental Psychology* 138.3, pp. 351–367.

Chater, Nick et al. (2010). “Bayesian models of cognition”. In: *WIREs Cogn Sci* 1.6, pp. 811–823.

Chemero, Anthony (2011). *Radical Embodied Cognitive Science.* Cambridge: MIT Press.

Chemero, Anthony and Michael Silberstein (2008). "After the Philosophy of Mind: Replacing Scholasticism with Science". In: *Philosophy of Science* 75.1, pp. 1–27.

Chiao, J and T Harada (2008). "Cultural neuroscience of consciousness: From visual perception to self-awareness". In: *Journal of Consciousness Studies* 15.10-11, pp. 58–69.

Church, Alonzo (1936). "An unsolvable problem of elementary number theory". In: *American Journal of Mathematics* 58.2, pp. 345–363.

Churchland, P S, V S Ramachandran, and Terrence J Sejnowski (1994). "A Critique of Pure Vision". In: *Large-scale neuronal theories of the brain.* Ed. by Christof Koch and Joel L. Davis. Cambridge: MIT Press.

Chwalisz, K, E Diener, and D Gallagher (1988). "Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured". In: *Journal of Personality and Social Psychology* 54, pp. 820–828.

Cisek, P (2007). "Cortical mechanisms of action selection: the affordance competition hypothesis". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1485, pp. 1585–1599.

Cisek, P and A Pastor-Bernier (2014). "On the challenges and mechanisms of embodied decisions". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.20130479, pp. 1–14.

Cisek, Paul (2012). "Making decisions through a distributed consensus". In: *Current Opinion in Neurobiology* 22.6, pp. 927–936.

Cisek, Paul and John Kalaska (2010). "Neural mechanisms for interacting with a world full of action choices". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 33, pp. 269–298.

Clark, Andy (1997a). *Being There: Putting Brain, Body, and World Together Again.* Cambridge: MIT Press.

— (1997b). "The Dynamical Challenge". In: *Cognitive Science* 21.4, pp. 461–481.

— (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* Oxford: Oxford University Press.

— (2013a). *Mindware: An introduction to the Philosophy of Cognitive Science.* 2nd ed. Oxford: Oxford University Press.

— (2013b). "The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science")". In: *Frontiers in Psychology* 4.270, pp. 1–9.

— (2013c). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36.03, pp. 181–204.

— (2015). "Embodied Prediction". In: *Open MIND* 7.T. Ed. by T. Metzinger and J. M. Windt, pp. 1–21. DOI: 10.15502/9783958570115.

— (2016a). "Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil". In: *Nous*, pp. 1–27. DOI: 10.1111/nous.12140.

— (2016b). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Clark, Andy and David Chalmers (1998). "The Extended Mind". In: *Analysis* 58.1, pp. 7–19.

Clark, Andy and Josefa Toribio (1994). "Doing without representing?" In: *Synthese* 101.3, pp. 401–431.

Cocchi, L et al. (2013). "Dynamic cooperation and competition between brain systems during cognitive control". In: *Trends in Cognitive Sciences* 17.10, pp. 493–501.

Colombetti, G (2008). “The Somatic Marker Hypotheses, and What the Iowa Gambling Task Does and Does not Show”. In: *The British Journal for the Philosophy of Science* 59.1, pp. 51–71.

Colombetti, Giovanna (2013). *The feeling body: Affective science meets the enactive mind*. Cambridge: MIT Press.

Colombo, M and P Seriès (2012). “Bayes in the Brain—on Bayesian Modelling in Neuroscience”. In: *The British Journal for the Philosophy of Science* 63, pp. 697–723.

Conant, R C and W Ross Ashby (1970). “Every good regulator of a system must be a model of that system”. In: *International Journal of Systems Science* 1.2, pp. 89–97.

Cos, I, J Duque, and P Cisek (2014). “Rapid prediction of biomechanical costs during action decisions”. In: *Journal of Neurophysiology* 112.6, pp. 1256–1266.

Cos, Ignasi, Nicolas Bélanger, and Paul Cisek (2011). “The influence of predicted arm biomechanics on decision making”. In: *Journal of Neurophysiology* 105.6, pp. 3022–3033.

Craig, A D (2002). “How do you feel? Interoception: the sense of the physiological condition of the body”. In: *Nature Reviews Neuroscience* 3.8, pp. 655–666.

Cummins, Robert (1989). *Meaning and Mental Representation.* Cambridge: MIT Press.

Dahl, A et al. (2013). “The Epigenesis of Wariness of Heights”. In: *Psychological Science* 24.7, pp. 1361–1367.

Dale, Rick (2008). “The possibility of a pluralist cognitive science”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 20.3, pp. 155–179.

Damasio, Antonio (1994). *Descartes' Error.* London: Vintage.

Davidson, Donald (1999). “The Emergence of Thought”. In: *Erkenntnis* 51.1, pp. 7–17.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature Neuroscience* 8.12, pp. 1704–1711.

Daw, Nathaniel D et al. (2011). "Model-Based Influences on Humans' Choices and Striatal Prediction Errors". In: *Neuron* 69.6, pp. 1204–1215.

Dayan, P (2012). "Twenty-five lessons from computational neuromodulation". In: *Neuron* 76.1, pp. 240–256.

Dayan, P et al. (1995). "The Helmholtz Machine". In: *Neural Computation* 7, pp. 889–904.

De Jaegher, Hanne and Ezequiel Di Paolo (2007). "Participatory sense-making". In: *Phenomenology and the Cognitive Sciences* 6.4, pp. 485–507.

Deneve, S (2008). "Bayesian spiking neurons I: inference". In: *Neural Computation* 20, pp. 91–117.

Dennett, Daniel (1969). *Content and Consciousness.* London: Routledge and Kegan Paul.

Dennett, Daniel C (1984). "Cognitive Wheels: The Frame problem in artificial intelligence". In: *Minds, Machines and Evolution.* Ed. by Christopher Hookway. Ablex, pp. 129–151.

Dietrich, Franz and Christian List (2016). "Mentalism versus behaviourism in economics: a philosophy-of-science perspective". In: *Economics and Philosophy* 32.2, pp. 249–281.

Doll, B B, D A Simon, and N D Daw (2012). "The ubiquity of model-based reinforcement learning". In: *Current Opinion in Neurobiology* 22.6, pp. 1–7.

Doya, Kenji et al., eds. (2007). *Bayesian brain: Probabilistic approaches to neural coding.* Cambridge: MIT Press.

Dretske, Fred (1981). *Knowledge and the Flow of Information.* Cambridge: MIT Press.

Eliasmith, Chris (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition.* Oxford: Oxford University Press.

Eliasmith, Chris and Charles H Anderson (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems.* Cambridge: MIT Press.

Ellsberg, Daniel (1961). "Risk, ambiguity, and the Savage axioms". In: *The quarterly journal of economics* 75.4, pp. 643–669.

Enel, Pierre et al. (2016). "Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex". In: *PLoS Computational Biology* 12.6, e1004967–35.

Engel, Andreas K, Karl J Friston, and Danica Kragic, eds. (2016). *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science.* Strungmann Forum Reports. Cambridge: MIT Press.

Engel, Andreas K et al. (2013). "Where's the action? The pragmatic turn in cognitive science". In: *Trends in Cognitive Sciences* 17.5, pp. 202–209.

Ernst, M O and M S Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870, pp. 429–433.

Fawcett, Tim W et al. (2014). "The evolution of decision rules in complex environments". In: *Trends in Cognitive Sciences* 18.3, pp. 153–161.

Feldman, H and K Friston (2010). "Attention, uncertainty, and free-energy". In: *Frontiers in Human Neuroscience* 4.215, pp. 1–23.

Feldman, Jacob (2013). "Tuning Your Priors to the World". In: *Topics in Cognitive Science* 5.1, pp. 13–34.

Field, Hartry H (1978). "Mental representation". In: *Erkenntnis* 13.1, pp. 9–61.

Fields, R. Douglas (2009). *The Other Brain.* New York: Simon and Schuster.

Fink, P W, P S Foo, and W H Warren (2009). "Catching fly balls in virtual reality: A critical test of the outfielder problem". In: *Journal of Vision* 9.13, pp. 14–14.

Fodor, J A (1980). "Methodological solipsism considered as a research strategy in cognitive psychology". In: *Behavioral and Brain Sciences* 3.1, pp. 63–109.

Fodor, J A and Z W Pylyshyn (1993). "Connectionism and cognitive architecture". In: *Cognition* 28.1, pp. 3–71.

Fodor, Jerry A (1975). *The Language of Thought.* Harvard University Press.

— (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge: MIT Press.

— (1992). *A Theory of Content and Other Essays.* Cambridge: MIT Press.

Freedman, David J and John A Assad (2011). "A proposed common neural mechanism for categorization and perceptual decisions". In: *Nature Neuroscience* 14.2, pp. 143–146.

Frijda, Nico H (1987). "Emotion, cognitive structure, and action tendency". In: *Cognition and emotion* 1.2, pp. 115–143.

— (2010). "Impulsive action and motivation". In: *Biological Psychology* 84.3, pp. 190–199.

Friston, K (2008). "Hierarchical Models in the Brain". In: *PLoS Computational Biology* 4.11, pp. 1–24.

— (2010). "The free-energy principle:a unified brain theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138.

— (2011a). "Embodied inference: Or I think therefore I am, if I am what I think". In: *The Implications of Embodiment (Cognition and Communication).* Ed. by Wolfgang Tschacher and Claudia Bergomi. Exeter: Imprint Academic, pp. 89–125.

— (2011b). "Functional and Effective Connectivity: A Review". In: *Brain Connectivity* 1.1, pp. 13–36.

— (2011c). "What Is Optimal about Motor Control?" In: *Neuron* 72.3, pp. 488–498.

— (2013). “Life as we know it”. In: *Journal of The Royal Society Interface* 10.20130475, pp. 1–12.

Friston, K, Jérémie Mattout, and James Kilner (2011). “Action understanding and active inference”. In: *Biological Cybernetics* 104.1-2, pp. 137–160.

Friston, K, Spyridon Samothrakis, and Read Montague (2012). “Active inference and agency: optimal control without cost functions”. In: *Biological Cybernetics* 106.8-9, pp. 523–541.

Friston, K, Christopher Thornton, and Andy Clark (2012). “Free-Energy Minimization and the Dark-Room Problem”. In: *Frontiers in Psychology* 3.130, pp. 1–7.

Friston, K et al. (2010). “Action and behavior: a free-energy formulation”. In: *Biological Cybernetics* 102.3, pp. 227–260.

Friston, K et al. (2012). “Dopamine, Affordance and Active Inference”. In: *PLoS Computational Biology* 8.1, pp. 1–20.

Friston, K et al. (2014). “The anatomy of choice: dopamine and decision-making”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, pp. 1–12.

Friston, K et al. (2015). “Active inference and epistemic value”. In: *Cognitive Neuroscience*, pp. 1–28. DOI: 10.1080/17588928.2015.1020053.

Friston, Karl J (1997). “Imaging cognitive anatomy”. In: *Trends in Cognitive Sciences* 1.1, pp. 21–27.

Fumagalli, R (2013). “The futile search for true utility”. In: *Economics and Philosophy* 29.3, pp. 325–347.

Gallese, V and Thomas Metzinger (2003). “Motor ontology: the representational reality of goals, actions and selves”. In: *Philosophical Psychology* 16.3, pp. 365–388.

Gardner, Howard (1985). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Geertz, Clifford (1973). *The interpretation of culture.* New York: Basic Books.

Gerard, Ralph W (1951). "Some of the problems concerning digital notions in the central nervous system". In: *Cybernetics: Circular causal and feedback mechanisms in biological and social systems. Transactions of the Seventh Conference.* Ed. by H V Foerster, M Mead, and H L Teuber. New York, Macy Foundation, pp. 11–57.

Gibson, James J. (1979). *The Ecological Approach to Visual Perception.* Boston, Houghton Mifflin.

Gigerenzer, G and P. M Todd (1999). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Gläscher, J et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Glimcher, Paul W (2011a). *Foundations of Neuroeconomic Analysis.* Oxford: Oxford University Press.

— (2011b). "Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis". In: *Proceedings of the National Academy of Sciences* 108.Supplement 3, pp. 15647–15654.

Glimcher, Paul W and Ernst Fehr (2014a). "Introduction: A Brief History of Neuroeconomics". In: *Neuroeconomics.* Ed. by Paul W Glimcher and Ernst Fehr. 2nd ed. Decision Making and the Brain. Academic Press, pp. xvii–xxvii.

— (2014b). *Neuroeconomics.* 2nd ed. Decision Making and the Brain. Academic Press.

Graziano, Michael S A (2016). "Ethological Action Maps: A Paradigm Shift for the Motor Cortex". In: *Trends in Cognitive Sciences* 20.2, pp. 121–132.

Graziano, Michael S A et al. (2002). "The Cortical Control of Movement Revisited". In: *Neuron* 36.3, pp. 349–362.

Gregory, R L (1980). "Perceptions as Hypotheses". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 290.1038, pp. 181–197.

Grush, Rick (2004). "The emulation theory of representation: Motor control, imagery, and perception". In: *Behavioral and Brain Sciences* 27.03, pp. 377–396.

Gu, Xiaosi et al. (2013). "Anterior insular cortex and emotional awareness". In: *Journal of Comparative Neurology* 521.15, pp. 3371–3388.

Hacking, Ian (1983). *Representing and Intervening.* Cambridge: Cambridge University Press.

Hammerstein, Peter and Jeffrey R Stevens (2012). *Evolution and the Mechanisms of Decision Making.* Cambridge: MIT Press.

Harnad, Stevan (1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1, pp. 335–346.

Hatfield, Gary (2002). "Perception as Unconscious Inference". In: *Perception and the Physical World.* Ed. by Dieter Heyer and Rainer Mausfeld. Oxford: Oxford University Press, pp. 115–143.

Haugeland, John (1985). *Artificial Intelligence: The Very Idea.* Cambridge: MIT Press.

— (2002). "Andy Clark on Cognition and Representation". In: *Philosophy of Mental Representation.* Ed. by Hugh Clapin. Oxford: Oxford University Press, pp. 24–36.

Hauk, Olaf, Ingrid Johnsrude, and Friedemann Pulvermüller (2004). "Somatotopic representation of action words in human motor and premotor cortex". In: *Neuron* 41.2, pp. 301–307.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Free Press.

Heyser, Charles J and Anthony Chemero (2012). "Novel object exploration in mice: Not all objects are created equal". In: *Behavioural Processes* 89.3, pp. 232–238.

Hinton, G E et al. (1995). "The "wake-sleep" algorithm for unsupervised neural networks". In: *Science* 268.5214, pp. 1158–1161.

Hobson, J A and K Friston (2012). "Waking and dreaming consciousness: Neurobiological and functional considerations". In: *Progress in Neurobiology* 98.1, pp. 82–98.

Hobson, J Allan and Karl Friston (2014). "Consciousness, Dreams, and Inference: The Cartesian Theatre Revisited". In: *Journal of Consciousness Studies* 21, pp. 6–32.

Hohmann, G W (1966). "Some effects of spinal cord lesions on experienced emotional feelings". In: *Psychophysiology* 3, pp. 143–156.

Hohwy, Jakob (2012). "Attention and conscious perception in the hypothesis testing brain". In: *Frontiers in Psychology* 3.96, pp. 1–14.

— (2013). *The Predictive Mind*. Oxford: Oxford University Press.

— (2014). "The Self-Evidencing Brain". In: *Nous*. DOI: 10.1111/nous.12062.

— (2015). "The Neural Organ Explains the Mind". In: *Open MIND* 19.T. Ed. by T. Metzinger and J. M. Windt, pp. 1–22. DOI: 10.15502/9783958570016.

— (2016). "The predictive processing hypothesis and 4e cognition". In: *The Oxford Handbook of Cognition: Embodied, Embedded, Enactive and Extended*. Ed. by A Newen, L Bruin, and S Gallagher. Oxford: Oxford University Press.

Hohwy, Jakob, A Roepstorff, and K Friston (2008). "Predictive coding explains binocular rivalry: an epistemological review". In: *Cognition* 108.3, pp. 687–701.

Hommel, Bernard et al. (2001). "The Theory of Event Coding (TEC): A framework for perception and action planning". In: *Behavioral and Brain Sciences* 24.5, pp. 849–937.

Hosoya, Toshihiko, Stephen A Baccus, and Markus Meister (2005). "Dynamic predictive coding by the retina". In: *Nature* 436.7047, pp. 71–77.

Hurley, Susan (1998). *Concsciousness in Action*. Harvard University Press.

Hutchins, E (2014). "The cultural ecosystem of human cognition". In: *Philosophical Psychology* 27.1, pp. 34–49.

Hutchins, Edward (1995). *Cognition in the Wild*. Cambridge: MIT Press.

Huth, Alexander G et al. (2016). "Natural speech reveals the semantic maps that tile human cerebral cortex". In: *Nature* 532.7600, pp. 453–458.

Hutto, Daniel D and Erik Myin (2013). *Radicalizing Enactivism*. Basic Minds Without Content. Cambridge: MIT Press.

James, William (1884). "What is an emotion?" In: *Mind* 9, pp. 188–205.

— (1890). *The Principles of Psychology*. New York: Dover.

Jeannerod, M (2006). *Motor Cognition*. Oxford: Oxford University Press.

Kanai, Ryota et al. (2015). "Cerebral hierarchies: predictive processing, precision and the pulvinar". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1668, pp. 20140169–20140169.

Kant, Immanuel (1781). *Critique of Pure Reason*. Cambridge: Hackett Publishing Company, Inc. [1996].

Kaplan, David Michael and William Bechtel (2011). "Dynamical models: an alternative or complement to mechanistic explanations?" In: *Topics in Cognitive Science* 3.2, pp. 438–444.

Kelso, J A S (2012). "Multistability and metastability: understanding dynamic coordination in the brain". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1591, pp. 906–918.

Kelso, J. A. Scott. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge: MIT Press.

Kiebel, Stefan J, Jean Daunizeau, and Karl J Friston (2008). “A hierarchy of time-scales and the brain”. In: *PLoS Computational Biology* 4.11, pp. 1–12. DOI: `10.1371/journal.pcbi.1000209`.

Kilner, James M et al. (2016). “Action-Oriented Models of Cognitive Processing”. In: *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science.* Ed. by Andreas K Engel, K Friston, and Danica Kragic, pp. 159–174.

Kirsh, David (2009). “Problem Solving and Situated Cognition”. In: *The Cambridge handbook of situated cognition.* Ed. by Philip Robbins and Murat Aydede. Cambridge: Cambridge University Press, pp. 264–305.

Klaes, C et al. (2011). “Choosing goals, not rules: deciding among rule-based action plans.” In: *Neuron* 70, pp. 536–548.

Knutson, Brian and Richard Peterson (2005). “Neurally reconstructing expected utility”. In: *Games and Economic Behavior* 52.2, pp. 305–315.

Kok, Peter, Janneke F M Jehee, and Floris P De Lange (2012). “Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex”. In: *Neuron* 75.2, pp. 265–270.

Kok, Peter et al. (2011). “Attention Reverses the Effect of Prediction in Silencing Sensory Signals”. In: *Cerebral Cortex* 22.9, pp. 1–10.

Körding, Konrad P and Daniel M Wolpert (2006). “Bayesian decision theory in sensorimotor control”. In: *Trends in Cognitive Sciences* 10.7, pp. 319–326.

Kuhn, Thomas (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Kveraga, Kestutis, Avniel S Ghuman, and Moshe Bar (2007). “Top-down predictions in the cognitive brain”. In: *Brain and Cognition* 65.2, pp. 145–168.

Lakoff, G and M Johnson (1980). “Conceptual metaphor in everyday language”. In: *The Journal of Philosophy.*

Landreth, Anthony and John Bickle (2008). “Neuroeconomics, Neurophysiology and the Common Currency Hypothesis”. In: *Economics and Philosophy* 24.03, pp. 419–429.

Lange, Carl G (1885). *Om sindsbevaegelser; et psyko-fysiologisk studie.* Copenhagen: Jacob Lunds.

Lazarus, Richard (1991). *Emotion and Adaptation.* Oxford: Oxford University Press.

Lee, Sang Wan, Shinsuke Shimojo, and John P O’Doherty (2014). “Neural Computations Underlying Arbitration between Model-Based and Model-free Learning”. In: *Neuron* 81.3, pp. 687–699.

Lende, Daniel H and Greg Downey, eds. (2012). *The Encultured Brain.* An Introduction to Neuroanthropology. Cambridge: MIT Press.

Lepora, Nathan F and Giovanni Pezzulo (2015). “Embodied Choice: How Action Influences Perceptual Decision Making”. In: *PLoS Computational Biology* 11.4, pp. 1–22. DOI: `10.1371/journal.pcbi.1004110`.

Lerner, Jennifer S et al. (2015). “Emotion and Decision Making”. In: *Annual Review of Psychology* 66.1, pp. 799–823.

Levy, D. J and P. W Glimcher (2012). “The root of all value: a neural common currency for choice”. In: *Current Opinion in Neurobiology* 22.6, pp. 1027–1038.

Lewis, Marc D and Rebecca M Todd (2005). “Getting Emotional”. In: *Journal of Consciousness Studies* 12.8-10, pp. 210–235.

Lindquist, Kristen A et al. (2012). “The brain basis of emotion: A meta-analytic review”. In: *Behavioral and Brain Sciences* 35.03, pp. 121–143.

Lipton, Peter (2004). *Inference to the Best Explanation.* London: Routledge.

Loewenstein, George and Jennifer S Lerner (2003). “The Role of Affect in Decision Making”. In: *Handbook of Affective Sciences.* Ed. by R J Davidson, K R Scherer, and H H Goldsmith. Oxford: Oxford University Press, pp. 619–642.

Machamer, Peter, Lindley Darden, and Carl F Craver (2000). “Thinking about Mechanisms”. In: *Philosophy of Science* 67.1, pp. 1–25.

Mandik, P (2005). “Action-oriented representation”. In: *Cognition and the brain: The philosophy and neuroscience movement.* Ed. by A Brook and K Akins. Cambridge University Press, pp. 284–305.

Maturana, H R and F J Varela (1980). *Autopoiesis and Cognition: The Realization of the Living.* Kluwer Academic Publishers Group.

McCulloch, Warren and Walter Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *Bulletin of Mathematical Biophysics* 5, pp. 115–133.

Menary, Richard (2007). *Cognitive Integration.* Palgrave Macmillan.

— (2010). “Introduction to the special issue on 4E cognition”. In: *Phenomenology and the Cognitive Sciences* 9.4, pp. 459–463.

— (2015). “Mathematical Cognition: A Case for Enculturation”. In: *Open MIND* 25.T. Ed. by T. Metzinger and J. M. Windt, pp. 1–20. DOI: 10.15502/9783958570818.

Miller, George A (2003). “The cognitive revolution: a historical perspective”. In: *Trends in Cognitive Sciences* 7.3, pp. 141–144.

Millikan, R G (1995). “Pushmi-pullyu representations”. In: *Philosophical Perspectives* 9, pp. 185–200.

Montague, P Read and Gregory S Berns (2002). “Neural economics and the biological substrates of valuation”. In: *Neuron* 36.2, pp. 265–284.

Moore, Kevin (2012). “Brains Don’t Predict, They Trial Actions”. In: *Frontiers in Psychology* 3, pp. 1–2. DOI: 10.3389/fpsyg.2012.00417.

Morgan, C. Lloyd (1894). *An Introduction to Comparative Psychology.* London W. Scott Ltd. URL: https://archive.org/details/anintroductiont00morggoog (visited on 06/17/2016).

Mumford, Tai Sing Lee David (2003). “Hierarchical Bayesian inference in the visual cortex”. In: *Journal of the Optical Society of America* 20.7, pp. 1434–1448.

Neisser, Ulric (1967). *Cognitive Psychology.* Classic Edition. New York: Psychology Press [2014].

Neumann, J von and O Morgenstern (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Newell, Allen (1980). “Physical Symbol Systems”. In: *Cognitive Science* 4.2, pp. 135–183.

Noe, Alva (2004). *Action in Perception.* Cambridge: MIT Press.

Okasha, S (2015). “On the interpretation of decision theory”. In: *Economics and Philosophy*, pp. 1–25. DOI: `10.1017/S0266267115000346`.

Open Science Collaboration (2015). “Estimating the reproducibility of psychological science”. In: *Science* 349.6251, pp. 1–8.

Oppenheimer, Daniel M and Evan Kelso (2015). “Information Processing as a Paradigm for Decision Making”. In: *Annual Review of Psychology* 66.1, pp. 277–294.

O’Regan, J Kevin and Alva Noe (2001). “A sensorimotor account of vision and visual consciousness”. In: *Behavioral and Brain Sciences* 24.05, pp. 939–973.

Orlandi, Nico (2014). *The innocent eye: why vision is not a cognitive process.* Oxford: Oxford University Press.

Ouden, H E M den et al. (2010). “Striatal Prediction Error Modulates Cortical Coupling”. In: *Journal of Neuroscience* 30.9, pp. 3210–3219.

Ouden, Hanneke E M den, Peter Kok, and Floris P De Lange (2012). “How prediction errors shape perception, attention, and motivation”. In: *Frontiers in Psychology* 3.548, pp. 1–12.

Padoa-Schioppa, C (2011). “Neurobiology of economic choice: a good-based model”. In: *Annual Review of Neuroscience* 34, pp. 333–359.

Park, D C and C M Huang (2010). “Culture wires the brain: A cognitive neuroscience perspective”. In: *Perspectives on Psychological Science* 5.4, pp. 391–400.

Park, H J and K Friston (2013). “Structural and Functional Brain Networks: From Connections to Cognition”. In: *Science* 342.6158, pp. 1238411–1238411.

Park, J. W and J Zak (2007). “Neuroeconomic studies”. In: *Analyse and Kritik* 29, pp. 47–59.

Pastor-Bernier, A and P Cisek (2011). “Neural correlates of biased competition in premotor cortex”. In: *The Journal of Neuroscience* 31.19, pp. 7083–7088.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Fransisco, Morgan Kaufmann Publishers.

Pessoa, Luiz (2009). “How do emotion and motivation direct executive control?” In: *Trends in Cognitive Sciences* 13.4, pp. 160–166.

— (2010). “Emotion and cognition and the amygdala: From “what is it?” to “what’s to be done?”” In: *Neuropsychologia* 48.12, pp. 3416–3429.

— (2013). *The Cognitive-Emotional Brain: From Interactions to Integration.* Cambridge: MIT Press.

— (2014). “Understanding brain networks and brain organization”. In: *Physics of Life Reviews* 11.3, pp. 400–435.

Peterson, Martin (2009). *An Introduction to Decision Theory.* Cambridge: Cambridge University Press.

Pezzulo, G, F Rigoli, and F Chersi (2013). “The mixed instrumental controller: using value of information to combine habitual choice and mental simulation”. In: *Frontiers in Psychology* 4.92, pp. 1–15.

Pezzulo, G et al. (2011). “The mechanics of embodiment: A dialog on embodiment and computational modeling”. In: *Frontiers in Psychology* 2.5, pp. 1–21. DOI: 10.3389/fpsyg.2011.00005.

Pezzulo, G et al. (2014). “The principles of goal-directed decision-making: from neural mechanisms to computation and robotics”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, pp. 1–6.

Pezzulo, Giovanni (2008). “Coordinating with the Future: The Anticipatory Nature of Representation”. In: *Minds & Machines* 18.2, pp. 179–225.

— (2011). “Grounding procedural and declarative knowledge in sensorimotor anticipation”. In: *Mind and Language* 26.1, pp. 78–114.

— (2012). “An Active Inference view of cognitive control”. In: *Frontiers in Psychology* 3.478, pp. 1–2.

Pezzulo, Giovanni, Francesco Rigoli, and K Friston (2015). “Active Inference, homeostatic regulation and adaptive behavioural control”. In: *Progress in Neurobiology* 134, pp. 17–35.

Pezzulo, Giovanni et al. (2016). “Active Inference, epistemic value, and vicarious trial and error”. In: *Learning and Memory* 23.7, pp. 322–338.

Pfeifer, Rolf and Josh Bongard (2007). *How the Body Shapes the Way We Think: A New View of Intelligence.* Cambridge: MIT Press.

Phelps, Elizabeth A, Karolina M Lempert, and Peter Sokol-Hessner (2014). “Emotion and Decision Making: Multiple Modulatory Neural Circuits”. In: *Annual Review of Neuroscience* 37.1, pp. 263–287.

Phillips, W A, Andy Clark, and S M Silverstein (2015). “On the functions, mechanisms, and malfunctions of intracortical contextual modulation”. In: *Neuroscience and Biobehavioral Reviews* 52, pp. 1–20.

Piccinini, Gualtiero and Carl Craver (2011). “Integrating psychology and neuroscience: functional analyses as mechanism sketches”. In: *Synthese* 183.3, pp. 283–311.

Piccinini, Gualtiero and Andrea Scarantino (2010). “Information processing, computation, and cognition”. In: *Journal of Biological Physics* 37.1, pp. 1–38.

Platt, M and C Padoa-Schioppa (2009). “Neuronal representations of value”. In: *Neuroeconomics*. Ed. by Paul W Glimcher et al. 1st ed. Academic Press, pp. 439–460.

Poldrack, R A (2010). “Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed?” In: *Perspectives on Psychological Science* 5.6, pp. 753–761.

Poldrack, Russell (2006). “Can cognitive processes be inferred from neuroimaging data?” In: *Trends in Cognitive Sciences* 10.2, pp. 59–63.

Poldrack, Russell A, Yaroslav O Halchenko, and Stephen José Hanson (2009). “Decoding the Large-Scale Structure of Brain Function by Classifying Mental States Across Individuals”. In: *Psychological Science* 20.11, pp. 1364–1372.

Pouget, Alexandre et al. (2013). “Probabilistic brains: knowns and unknowns”. In: *Nature Reviews Neuroscience* 16.9, pp. 1170–1178.

Powers, William Treval (1973). *Behavior: The Control of Perception.* London: Wildwood House.

Price, C J and K Friston (2002). “Degeneracy and cognitive anatomy”. In: *Trends in Cognitive Sciences.*

Price, Cathy J and K Friston (2005). “Functional ontologies for cognition: The systematic definition of structure and function”. In: *Cognitive Neuropsychology* 22.3-4, pp. 262–275.

Prinz, Jesse J (2004). *Gut Reactions: A Perceptual Theory of Emotions.* Oxford: Oxford University Press.

Prinz, Jesse J and Lawrence Barsalou (2000). “Steering a Course for Embodied Representation”. In: *Cognitive Dynamics: Conceptual Change in Humans and Machines.* Ed. by Eric Dietrich and Arthur B. Markman. Lawrence Erlbaum Associates, Inc., pp. 51–78.

Pulvermüller, Friedemann (1999). “Words in the brain’s language”. In: *Behavioral and Brain Sciences* 22.02, pp. 253–279.

Putnam, Hilary (1991). *Representation and Reality.* Cambridge: MIT Press.

Pylyshyn, Zenon Walter (1984). *Computation and cognition.* Cambridge: Cambridge University Press.

Qiu, C, R S Shivacharan, and M Zhang (2015). “Can Neural Activity Propagate by Endogenous Electrical Field?” In: *The Journal of Neuroscience* 35.48, pp. 15800–15811.

Ramsey, William (2015). “Must cognition be representational?” In: *Synthese*, pp. 1–18. DOI: 10.1007/s11229-014-0644-6.

Ransom, Madeleine, Sina Fazelpour, and Christopher Mole (2016). “Attention in the predictive mind”. In: *Consciousness and Cognition*, pp. 1–14. DOI: 10.1016/j.concog.2016.06.011.

Rao, Rajesh P N and Dana H Ballard (1999). “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature Neuroscience* 2.1, pp. 79–87.

Ratcliff, Roger (1978). “A theory of memory retrieval.” In: *Psychological Review* 85.2, p. 59.

Redgrave, P and K Gurney (2006). “The short-latency dopamine signal: a role in discovering novel actions?” In: *Nature Reviews Neuroscience* 7, pp. 967–975.

Rigotti, Mattia et al. (2013). “The importance of mixed selectivity in complex cognitive tasks”. In: *Nature* 497.7451, pp. 585–590.

Robbins, Philip and Murat Aydede (2009). “A short primer on situated cognition”. In: *The Cambridge handbook of situated cognition.* Ed. by Philip Robbins and Murat Aydede. Cambridge: Cambridge University Press, pp. 3–10.

Ross, Don and James Ladyman (2010). “The Alleged Coupling-Constitution Fallacy and the Mature Sciences”. In: *The Extended Mind.* Ed. by Richard Menary. Cambridge: MIT Press, pp. 155–166.

Russell, James A and Lisa Feldman Barrett (1999). “Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.” In: *Journal of Personality and Social Psychology* 76.5, pp. 805–819.

— (2009). “Core Affect”. In: *Oxford Companion to Emotion and the Affective Sciences.* Ed. by David Sander and Klaus Scherer. Oxford: Oxford University Press.

Samuelson, Paul (1938). “A note on the pure theory of consumer’s behaviour”. In: *Economica.* 51.17, pp. 61–71.

Savage, Leonard (1954). *The Foundations of Statistics.* Second. Dover: John Wiley and Sons [1972].

Searle, John R (1980). “Minds, brains, and programs”. In: *Behavioral and Brain Sciences* 3.03, pp. 417–424.

Sen, A (1971). “Choice functions and revealed preference”. In: *Review of Economic Studies* 38, pp. 307–317.

Seth, Anil K (2013). “Interoceptive inference, emotion, and the embodied self”. In: *Trends in Cognitive Sciences* 17.11, pp. 565–573.

— (2014). “A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia”. In: *Cognitive Neuroscience* 5.2, pp. 97–118.

— (2015). “The Cybernetic Bayesian Brain”. In: *Open MIND.* Ed. by T. Metzinger and J. M. Windt. Frankfurt am Main: MIND Group, pp. 1–52.

Shapiro, Larry (2007). “The Embodied Cognition Research Programme”. In: *Philosophy Compass* 2.2, pp. 338–346.

Shapiro, Lawrence (2011). *Embodied Cognition.* London: Routledge.

Sherrington, Charles (1947). *The Integrative Action of the Nervous System.* Cambridge University Press.

Shipp, Stewart, Rick A Adams, and K Friston (2013). "Reflections on agranular architecture: predictive coding in the motor cortex". In: *TRENDS in Neurosciences* 36.12, pp. 706–716.

Simon, Herbert (1983). *Reason in Human Affairs.* Stanford University Press.

Simon, Herbert A (1990). "Invariants of human behavior". In: *Annual Review of Psychology* 41.1, pp. 1–20.

Simon, Herbert A and Allen Newell (1971). "Human problem solving: The state of the theory in 1970." In: *American Psychologist* 26.2, pp. 145–159.

Smith, Tom et al. (2002). "Neuronal plasticity and temporal adaptivity: GasNet robot control networks". In: *Adaptive Behavior* 10.3-4, pp. 161–183.

Soliman, Tamer and Arthur M Glenberg (2014). "The Embodiment of Culture". In: *The Routledge Handbook of Embodied Cognition.* Ed. by Lawrence Shapiro. London: Routledge, pp. 207–219.

Sporns, Olaf (2011). *Networks of the Brain.* Cambridge: MIT Press.

Spratling, Michael W (2013). "Distinguishing theory from implementation in predictive coding accounts of brain function". In: *Behavioral and Brain Sciences* 36.3, pp. 231–232.

Stapleton, Mog (2016). "Leaky Levels and the Case for Proper Embodiment". In: *Embodiment in Evolution and Culture.* Ed. by G Etzelmuller and C Tewes. Tuebingen: Mohr Siebeck., pp. 17–30.

Steels, Luc (2003). "Intelligence with representation". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 361.1811, pp. 2381–2395.

Stephens, David W (2008). "Decision ecology: Foraging and the ecology of animal decision making". In: *Cognitive Affective and Behavioural Neuroscience* 8.4, pp. 475–484.

Stepp, N, A Chemero, and M T Turvey (2011). "Philosophy for the rest of cognitive science". In: *Topics in Cognitive Science* 3.2, pp. 425–437.

Sterelny, K (2010). "Minds: extended or scaffolded?" In: *Phenomenology and the Cognitive Sciences* 9, pp. 465–481.

Stern, Robert (2004). "Does 'ought' imply 'can'? And did Kant think it does?" In: *Utilitas* 16.01, pp. 42–61.

Stich, Stephen P (1978). "Beliefs and Subdoxastic States". In: *Philosophy of Science* 45.4, pp. 499–518.

Sugrue, Leo P, Greg S Corrado, and William T Newsome (2005). "Choosing the greater of two goods: neural currencies for valuation and decision making". In: *Nature Reviews Neuroscience* 6.5, pp. 363–375.

Summerfield, Chris and Konstantinos Tsetsos (2015). "Do humans make good decisions?" In: *Trends in Cognitive Sciences* 19.1, pp. 27–34.

Tarsitano, Michael S and Richard Andrew (1999). "Scanning and route selection in the jumping spider Portia labiata". In: *Animal Behaviour* 58.2, pp. 255–265.

Tervaniemi, Mari, Sini Maury, and Risto Näätänen (1994). "Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity." In: *Neuroreport* 5.7, pp. 844–846.

Thelen, Esther and Linda Smith (1994). *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge: MIT Press.

Thompson, Evan (2004). "Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela". In: *Phenomenology and the Cognitive Sciences* 3.4, pp. 381–398.

— (2007). *Mind in Life: Biology, Phenomenology and the Sciences of Mind*. Harvard University Press.

Thura, D and P Cisek (2014). "Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making". In: *Neuron* 81.6, pp. 1401–1416.

Treue, S (2001). "Neural correlates of attention in primate visual cortex". In: *TRENDS in Neurosciences* 24.5, pp. 295–300.

— (2003). "Visual attention: the where, what, how and why of saliency". In: *Current Opinion in Neurobiology* 13.4, pp. 428–432.

Turing, A M (1936/37). "On computable numbers, with an application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* 42.1, pp. 230–265.

Tversky, Amos and Daniel Kahneman (1981). "The framing of decisions and the psychology of choice". In: *Science* 211, pp. 453–458.

— (1986). "Rational choice and the framing of decisions". In: *Journal of Business* 59.4, pp. 251–278.

Varela, Francisco J, Evan Thompson, and Eleanor Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press.

Vlaev, Ivo et al. (2011). "Does the brain calculate value?" In: *Trends in Cognitive Sciences* 15.11, pp. 546–554.

Wells, Andy (2005). *Rethinking cognitive computation: Turing and the science of the mind*. London: Palgrave Macmillan.

Wheeler, M and Andy Clark (1999). "Genic representation: reconciling content and causal complexity". In: *The British Journal for the Philosophy of Science* 50.1, pp. 103–135.

Wilson, A D and S Golonka (2013). "Embodied cognition is not what you think it is". In: *Frontiers in Psychology* 4.58. DOI: 10.3389/fpsyg.2013.00058.

Wilson, Margaret (2002). “Six views of embodied cognition”. In: *Psychonomic Bulletin and Review* 9.4, pp. 625–636.

Withagen, Rob et al. (2012). “Affordances can invite behavior: Reconsidering the relationship between affordances and agency”. In: *New Ideas in Psychology* 30.2, pp. 250–258.

Wolpert, Daniel M and Michael S Landy (2012). “Motor control is decision-making”. In: *Current Opinion in Neurobiology* 22.6, pp. 1–8.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanations.* Oxford: Oxford University Press.