From Life-Like to Mind-Like Explanation: Natural Agency and the Cognitive Sciences

by

Alex Djedovic

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Institute for the History and Philosophy of Science and Technology
University of Toronto

From Life-Like to Mind-Like Explanation: Natural Agency and the Cognitive Sciences

Alex Djedovic

Doctor of Philosophy

Institute for the History and Philosophy of Science and Technology
University of Toronto

2020

Abstract

This dissertation argues that cognition is a kind of natural agency. Natural agency is the capacity that certain systems have to act in accordance with their own norms. Natural agents are systems that bias their repertoires in response to affordances in the pursuit of their goals. Cognition is a special mode of this general phenomenon. Cognitive systems are agents that have the additional capacity to actively take their worlds to be certain ways, regardless of whether the world is really that way. In this way, cognitive systems are *desituated*. Desituatedness is the root of specifically cognitive capacities for representation and abstraction. There are two main reasons why this view needs defending. First, natural agency is typically viewed as incompatible with natural science because it is committed to a teleological mode of explanation. Second, cognition is typically held to be categorically distinct from natural agency. This dissertation argues against both of these views. It argues against the incompatibility of agency and natural science by demonstrating that systems biology, general systems theory, and sciences that deal with complex systems have typically underappreciated conceptual and theoretical resources for grounding agency in the causal structure of the world. These conceptual resources do not, however, reduce agency to systems theory because the normativity inherent in agency demands descriptive resources beyond those of even the most sophisticated systems theory. It argues against the categorical

difference between natural agency and cognition by pointing out that separating cognition from a richer web of situated, ecologically embedded relations between the agent and the world generates the frame problem, which is an insuperable obstacle to making cognition that is sufficiently responsive to the complexity of the world. Rooting cognition in natural agency is a more robust empirical bet for theorizing cognition and artificial intelligence.

Acknowledgments

The process of writing this dissertation illustrates how our environments are intimately tied to our agency. It would not have been possible to write it without the help of a complex, dynamic, adaptive, and wonderful ecosystem of fellow agents. I have been extraordinarily fortunate to interact with them all.

My first and foremost gratitude goes to my dissertation committee. From my supervisor Denis Walsh I learned patience, precision, and philosophical craftmanship. My deep thanks for his unwavering, dedicated, and constructive support of this project. From John Vervaeke I learned boldness of thought and to appreciate the richness and nobility of philosophy. From Bill Seager I learned to temper boldness with precision. From Brian Cantwell Smith I learned deep intellectual humility. Their contributions have made this work immeasurably better.

I also owe a deep debt of gratitude to the cognitive science community at the University of Toronto. The enthusiasm and acumen of fellow students, former students, teaching colleagues, and philosophical fellow-travelers kept me going through the darkest impasses of dissertation writing. I wish to thank Jesse Berlin, Juensung Kim, Michael Galang, Aaron Henry, Simon Cook, Filip Miscevic, Hannah Cho, Emma Pask, Brendan Smith, Thomas Stanuolis, Cassandra Williams, Elizabeth Long, and Philip Rajewicz for stimulating and ongoing conversations about many of the issues discussed in this dissertation. My fellow teaching assistants Susan Gillingham, Justina Zatzman, Anderson Todd, Kimberley Cho, Wook Yang helped solidify much of my understanding through teaching. I wish to particularly recognize Amogh Sahu and Jana Kurrek for reading long chunks of earlier versions of this dissertation with enthusiasm and critical acumen.

I was also part of a vibrant community of graduate students at the IHPST. Thanks to Cory Lewis, Fermin Fulda, Michael Cournoyea, Rebecca Moore, Xavi Lanao, Melissa Charenko, Greg Lusk, Gwyndaf Garbutt, Agnes Bolinska, Esther Atkinson, Kira Lussier, Paul Greenham, Sarah Qidwai, Nicholas Overgaard, Chris Belanger, Isaac Record, Ellie Louson, Curtis Forbes, and Paul Patton for conversations and support that were pivotal in my intellectual development.

I also wish to thank colleagues that informed my intellectual standards though their demonstrations of true philosophical, intellectual and human excellence: Eugene Earnshaw, Jacob Stegenga, Sean Smith, Zach Irving, Dave Suarez, and Parisa Moosavi.

My thanks to Evan Thompson, Yiftach Fehige, Hakob Barseghyan, Brian Baigrie, Marga Vicedo, Andrew Buskell, Rasmus Winther, Stuart Newman, Jossi Berkovitz, Sonia Sultan, and Lindley Darden for important conversations on many philosophical topics, big and small.

I also wish to collectively recognize the membership of CUPE Local 3902 in fighting for the economic stability for academic workers without which this work would not have been possible.

Deep thanks to my family for their long, unwavering support of this difficult-to-articulate project. I dedicate the dissertation to my grandfather, who sadly did not live to see its completion. In the end, my deepest gratitude goes to my partner Megan Harris, who understands very well the toll this work took, and who bore with me anyway.

Table of Contents

List of Figures

Introduction

This dissertation is an attempt to articulate the place of mind in nature. It is a contribution to the project of naturalizing mind.

Naturalization is the assimilation of mind to the picture of the world given by natural science. This has traditionally meant assimilating mind to the ontology given by physics, or finding a place for mind in the mechanism of nature. This dissertation proposes a radically different route of naturalization. Instead of assimilating mind to physics, it seeks to assimilate mind to life.[1]

The fundamental claim in this dissertation is that *living things are agents and minded things are just agents of a sophisticated, cognitive sort.* That is, living things have the requisite properties for grounding cognition. If this is right, there is no gulf between life and mind, only a spectrum of complexity and sophistication.[2] The fact that cognition seems to us categorically different from other forms of natural agency stems from how salient cognition is to cognitive agents.

Invoking agency clarifies two related aspects of assimilating mind to life. Agency offers a battery of concepts for understanding life on the one hand and cognition on the other. Living things are agents of a sort, and cognition is a sophisticated form of agency.

There are two questions on how agency relates to life and cognition. (1) What are living things such that they make appropriate the agential concepts we bring to bear on them? (2) How are basic agential concepts related to cognition?

---

[1] In this, the dissertation operates within the aims of a broadly enactivist research program articulated in Thompson (2007), Barandiaran et al (2009), Moreno & Mossio (2015), and Di Paolo et al (2017).

[2] This view has three main sources. (1) the thesis of the deep continuity of life and mind (Varela et al 1991; Di Paolo 2005; Thompson 2007; Hutto & Myin 2013, 2017; Moreno & Mossio 2015; Di Paolo et al 2017). (2) Embodied, embedded, extended, and enactive approaches to cognition (Brooks 1991; Damasio 2000; Anderson 2003; Wheeler 2005, Chemero 2009; Clark 1996, 2013). (3) The view that organisms constitute a basis for natural agency (Skewes & Hooker 2009; Christensen 2012; Walsh 2006, 2008, 2012, 2015, 2018; Fulda 2016; Jones 2016; Nicholson 2013, 2018).

For (1), living things are complex systems that realize adaptive autopoiesis.[3] Autopoiesis is a set of capacities that allows a system to self-organize as a coherent whole within the constraints of physical necessity. Adaptivity is the capacity to act on the environment in pursuit of autopoiesis. When adaptive autopoietic systems are situated in their environments, the agential conceptual web of goals, repertoire, and affordances is necessary for understanding the behaviour of that system. Agents are systems that adaptively respond to what their conditions afford.

For (2), agency grounds cognition but is not identical with it. Cognition is a complex form of agency. The response of simple agents to their affordances is intimate and direct, whereas complex agents have indeterminate or open-ended goals. To the extent they are complex, agents operate under increasing uncertainty about what their conditions afford. In order to act, complex agents must on occasion *take* the world to be a certain way quite independent of the way the world is. Complex agents thus respond primarily to how they take the world to be, not to what the conditions afford. Cognition is grounded on the capacity to take the world to be certain ways. It underwrites a *desituated* relation to the world, in contrast to the situatedness of simple agents.

The four chapters of this dissertation each make a distinct contribution to substantiating the assimilation of cognition to life. Very roughly, Chapter 1 articulates already developed heterodox views of agency, mind, natural science, and life that support the fundamental claim. Chapter 2 articulates how agency conceptually maps onto life. Chapter 3 articulates a version of natural agency that is palatable to the mechanistic tenor of natural science in general and to cognitive science in particular. Chapter 4 articulates the mapping of agency onto cognition, specifically the difference between the life-like explanatory framework appropriate for simple agents and the mind-like explanatory framework appropriate for complex agents.

This introduction gives an overview of agency, life, cognition, and the details of the assimilation relation between cognition and life. It then offers a brief overview of the more specific arguments found in the four main chapters of the dissertation.

---

[3] Autopoiesis is described in Maturana & Varela (1980), Varela et al (1991), Di Paolo (2005), Thompson (2007), and Moreno & Mossio (2015).

Agency is a readily observable gross behavioural property of living things. It is the phenomenon of a suitably integrated whole (the agent) exerting appropriate control over itself and (to a lesser extent) the world. For example, bacteria break down nutrients, squirrels bury acorns, and persons behave appropriately or inappropriately.

More specifically, agents are systems that have goals, repertoire, and affordances. Goals are states of affairs that are plastically and reliably attained by agents. Repertoire is the agent's range of potential actions in a given situation. Affordances are what the environment offers to an agent given its goals and repertoire. Taken together, agents are systems that bias their repertoires in response to affordances in the pursuit of their goals. In giving goals explanatory power, agential thinking is inescapably teleological.[4]

What is it to naturalize mind? Naturalism comes in many forms.[5] Typically, we distinguish ontological and methodological naturalism.[6]

Roughly, ontological naturalism is the view that supernatural entities do not exist. Methodological naturalism is the view that understanding the natural world requires no reference to anything supernatural.

However, it is not clear that this distinction is useful. In this dissertation I subscribe to what I call liberal naturalism. This is the view that the natural sciences place constraints on what we can say there is, and offer guidance on how to study it. Nothing can be admitted to our ontology that is inconsistent with our best science. Nor is it appropriate to study the world in ways that are inconsistent with our best sciences. But these constraints leave significant scope for expanding our ontology and methodology. Liberal naturalism also holds that the explanatory indispensability of concepts in a field of study gives *prima facie* evidence that the world is the

---

[4] The analysis of agency and its cognate concepts (goals, repertoire, affordances, normativity) can be found in Walsh (2006, 2008, 2012, 2015, 2018), and Fulda (2016, 2017). Other work on the concept of agency within this tradition can be found in Jonas (1966), Ayala (1970), Bedau (1991, 1992), Juarrero (1999), Thompson (2007), Burge (2009, 2010), Christensen (2012) Di Paolo (2005), Barandiaran et al (2009), Hanna & Maiese (2009), Nicholson (2013), Di Paolo et al (2017).

[5] For an overview, see Lewens (2012).

[6] See Van Riel & Van Gulick (2019) for an overview. The attitude of liberal naturalism is well-expressed in Shaffner (2006), but see Winther (2009) for complications.

way the concepts describe it to be. If our best science cannot do without it a particular phenomenon, then our best science tells us the phenomenon is real.

This liberal naturalism underwrites the appeal to agency. Agential concepts are indispensable in biology and cognitive science. Despite this indispensability, less liberal versions of naturalism hold to instrumentalism about agency.

Instrumentalism is the view that we treat certain physical systems as agents due to the intractable complexity of their actual workings.[7] This instrumentalism only makes sense on the supposition of a hard distinction between ontological and methodological naturalism. An instrumentalist may hold agency to be methodologically indispensable but ontologically dispensable. Liberal naturalism does not admit such a distinction.

On a liberal naturalist view, although the existence of agents is in tension with the physicalist position of treating nature as a vast mechanism, the tension does not entail that agency should be reduced or eliminated. Jonas (1966) articulates this tension:

> How this finalism tallies, in the same world, with mechanical causality whose reality cannot be denied either is a problem not to be "solved" by sacrificing an evidence (purposiveness) to a theorem (exclusiveness of *causa efficiens*) which was derived by generalization from another evidence; but, if solvable at all, only by treating it as the profoundly challenging and as yet completely unsettled problem it is. (Jonas 1966, 90-91)

Liberal naturalism seeks to in part resolve the tension between the finalism of agency and the pervasiveness of mechanism by taking naturalism to be a commodious view. Agential concepts earn the right to be part of the scientific picture of the world by their usefulness to biology and cognitive science, even if mechanism is truly all-pervasive.

There is an even broader question here: *what motivates assimilating mind to agency*? Briefly, the approach outlined here is an alternative route for understanding the mind's place in nature. This is because despite operating at an incredible degree of dialectical refinement and sophistication,

---

[7] This view is practically universal in biology in the form of treating agency as *teleonomy* (Okasha 2018) and commands considerable support in cognitive science as the intentional stance (e.g. Dennett 1989).

the attempt to assimilate mind to nature, understood as a collection of physical or mechanical processes, have arguably reached a stalemate.[8] Assimilating mind to agency is an approach with considerably more open questions and opportunities.

This claim is, to some extent, a matter of philosophical temperament and so must form a basic foundation of the project. That said, three broad considerations motivate the alternative approach in this dissertation.

First, the view that assimilating mind to nature means assimilating mind to physics presupposes that there is a grand reduction of the sciences to physics. However, the alternative view that there are multiple levels of explanation even in physics, and more so for other sciences, has been gaining currency for quite some time in philosophy of science. What the varieties of anti-reductionism mean is contested. But for present purposes, all that is needed to motivate the project is commitment to anti-reductionism in some form.[9] In this way, the explanatory power of life and agency is to some extent autonomous from its obvious implementation in physical mechanism.

Second, the received view of the mind as a representational process may not be the easiest avenue for naturalizing mind. Representationalism is the view that cognition is the manipulation of representations, which are states that stand in for aspects of the world. On this view, minds are systems that manipulate representational states that stand in intentional relations to the world, where states of the cognitive system are *about* states of the world. The criticisms of representationalism are variegated, but they all share the supposition that the capacity to represent is not *the* fundamental feature of mind.[10] It is, if anything, a symptom of a more primitive capacity which I call *taking* the world to be a certain way. The assimilation of mind to agency offers a novel route for approaching this more primitive taking capacity. Representationalism may or may not be vindicated by this taking capacity; that question is beyond the scope of this dissertation.

---

[8] For an overview of the philosophical issues in physicalism, see Stoljar (2017).

[9] See Van Riel & Van Gulick (2019) for an overview of the varieties of anti-reductionism.

[10] For an overview of the criticism, see Ramsey (2017), Hutto & Satne (2015), and Steiner (2014).

Third, the computational approach to mind is also not the only game in town. One important feature that computational approaches miss is that cognition happens in organisms, which, as we will see, have properties more involved that the input/output or read/write dynamics of computational systems. Computation is ultimately compatible with the organicist view, and this dissertation takes a few steps in arguing this, but developing this claim is also beyond the scope of the present project.

Chapter 1 surveys four heterodox views that are prerequisites for making the fundamental claim: (i) goal-centered agency, (ii) situated cognition, (iii) complexity science, and (iv) situated Darwinism. The take-aways from Chapter 1 are (i) that agency must be understood as the capacity to pursue goals, not the capacity to respond to reasons, (ii) that explaining cognition requires a ser of situated relations between the cognitive system and the world in addition to the more familiar representational ones, (iii) that natural science makes indispensable use of emergent properties, rather than requiring reduction to mechanisms, (iv) that the evolutionary process itself requires appeal to the agency of organisms, rather than to the workings of their parts. Taken together, these four views, which have been developed deeply elsewhere, motivate the more specific question of how mind is to be reduced to agency.

Chapter 2 critically examines two accounts of agency: the Organizational Account of Agency (OAA) and the Ecological Account of Agency (EAA). The OAA holds that agency inheres in the physical properties of certain kinds of complex systems. The EAA holds that agency is a gross behavioral profile that demands a certain kind of explanation. The chapter argues that the OAA can only ground the basic goal of the persistence of an agent, and so is not enough to ground the rich forms of agency that are required to ground mind. However, the EAA is an indispensable and complementary framework for the OAA if it is to explain mind.

Chapter 3 is an attempt to show how liberal naturalism can reconcile agency with our best natural sciences. Understanding agency requires commitment to teleological explanation. This chapter argues, following recent philosophical developments, that there is an understanding of teleology that is perfectly natural. By way of further motivating the commitment to teleology, the chapter argues that this understanding of teleology lessens the theoretical impasse of the frame problem in cognitive science. While the frame problem is not solved by appeal to teleology, it is

transformed into a series of framing issues, which constitutes a modest measure of progress on understanding the mind's supple context-sensitivity.

Chapter 4 sketches the difference between agency in general and cognitive agency in particular. The difference between agents in general and cognitive agents is *desituatedness*. All agents are ecologically situated systems, but cognitive agents are those agents that achieve a degree of desituatedness in how they exercise their capacities. Desituatedness is not yet a sharp categorical cutoff, but the difference between situated and desituated capacities can be seen at the far ends of the spectrum of complexity that exists between simple and sophisticated natural agents.

In sum, this dissertation articulates a heterodox approach to agency, cognition, and scientific explanation. On this view, agency, life, mind, and nature hang together in intimate relations of mutual explanatory illumination. The potential payoff for developing the sciences of the mind is immense.

The fundamental claim is that cognition is a kind of agency in which the agent achieves a desituated relation between itself and its world. Agency, on this understanding, is a phenomenon that occurs only at certain levels of complexity in the natural world. The task of the dissertation is to articulate this view that life is agency, that cognition is agency, and that cognition is grounded in life.

1 Four Heterodoxies for Natural Agency

This dissertation is a contribution to the project of integrating cognition into the natural agency of organisms. The basic idea is that organisms are natural agents, and cognition is a sophisticated type of agency. Cognition is one way in which organisms pursue their natural purposes. This idea already has some currency in philosophy, biology, and the cognitive sciences.[11] The aim of what follows is to synthesize and refine this integrative approach.

This integration of cognition into natural agency faces a *prima facie* explanatory gap. This gap is the intuition that natural agency and cognition are quite different phenomena. More specifically, it is the intuition that the explanatory tools of natural science cannot do justice to the essential aspects of cognition. Behind this intuition is the further intuition that cognitive agents have intentional relations to their world, whereas putative natural agents are the result of causes and conditions.[12] These intentional relations seem to be unconstrained by the more familiar causal relations. So, it seems that the explanatory frameworks of natural science cannot be brought to bear on the cognitive processes of thinking things.

The basic claim of this chapter is that this explanatory gap is usefully recast using the phenomena of *complex systems* and *agency*.

Complex systems enrich the explanatory resources for thinking about cognitive agents as physical entities. These systems are not like the simple, linear causal system that our mechanistic sciences usually deal with. Their existence suggests that there is more to the explanation of certain physical systems that giving mechanisms. The sorts of explanations we give of complex systems stretches the limits of physical explanation.

---

[11] The view has three main sources. (1) the thesis of the deep continuity of life and mind (Varela et al 1991; Di Paolo 2005; Thompson 2007; Hutto & Myin 2013, 2017; Moreno & Mossio 2015; Di Paolo et al 2017). (2) Embodied, embedded, extended, and enactive approaches to cognition (Brooks 1991; Damasio 2000; Anderson 2003; Wheeler 2005, Chemero 2009; Clark 1996, 2013). (3) The view that organisms constitute a basis for natural agency (Skewes & Hooker 2009; Christensen 2012; Walsh 2006, 2008, 2012, 2015, 2018; Fulda 2016; Jones 2016; Nicholson 2013, 2018).

[12] I take intentionality in its broad meaning, as the basic *aboutness* of certain cognitive states. The intentional relation does not immediately entail object-directedness. A cognitive agent's basic openness to a world and some of its actions need not be object-directed (Hutto & Myin 2015).

Agency captures the intuition that in some systems more than just physical explanation is required to explain them. Where physical explanations can tell us *how* a physical system acts, agential explanation tells us *why* it acts. Agency is an entity's capacity to act. It is a gross behavioural property[13] of natural systems that has important parallels to the intentionality of thinking things.[14] Acting involves pursuing one's goals, which induces norms. Norms make certain actions appropriate or not given the agent's goals in a situation.[15] Norms are sets of rules that determine the good-making features of particular actions. Because acting in pursuit of goals induces norms of appropriate action, agents are rule-governed, rather than merely rule-described. Given its goals there are things that an agent ought to do. Agency contains enough normativity to potentially do justice to the obvious normativity of cognitive life.[16]

So, in sum, complex systems can demonstrate how agents can be naturally realized, and the view of agents just sketched illustrates how agential explanations apply to those systems.

The following line of thought explicates the basic logic of the issue:

(1) Complex systems are kinds of natural systems.
(2) Natural systems are amenable to naturalistic modes of explanation.
(3) Organisms are particular kinds of complex systems.
(4) Organisms are agents.
(5) Cognitive agents are kinds of organisms.
(6) If a system is amenable to a naturalistic mode of explanation, its specific kinds are amenable in the same way.

---

[13] In this context, gross behavioural properties contrast with covert behavioral properties on the one hand, and with micro-level dynamical properties on the other. In other words, we observe agency directly from how creatures act, not by inference or generalization. We also observe agents as wholes, not as parts; we observe X raising their arm, not the lower-level realizers of the arm-raising.

[14] Such a view is defended in Walsh (2012, 2013, 2018), and Fulda (2016, 2017).

[15] Norms here should be understood in the sense of hypothetical normativity. Hypothetical norms are of the following form: given that an entity has goal Z, action Y is good for it. This is a far cry from making states of affairs intrinsically evaluable.

[16] Goals are ontologically peculiar. Given a certain kind of complex system in a certain kind of environment, there are facts of the matter about what is appropriate for that system to do. But a considerable amount of ground-clearing is needed to establish this claim. Chapter 3 develops this view at length.

Therefore,

(C) Thinking things can be explained by naturalistic modes of explanation.

This argument is a radical view of the relationship between agency, cognition, and naturalistic explanation. Being radical, it is not convincing at first. This chapter aims to radically increase its plausibility.

The thesis of this chapter is that *four heterodox scientific and philosophical views are necessary to run the above argumen*t; without any of the four views, encompassing complex systems and cognitive-level agents with the same explanatory framework will be a non-starter. The heterodox views are: (H1) *goal-centered agency*, (H2) *situated cognition*, (H3) *complexity science*, and (H4) *situated Darwinism*. They are necessary because they neutralize opposing deep intuitions from four corresponding orthodox views: (O1) *reason-centered agency*, (O2) *Cartesian cognitive science*, (O3) *mechanism*, and (O4) *replicator biology*. In short, getting explanatory value out of agency requires a far-reaching, controversial, but internally coherent revision of views about action, cognition, explanation, and evolution.

What are these four orthodoxy/heterodoxy pairs?

The first orthodoxy is reason-centered agency.[17] On this view, what it is to be an agent is to be responsive to reasons. Reason-centered agency calls (4) above directly into question. If one holds reason-centered agency, then the sort of agency that organisms purportedly have is categorically different from the sort of agency that counts. On this view, the agency of organisms is ersatz-agency, and (4) fails due to equivocation between real agency and ersatz-agency.

What counteracts reason-centered agency is goal-centered agency. On this view, agency is grounded in goal-directed activity. Reason-responsiveness is not constitutive of agency; goal-directedness is. Reason-responsive systems are a special class of agent.

---

[17] The classic statement of this view is found in Anscombe (1957), and Davidson (1963/2001).

The second orthodoxy is a Cartesian picture of the mind.[18] On the Cartesian view, while all existing cognitive agents are organisms, this is merely a contingent connection. This claim calls (5) into question. While (5) may be true, on the orthodox view it is only contingently true. This blocks the connection between the explanatory modes used to understand organisms and the explanatory modes used to understand minds.

What counteracts Cartesianism is the heterodoxy of situated cognition. On this view, cognition is a more capacious notion than the manipulation of states that represent the world. Direct relations between cognition and the world are also essential for explaining cognition. A situated view of cognition allows for a more intimate relation between agency in general and the peculiarities of cognitive agency.

The third orthodoxy is mechanism. This is a view that entails that naturalistic explanations of the properties of whole systems must advert to the local interactions of a system's parts.[19] For the mechanistic view the properties of systems as wholes are explanatorily otiose. Mechanism entails that (1), (3), (4) and (5) are descriptively useful but explanatorily insufficient. This is because on the mechanistic view any phenomenon that can be explained by adverting to wholes can be explained by appeal to the system's parts. As such, complex systems, organisms and cognitive systems amount to redescriptions of mechanisms, not useful explanatory entities.

What counteracts mechanism is the heterodoxy of complexity science. Complexity science uses an anti-mechanistic methodology in which for complex systems the properties of wholes *as wholes* are explanatorily essential and in certain cases have explanatory priority over the properties of the parts. Complexity science holds that the properties of complex systems, organisms, and cognitive entities are indispensable for full understanding. The characterization of whole-system properties is essential to explaining many sorts of complex systems, of which organisms and cognitive agents are instances.[20]

---

[18] Wheeler (2005) contains a compact discussion of Cartesianism in cognitive science in its many manifestations.

[19] See Machamer et al (2000) and Glennan (2017) for an overview of contemporary mechanistic philosophy.

[20] As indicated in the introduction, these epistemic virtues redound back to ontology. That is, there is some reason to believe that our best science tells us that the phenomena countenanced by complexity science are real. However, the relation between mechanistic explanation and the teleological explanations found in agency have an intertwined relationship. This relationship will be explored in Chapter 2 and Chapter 3.

The fourth orthodoxy is replicator biology. This is the view that organisms are nothing but aggregates of replicator entities. Replicator biology views organisms as just complex mechanisms shaped by evolution by natural selection. Replicator biology calls (4) directly into question. If organisms are not agents, the assimilation of cognitive agency to agency is nothing more than reduction to mechanisms.

What counteracts replicator biology is Situated Darwinism. Situated Darwinism is the expression of the explanatory commitments of H1-H3 in the domain of biology. For situated Darwinism, whole-organism properties are essential for understanding the evolutionary process.

The four orthodoxies conspire to generate the *prima facie* explanatory gap. O1 says that what constitutes agency is responsiveness to the norms of reason. O2 says that what constitutes cognition is the manipulation of representational states. O3 says that any naturalistic explanation must cite causal mechanisms. Taken together, O1-O3 entail a hard distinction between agency and cognition, and require the use of mechanistic decomposition for both agency and cognition. O4 adds to the gap by dissolving the relevant organismic wholes into a mechanistic collection of replicators. All in all, agency and cognition are distinct, and agency is not explanatorily useful. I take some version of O1-O4 to be the set of assumptions most people hold.

The four heterodoxies constitute an alternative, cohesive system of thought.

The goal of this chapter is to offer an exposition of and motivation for holding the four heterodoxies. Section 1.1 develops and motivates organism-centered agency. Section 1.2 describes situated cognition and the essential link to biological embodiment that falls out of that framework. Section 1.3 describes complexity science and the hierarchy of kinds of complex systems that will be relevant to the explanatory claims of the above argument. Section 1.4 contrasts Situated Darwinism and replicator biology. Section 1.5 integrates the essential conclusions of 1.1-1.4.

1.1 Heterodoxy 1: Goal-Centered Agency

The thesis of this section is that what makes an entity an agent is having its own goals. All and only such entities can have dynamics that can be considered proper actions because only such entities can pursue goals. This is the goal-directedness view of agency. The goal-directedness

view contrasts with the orthodox intuition that what makes an entity an agent is its reason-responsiveness.

Taking on goal-directed agency is essential for naturalizing it. This is because a naturalistic account of goal-directedness is considerably more accessible than a naturalistic account of reason-responsiveness. If agency is goal-directedness, its roots lie with organisms. On this picture, reason-responsiveness is a sophisticated *type* of agency, not the base case of agency.

If the mark of agency is that agents act in order to pursue goals, then organisms manifest this feature just as much as rational creatures do.

This section focuses on substantiating the basic premise that agency is a matter of goal-directedness. Section 1.1.1 unpacks the Anscombe-Davidson thesis about intentional action, the view behind the reason-responsiveness criterion of agency. Section 1.1.2 calls the Anscombe-Davidson thesis into question because its focus is parochial, unwarrantedly restricting the extension of agency in nature. Section 1.1.3 sketches the alternative view of agency rooted in goal-directedness.

1.1.1 Agency as Reason-Responsiveness

The reason-responsiveness view of agency holds that agents are entities that act for their own reasons.[21] Being an agent means being a system capable of apprehending reasons for action, and making one's activities responsive to the norms imposed by reasons. Citing reasons thus makes sense of the actions of agents. Reasons are typically internal states that are propositional attitudes that articulate the appropriate intentions of the agent.[22]

---

[21] The view finds classic philosophical articulation in Anscombe (1957) and Davidson (1963/2001). The thesis is foundational to debates in the philosophy of action (Smith 2012, Setiya 2018). See also Dretske (1988), and Jones (2016).

[22] Terminologically, the class of actions is broader than the class of intentional actions. But according to the reason-responsiveness account the class of *non-intentional actions* is a misnomer. Strictly speaking, such actions are happenings rather than actions, since they are not connected to agents in the right way. See Jones (2016) for a nuanced discussion of this ambiguity.

The Anscombe-Davidson thesis is the philosophical refinement of the reason-responsiveness criterion of agency. This thesis has two parts: (1) reason-responsiveness is a criterion of demarcation between action and non-action, and (2) a specification of reasons.

Anscombe (1957) offers an approach to (1):

> What distinguishes actions which are intentional from those which are not? The answer that I shall suggest is that they are the actions to which a certain sense of the question "Why?" is given application; the sense is of course that in which the answer, if positive, gives a reason for acting. But this is not a sufficient statement, because the questions "What is the relevant sense of the question "Why?"" and "What is meant by "reason for acting"?" are one and the same. (Anscombe 1957, 9)[23]

Anscombe holds that reasons are answers to particular sorts of why-questions. For example, if I ask why agent S is walking towards the ice cream truck, saying that S wants to get ice cream gives a reason for the action. From such reasons for acting, one can derive the appropriate action from a practical syllogism.

The practical syllogism here is roughly: (P1) Walking to the ice cream truck is a good way to get ice cream, (P2) S wants ice cream, so (A) S walks to the ice cream truck. The conclusion of the practical syllogism is the action S undertakes. In walking to the ice cream truck, S responds to the reasons contained in P1 and P2.[24]

In this way, A can be made sense of with a reason-giving explanation. Agents are systems for which reason-giving explanation is appropriate. The explanatory relation between S's action and S's reasons is quasi-logical.

Davidson contributes to the Anscombe-Davidson thesis by (i) further explicating "reason for acting" and (ii) adding in a requirement that reasons cause actions.

---

[23] Anscombe's concern here is with distinguishing intentional action from a class of wider, non-intentional actions. However, as it functions in the Anscombe-Davidson thesis, this has come be a criterion of action itself. Detaching Anscombe from Davidson yields a broadened conception of agency not unlike the one defended in this dissertation. See also the discussion of Anscombe and Davidson in Chapter 4.

[24] This is a very rough sketch of practical reasoning. Further discussion of practical reasoning can be found in Davidson (1963/2001) Broome (1999) and Wallace (2020).

Reasons for acting unpack to primary reasons:

> R is a primary reason why an agent performed action A under description d only if R consists of a pro attitude of the agent towards actions of a certain property and a belief of the agent that A under the description d has that property. (Davidson 1963/2001, 3)

So, having a primary reason for S consists of having a pro-attitude towards ice-cream-attaining action (S wanting ice cream), and knowing that walking in that direction is an ice-cream-attaining action (S apprehending that ice cream trucks are good places to get ice cream). Typically, the components of a primary reason are expressed only partially, such as in saying that S wants ice cream makes sense of S doing A. Responding to the primary reason makes A intentional, i.e., the action of an agent.

For Davidson, primary reasons are causes of actions. A primary reason describes an internal propositional attitude, a belief-desire pair, that suffices to bring about S's action.[25] Reason-giving explanation is a species of ordinary causal explanation.

Anscombe and Davidson both think that agency requires: (i) having a propositional attitude toward a state of affairs, and (ii) apprehending that this attitude applies some normative requirement to act in accordance with the attitude. The explanation of S's action is an intellectualist one in citing S's propositional attitudes towards states of affairs, and what makes S susceptible to this form of explanation is that S has these propositional attitudes.

However, Anscombe and Davidson also part ways in some aspects of their account of reasons.

For Davidson, the relation between action and reason is causal. Furthermore, it is the relation of an internal propositional attitude to an action. The ability to have primary reasons demands sophisticated deliberative, self-conscious capacities. This has the effect of making language-like cognition a criterion of agency itself.[26]

---

[25] Davidson (1963/2001).

[26] Davidson (1963/2001).

For Anscombe, the relation between reasons and actions is quasi-logical and axiological, not causal. It describes what would be good for the agent in the action, presenting the action under the guise of the good. Anscombe's analysis is intended to describe the logic of the psychological concept of intentional action.[27] This does not fit with the causal aspect introduced by Davidson.

> [I]f [this] account were supposed to describe actual mental processes, it would in general be absurd. The interest of the account is that it describes an order which is there whenever actions are done with intentions. (Anscombe 1957, 80)

This difference will be important in accounting for cognitive agency later in this dissertation. But for present purposes, even Anscombe's definition of intention requires providing reasons as answers to why an agent acts.

In sum, the Anscombe-Davidson thesis is the orthodox starting point for theorizing agency. According to it, the appropriateness of reason-giving explanation is the mark of agency. All and only reason-responsive entities are agents. If a process cannot be fitted under the guise of reason-responsiveness, it is not the action of an agent. This means that accounting for agency requires the apparatus of propositional attitudes, making the criterion of agency an intellectualist one.

1.1.2 The Limits of Reason-Responsiveness

The thesis of this section is that the Anscombe-Davidson thesis expresses a parochial bias. The parochial bias comes from the definition of primary reasons, which limits the extension of natural agency. Many who have thought about the roots of action theory have expressed concern that the reason-responsiveness criterion of agency is too restrictive.[28]

---

[27] Wiseman (2016).

[28] For example, Burge (2009, 2010) has defended a view of primitive agency that extends to many organisms. Lyon (2006) has argued for such a concept as a novel foundation of cognitive explanation. Additionally, Juarrero (1999), Hanna & Maiese (2009), Velleman (2010) offer disparate critical takes on the foundations of the Anscombe-Davidson thesis. Juarrero (1999) and Hanna & Maiese (2009) argue that the ambit of action must include many sorts of typically neglected bodily processes. Velleman (2010) argues that the notion of "reason for acting" is internally inconsistent.

Frankfurt (1978) articulates the general attitude of the objection well. He holds that the difference between reason-responsiveness and mere goal-directedness should not be foundational to how action and agency is conceived:

> We [humans] are far from being unique either in the purposiveness of our behaviour or in its intentionality. There is a tendency among philosophers to discuss the nature of action as though agency presupposes characteristics which cannot be attributed to members of species other than our own. But in fact the contrast between actions and mere happenings can readily be discerned elsewhere than in the lives of people. (Frankfurt 1978, 78).

And so:

> …[w]e must be careful that the ways in which we construe agency and define its nature do not conceal a parochial bias, which causes us to neglect the extent to which the concept of human action is no more than a special case of another concept whose range is much wider. (Frankfurt 1978, 79).

This parochial bias is the requirement that agency is reason-responsiveness as understood in the Anscombe-Davidson thesis. This means that every action must be caused by a primary reason. The existence of a primary reason entails that the agent takes a deliberative, self-conscious attitude to the action.

As Davidson puts it:

> [A]n agent A intentionally φ-s just in case A's φ-ing is caused by A's intention to φ, where an intention is an all-out evaluative judgment about the desirability of the intended action (Jones 2016, 3)

This is an intellectualist view of agency. On this view, the capacity to act is closely related to the capacity for such acts as all-out judgments, if not the exercise of them.[29] But there is no reason, other than strong presupposition, to demarcate the bounds of agency in this way.

---

[29] See Jones (2016).

What does this broader sense of action look like? Two classes of actions point to a view of agency that does not depend on reason-responsiveness: the active movements of nonhuman animals, and the purportedly mindless actions of human agents.

Nonhuman animals of all sorts engage in activities that are intelligible in light of their various goals. For example, gazelles jump away from predators mostly in cases where such actions are appropriate. However, nonhuman animals arguably do not have the capacity for all-out judgments about the desirability of their intended actions. Nevertheless, the behaviours are clearly the animal's actions.[30] So there seems to be a class of actions that is not reason-responsive. The lack of reason-responsiveness in the Anscombe-Davidson sense does not invalidate the agency of nonhuman animals.[31] The gazelle as a non-linguistic creature is not capable of having propositional attitudes to its internal states. Nevertheless, the property of being appropriately in touch with the unfolding of its body movements is a common feature of the capacities of humans and gazelles as similar sorts of organisms.[32]

For mindless action of cognitive agents, there are broad classes of behaviours, movements, and even thought patterns that are clearly attributable to an agent, but are not explainable by criteria of reason-responsiveness. These actions are spontaneous actions, akratic actions, and desire-overriding pro-attitudes. For example, agent S drumming her fingers on the table, or S engaging in procrastination, or S powering through writer's block on nothing but determination.[33]

The intuition behind assigning agency to this broader class of actions is well-articulated by O-Shaughnessy (2008):

---

[30] The formal criteria for this claim will be unpacked in great detail in Chapter 2.

[31] Here one might object that the actions of the gazelle are merely attributed, not intrinsic to the gazelle. But to be consistent about that objection the skepticism about action must also apply to other humans as well, which begs the questions against theorizing agency at all.

[32] Here I must note an important difference between Davidson and Anscombe. Whereas Davidson holds self-conscious deliberation to be necessary for action, Anscombe holds that many agents can have reasons without them being articulable. The discussion of goals in agents captures this sense of non-articulable reason, and the difference in views might amount to a terminological one.

[33] Anti-Davidsonian arguments from spontaneity, akrasia, and desire-overriding reasons are developed Hanna & Maiese (2009), but are beyond the scope of the present discussion.

> Does nothing lie between a corpse-like graven image and a vehicle for reason? How else but as action is one to characterize the making of these movements, and to what but the person is one to attribute them? One can hardly telescope them into mere spasms [...] They relate to standard examples of action somewhat as do objects that are mere lumps of stuff, say raw diamonds, to objects that are mere lumps of stuff and more, e.g., artifacts, natural kinds. (O'Shaughnessy 2008, 54-55)

There is no knock-down argument against the reason-responsiveness view of agency. The way is open to the proponent of reason-responsiveness to produce a wider, liberal understanding of pro-attitudes and mental states. But to the extent that one liberalizes pro-attitudes and mental states is the extent to which one drifts towards the alternative intuition of goal-directed agency.

Taken together, the phenomena of nonhuman animal agency and of "arational" or otherwise non-reason-responsive cognitive agency point to a broad class of action that is missed by the Anscombe-Davidson thesis. There are many creatures that act but not for reasons. Even for creatures capable of apprehending reasons as reasons, the Anscombe-Davidson characterization of their actions is too narrow. In sum, many natural systems—in particular complex organisms—act as agents by pursuing goals, but do not meet the stringent requirements on agency imposed by the Anscombe-Davidson thesis. Such limitations in part motivate an alternative account of agency.

1.1.3 Goal-Centered Agency

The thesis of this section is that goal-directedness is the appropriate starting point for agency. More precisely, rather than basing agency on reason-responsiveness, agency is based on goal-directedness. On this view, all agents are goal-directed and a process counts as an action if it contributes to the agent's pursuit of its goals. Organisms are the most general class of entities that pursue goals.[34]

[34] The thesis that organisms are goal-directed and thereby agents is technically independent of the thesis that agents are goal-directed. The essential goal-directedness of organisms will be developed in 1.4. Here the focus is on goal-directedness itself.

The goal-directedness account of agency denies the Davidsonian requirement that agency requires a capacity to have primary reasons. Other sorts of motivational states can explain the actions of agents. As discussed in 1.1.2, several classes of actions fall short of being articulable in all-out evaluative judgments but nevertheless seem like actions of an agent in a less stringent sense. There are likely several distinct ways in which the above-discussed states fall short of being articulable as all-out judgment. Nevertheless, they are appropriately integrated into the totality of an agent's goals. For example, the gazelle that jumps away from the predator has the goal of predator-avoidance, the jumping contributes to attaining that goal, and this makes the jump the gazelle's action. Similarly, procrastinating serves a function within S's overall structure of goals, even if not articulable by S herself.

While dropping the Davidsonian requirement, the goal-directedness account of agency is closer to Anscombe's view in that actions are demarcated by their suitability to a certain kind of teleological explanation. What unifies the problematic cases is that they can be explained by appealing to an agent's pursuit of its goals. This suggests that the proper account of agency appeals to goals, not reasons.

An explanation that cites goals is teleological. A teleological explanation is one where the presence, occurrence, or character of some process is explained by the end or goal that it serves. In the context of action explanation, a teleological explanation answers a why-question in the same way that a rationalizing explanation does. It cites a goal as a way of making sense of an occurring action.[35]

Although rationalizing explanations have different content, the form of both rationalizing and teleological explanations is the same. In terms of content: in the rationalizing case, one appeals to primary reason, whereas in the teleological case, one appeals to the goal. The differences between the two explanatory styles come down to the difference in the capacities of the entities that are being explained, *not* to the sense of "why?" being answered.[36] Reason-giving

[35] Teleological explanation is discussed extensively in Taylor (1964), Bedau (1991, 1992), Juarrero (1999), Walsh (2008, 2012), and Fulda (2016).

[36] Anscombe (1957) says that the same sense of "Why?" applies in explaining human behaviour and the behaviour of some animals (1957, 86).

explanation is a rarefied type of teleological explanation on this view, one that fails to apply to certain cases that intuition classifies as actions. In a slogan: teleological and rationalizing explanations have different content but the same form.

A non-parochial account of agency naturally follows. The same teleological form of explanation applies in both cases. Both humans and nonhuman animals are goal-directed, at least in basic ways.

The basic idea of an alternative approach is that cognitive agents are constitutively bound by wider, biological agency.[37] That is, reason-responsiveness operates against a background of other kinds of normativity.

Just as having intentions and desires introduces normative demands on an agent, so staying alive and pursuing one's biological way of life does too. If the mark of agency is that agents can make things happen that otherwise would not happen, in order to pursue a goal, then organisms manifest this feature just as much as rational creatures do.

So, the heterodoxy is that to be an agent is to be a system that pursues its own intrinsic goals. Reason-responsiveness is just a particular form of the more general, goal-directed capacity. This view of agency offers a natural conceptual framework within which more specific accounts of broad phenomena that are intuitively actions can be offered.[38]

In sum, the limitations of reason-responsiveness as an account of agency suggest an alternate, goal-directedness-based, account of agency. Actions are processes in organisms that can be effectively explained teleologically.[39] This view of agency entails an intimate relationship between goals in general and reasons in particular; reasons are sophisticated, rarified goals.

---

[37] The idea of bio-agency is explored in Skewes & Hooker (2009).

[38] In chapter 4 of this dissertation, I will take steps towards re-integrating reason-responsiveness into an account of agency. But an account of spontaneity, akrasia, and desire-overriding reasons are beyond the scope of the present project.

[39] Teleology explains not just actions, but developmental processes as well. Developmental processes are arguably not actions. As discussed in section 1.3, and Chapter 3, actions are goal-directed processes with additional features.

1.2 Heterodoxy 2: Situated Cognitive Science

The thesis of this section is that cognition is a situated phenomenon. It can only be understood in the context of an agent embedded in its setting because the setting is partially constitutive of cognition itself. Treating cognition as a situated phenomenon is a radically heterodox move. Situated cognition is a systematic rejection of Cartesianism about cognition. The most important immediate consequence of endorsing situated cognition is that life-like and mind-like explanation can be thought of as continuous.[40] The most important consequence of situated cognition is that complexity science and systems biology are explaining aspects of cognition, not just implementation details as the Cartesian view would hold.[41] Certain embedded complex dynamics are partially constitutive of cognition. Similarly, the capacities of organisms as organisms are partially constitutive of cognition.

This section sketches the two frameworks of Cartesianism and situated cognition. The choice between the two frameworks is in some sense a matter of philosophical temperament. The aim of this section is to make the contrast sharp, and to indicate that both frameworks are internally consistent.

1.2.1 Orthodoxy: Cartesianism about Cognition

Cartesianism is both a metaphysical stance on the mind and a methodological stance on what sort of activity counts as cognitive and how one goes about identifying it.[42] Nobody in cognitive science holds to metaphysical Cartesianism, but methodological Cartesianism is widespread.[43] Methodological Cartesianism is an *explanatory dualism* between cognition and its implementation. This is the view that:

---

[40] Mind-like explanation covers what intuitively goes under psychological, information-processing, folk-psychological, or intentional stance explanation. These explanatory frameworks are themselves not typically demarcated cleanly. For an overview, see Bennett et al (2007), Dawson (2013) and Bennett & Hacker (2008).

[41] As chapter 4 will discuss, the situated view of cognition is in *prima facie* tension with the view that cognition is desituated agency. But after fully outlining natural agency, this tension will be shown to be inessential.

[42] For a comprehensive discussion of Cartesianism in cognitive science, see Wheeler (1997, 2005, 2011). See also the discussion of neo-Cartesianism in Haugeland (1990). Cartesian cognitive science is explicitly expressed in Fodor (1975, 1980, 1981, 2008), Fodor & Pylyshyn (1988).

[43] Some references to this position use the term neo-Cartesianism (e.g., Haugeland 1990). Nothing hangs on this terminological difference in this context.

> [T]o explain physical phenomena, one need appeal only to specific physical entities or states, and to specifically physical laws; (ii) to explain psychological phenomena, one need appeal only to specifically mental entities and states, and to specifically mental laws. (Wheeler 1997, 3)[44]

Cartesianism entails that although cognition can be *realized* or *implemented* in organisms, the generalizations and invariances which we use to explain cognition are autonomous from the physical laws operating on the realizer. Invoking the physical laws operating in a cognitive organism offers at best a shallow explanation of that organism's cognition. This amounts to denying that naturalistic explanation at the organismic level transfers to the cognitive level.[45]

The Cartesian explanatory framework as a conjunction of eight theses:

> (C1) The primary characteristic of a cognitive agent's epistemic situation is the experience of a subject-object dichotomy.
>
> (C2) Explaining mental processes means explaining the manipulation of representational states within the cognizer.
>
> (C3) The bulk of intelligent action is the product of general-purpose reason; i.e., reasoning processes that adaptively retrieve and modify mental representations.
>
> (C4) Perception is essentially inferential in nature. Cognitive systems use representational manipulation to fill in gaps in input and construct a coherent representation of the world.[46]
>
> (C5) Perceptually guided intelligent action is organized based on a fundamental sense-represent-plan-move motif.
>
> (C6) The intelligent agent is fundamentally explanatorily disembedded from its environment. The environment impinges on the agent by providing (i) problems for the

---

[44] Talk of "laws" is unnecessarily restrictive in this context. Any general invariance relation that is suitable for explanation works in place of laws here. Very few psychological explanations can be assimilated to the model of explanation by laws. For an overview of the invariance-based view of explanation, see Woodward (2014).

[45] More precisely, Cartesianism entails that the biogenic approach to cognition is a non-starter (Lyon 2006). This is an approach that locates the essential features of cognition in life and builds out from there.

[46] "The current Establishment theory (sometimes referred to as the "information processing" view) is that perception depends, … upon inferences. … And, finally, the Establishment holds that the psychological mechanism of inference is the transformation of mental representations. It follows that perception is in relevant respects a computational process." (Fodor & Pylyshyn 2002: 167-8)

> agent to solve, (ii) information to the mind, and (iii) a stage on which actions are executed.
>
> (C7) The operating principles of the generation of intelligent action are conceptually and theoretically independent of the understanding of the agent's physical embodiment.
>
> (C8) Psychological explanation is temporally austere; good scientific explanations of mental phenomena do not typically appeal to richly temporal processes. (Paraphrased from Wheeler 2005, 21-53)[47]

These eight theses support each other tightly. The ontology of cognition explicit in Cartesianism entails that the realization of cognition in organisms vastly underdetermines how cognition is to be explained.[48] Cartesian cognitive science treats situatedness as merely a question of how to get a cognitive architecture linked up to the body and world in the right way.

Because of the interlocking quality of the framework Cartesianism cannot be rejected piecemeal, but only systematically.

Cartesianism entails that it is futile to assimilate cognitive agency to agency writ large. Even if all cognitive agents are organisms, and even if organisms are essentially agents (as suggested in 1.1), it is still an open possibility that cognitive organisms are merely a particular implementation of an independent cognitive architecture. What it is to be a living thing has no bearing on what it is to be a cognitive thing for Cartesianism.

1.2.2 Situated Cognition

Situated cognition is here a catch-all term for a view of cognition that systematically rejects Cartesianism.[49] Haugeland (1998) illustrates the essence of situated cognition using slightly different terms:

---

[47] This interpretation of Cartesianism in cognitive science is the steel-manned version. C1-C8 are broadly inclusive and softer than some of Descartes' own positions. For a discussion of the exegetical subtleties in characterizing Cartesianism, see Wheeler (2005), Chapter 2.

[48] For example, all existing cognitive agents are situated and that situatedness places definite constraints on their information processing, but for Cartesianism issues around such situatedness are peripheral to explaining cognition as such. The Cartesian strategy in instances where situatedness matters is to shunt the matter off to the issue of "transduction". Issues of transduction have to do with how one hooks up a cognitive system to the world in the right way. This, of course, presupposes that cognition can be characterized independently of such interaction.

[49] Situated cognition covers a wide cluster of non-Cartesian approaches in cognitive science. Many terms indicate pieces of this view: anti-representationalism (Brooks 1991, Varela et al 1991, Chemero 2009, Hutto & Myin 2013),

> The contrary of this [Cartesian] separation ... is something I would like to call intimacy of the mind's embodiment and embeddedness in the world. The term "intimacy" is meant to suggest more than just necessary interrelation or interdependence but a kind of commingling or integralness of mind, body and world—that is to undermine their very distinctness. (Haugeland 1998, 208)

More explicitly, situated cognition consists of eight theses paralleling the Cartesian ones:

> (S1) The primary characteristic of a cognizer's epistemic situation is absorbed coping with its world.
>
> (S2) Explaining mental processes means primarily explaining how a cognizer makes sense of its situation. Explanation in terms of the manipulation of representations depends on the logically prior process of sense-making.
>
> (S3) The distinction between general-purpose reason and special-purpose cognitive faculties is not a strict dichotomy. All cognitive faculties are limited in some sense.
>
> (S4) Perception is not inferential in nature. Perception is direct, for action, and of affordances.[50]
>
> (S5) Perceptually guided intelligent action is based on a fundamental sense-act motif.[51]

---

embodied, embedded, extended, or enactive (E4) cognition (Wheeler 2015), dynamical cognitive science (Juarrero 1999, Thompson 2007), Heideggerian AI (Wheeler 2005, Dreyfus 2005). There are a number of characterizations of this position in the literature, all sharing a family resemblance. The most systematic accounts can be found in Hutto & Myin (2013) and Gallagher (2017, 2019). It is worth noting that Wheeler (2005) would not endorse all of S1-S8. See Steiner (2019) for an assessment of the matter from the perspective of philosophy of science.

[50] "Perception is not an inner representation of an objective world, but a relation of inhabiting a world... the 'world' for us is more than simply the spatial container of our existence. It is the sphere of our lives as active, purposive beings. .... the world is the place that we 'inhabit', rather than simply a set of objects that we represent to ourselves in a purely detached way" (Matthews 2002, 49).

[51] "An agent's environment is filled with light, reflected off its various surfaces. Because of the surfaces and textures in the environment, this light has a determinate structure—the ambient optical array. Vision occurs directly, by an agent 'sampling' the ambient optical array. The agent samples the array by moving through it. Movement exposes the invariant structures of the environment. Edges, planes, barriers are detected by seeing the way their agent-centered relations change as the agent moves. Surfaces that had been occluded by others become disoccluded and

> (S6) The intelligent agent is fundamentally situated in its environment. Agent and environment involve relations of both rich feedback and co-constitution.
>
> (S7) The operating principles of the generation of intelligent action are commingled with the understanding of the agent's physical embodiment.[52]
>
> (S8) Psychological explanation allows for rich temporality; good scientific explanation of mental phenomena will typically advert to the rich temporality of situated processes.

Whereas the Cartesian view is familiar, the Situated view needs some unpacking.

S1, absorbed coping, is a basic claim about what phenomena constitute cognition. It opposes the Cartesian view that cognition is constituted by the internal activities of a subject as it deals with a world. While common sense takes cognitive agents as primitively set against a world of objects, the situated perspective breaks below common sense and holds that cognition is constituted by absorbed skillful coping with the world. The background conditions of absorbed coping with the world make the common-sense view possible.

> [T]he background is the vast, holistic, indeterminate, and therefore *unrepresentable*, web of embodied, psychological, social, and cultural structures that constitute one's world and that are implicitly presupposed by concrete examples of human sense-making. (Wheeler 2016, 98)

In situated cognition, absorbed coping is more fundamental than experience structured in the subject-object mode. As such, absorbed coping is critical for a full explanation of cognition.[53]

---

vice versa. The key to perception is action: it proceeds through action and it delivers action-oriented content." (Walsh 2012, 14)

[52] Co-constitution is a tricky concept. Agents and environments are not constituted in a synchronic fashion, but diachronically. This allows one to hold the more sensible view that agents and environments can exist independently of each other, at least for a time.

[53] This point has been made in many contexts. In the present context, the Heideggerian critique of artificial intelligence relies on this observation. The Heideggerian critique of AI draws on Heidegger and Merleau-Ponty's analyses of the background coping as a requisite for cognition. See Dreyfus (1972, 1994, 2007, 2013), Wheeler (2005), Thompson (2007), Schear (2013). On the limitations of the subject-object split as the basis of cognition, see Zahavi (2018).

S2, sense-making, is a claim about psychological explanation.[54] In situated cognition, explanation that cites the manipulation of representations depends on a prior process of sense-making. Representations cannot explain sense-making because their content does not have the right sort of context-sensitivity. The very notion of stable content—even context-sensitive content—presupposes that a cognitive state is the same state across multiple contexts. In contrast, sense-making is a more thoroughly context-sensitive process. Sense-making arises out of the way that an agent copes with, exploits, ameliorates the implication of its setting or the pursuit of its goals. On this view sense-making is non-representational. The situated view of cognition takes the relative context-independence of representational states as something to be explained, not the primitive explainer of mental processes.

The frame problem is an illustration of the necessity of sense-making. It is the problem of how a cognitive system can capture context-sensitive relevance. The frame problem is something that any cognitive system of limited capacities seems to encounter, and solve. Any action of a cognitive system causes many unexpected side-effects. In order to manage action, limited cognitive agents must ignore the vast majority of these side-effects, otherwise they risk cognitive paralysis because there are simply too many side-effects to consider.[55]

Cartesianism leads to the presupposition that in order to deal with or ignore a feature of its environment, an agent must first explicitly represent it, and then determine its relevance for the agent, and then determine what action would be appropriate given the feature and its relevance. The richness of the environment and the various ways in which each feature might be relevant vastly outstrip the cognitive capacities of any finite agent.[56]

The frame problem might be usefully approached by opening up cognitive explanation to processes that are non-representational. In particular, massively parallel, context-sensitive, and sub-semantic processes offer some hope of sidestepping the frame problem. In this way, the

---

[54] For an overview of sense-making, see Thompson (2007), Thompson & Stapleton (2009).

[55] See Dennett (1987), Wheeler (2005), Vervaeke et al (2009) and Shanahan (2016) for extended illustrations of the frame problem.

[56] Shanahan (2016) glosses the frame problem thus: "The epistemological problem is this: How is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language-like representations of the sort used in classical AI?"

frame problem offers a *prima facie* motivation for non-representational states in explanations of cognition.[57] The upshot here is that situatedness, which posits absorbed coping as partly constitutive of cognition, has a set of relations that avoid the need to represent features of situations. Absorbed coping takes some of the explanatory burden from the cognitive system's framing of situations.

S3, special-purpose reason, blurs the dichotomy between general-purpose and special-purpose cognitive faculties. In situated cognition there is no general-purpose reason. The situated view emphasizes the shared characteristics of all cognitive processes, namely that they are inherently prone to blind spots and are organized to trade off strengths and weaknesses with other cognitive faculties.[58]

For example, there is an inherent trade-off for a cognitive system between exploiting present information and exploring to find new information. Neither strategy is optimal in all situations. A general architecture of opponent processing preserves both cognitive faculties while allowing for a degree of flexibility in when and to what extent such strategies are applied.[59]

S4, direct perception, denies the view that perception is inferential; i.e., that it is a process of using representations to fill in gaps in the information received from the world. S4 supports the ecological approach to perception. On this view, perception is (i) direct, (ii) for action, and (iii) of affordances.[60] Affordances are the "object" of perception. The primary situation of a perceiving agent is to notice the opportunities for its own actions. Affordances are relational properties between the capacities of the agent and the features of its environment.[61] This feature

---

[57] See Wheeler (2008).

[58] See Vervaeke et al (2009) for a discussion of the conceptual pervasiveness of trade-offs in learning as well as meta-leaning strategies in cognitive science. The details are beyond the scope of the discussion here.

[59] See Vervaeke et al (2009). For an example from neuroscience, see Barbey (2018).

[60] For an overview of the ecological approach to perception, see Chemero (2009), Walsh (2012) and Lobo et al (2018). Denying that perception is inferential does not immediately entail the ecological interpretation of perception, but here I assume that the ecological approach is the most promising denial of inferential perception.

[61] For a discussion of the ontology of affordances, see Chemero (2003, 2009). The view of affordances as inherently relational properties follows Chemero's conception.

of affordances makes them inherently contextual unlike representational states whose content is to some degree context-independent.

S5, the sense-act motif, holds that the sense-act cycle is the basic unit of cognition. On the situated view, a relatively direct connection between sensing and acting is all that is needed for most cognitive processes, particularly those that constitute absorbed coping. This denies the Cartesian assumption that representation and planning are necessary for action.[62]

S6, co-constitution, is the claim that cognition is richly embedded in the world. This means that the Cartesian explanatory strategy of abstracting away from the particulars of implementation will lead to failure of descriptive adequacy. S6 suggests a new ontology of cognitive processes. The most salient consequence of S6 is that it complicates the distinction between what constitutes cognition and what is merely a causal input into cognition. One way to understand the breakdown of the distinction is to view cognition and the world as co-constituting. The agent's world is best understood as made by the agent's active coping with it.

S7, embodiment, is the claim that embodiment and its constraints are partly constitutive of the operating principles of intelligent action. Because embodiment is essentially involved in cognition, these operating principles of intelligent action are sub-semantic, not semantic.

Sub-semantic operating principles of intelligent action are well-illustrated in the idea that cognition aims to optimize constraints and trade-offs between myriad bioeconomic properties. Bioeconomics has to do with the way biological systems optimize, manage, and maintain energetic and metabolic resources.[63] The relevance of bioeconomics to cognition would not make sense unless embodiment was partially constitutive of cognitive operation.

The theory of relevance realization exemplifies explanation in bioeconomic terms.[64] It is a theory of the ways in which dynamical properties and trade-offs within a cognitive agent result,

---

[62] The details of this are explicated well in Wheeler (2005).

[63] These ideas are discussed in various Free Energy approaches to brain function (Friston 2009, 2010), and in theories of cognition that take energetic constraints as partially constitutive of cognitive processing (Vervaeke et al 2009).

[64] See Vervaeke et al (2009).

in a process strongly analogous to evolution by natural selection, in an optimization of cognitive resources in myriad situations. This is required for agents to achieve appropriately context-sensitive relevance which serves as a solution to the frame problem. The theory works at a level less abstract than Cartesian cognitive architecture, but also more abstractly than at the level of the details of embodiment.

Such commingling of cognitive and biological explanation relies on the thesis of the deep continuity of life and mind.[65] Deep continuity is the view that that the essential organizational features that make cognitive agents are an elaborated version of the essential organizational features that make an entity living.

If S7 is true it entails a qualified version of the multiple realizability thesis. Situated multiple realizability needs to include certain non-formal parameters of timing and appropriateness in addition to formal criteria when determining whether the same cognitive system is realized in multiple media. Robots can be cognitive agents, but they will be so in virtue of the realizers of their situatedness in addition to their information-processing capabilities.

S8, rich temporality, is a claim about the psychological explanations that will do the most justice to their target phenomena. It is an empirical bet that for phenomena of absorbed coping, adverting to the rich dynamics of situated agents will constitute better explanations. It is a claim that it is explanatorily inappropriate to abstract away from corporeality and temporality.

S1-S8 form a set of mutually supportive claims. The set contains both a claim about what constitutes cognition and how one goes about studying it. For instance, appeal to bioeconomic properties as a fundamental explanatory motif for cognition only makes sense within a situated framework. Conversely, without the empirical work generated by the more specific theoretical commitments, the fundamental extension of cognition to phenomena that do not fit the subject-object structure of experience appear stipulative. But as a unified set of theses situated cognition is a thoroughgoing rejection of Cartesianism.

---

[65] See Thompson (2007). For weaker versions of life-mind continuity, see Wheeler (2012) and Godfrey-Smith (1996).

Cognitive systems, on the situated view, are partially constituted by processes that Cartesianism would classify as inputs into cognition.[66] The upshot of that the situated framework the cognitive agent and its world are co-constituted. The idea of co-constitution is best approached through the agent's relations of affordance with respect to the environment. The world of a situated cognitive agent is determined in part by the capacities of the cognitive entity. This theoretical commitment opens up a radical revision of the relationship between agents and environments.[67]

The operation of cognition in the Cartesian sense, i.e., in the subject-object mode, depends on this more primitive co-constitution. The commitments of Cartesianism can be seen as a powerful but incomplete set of methodological commitments in the characterization and explanation of cognition. The salience of the subject-object mode of cognition is insufficient to motivate the subject-object mode as a demarcation of cognition.

There is no definitive reason to adopt the Cartesian or situated view.[68] However, the relevance of the alleged success, failure, or standstill of particular research projects in cognitive science depends on prior commitments to one view or the other.[69] In this situation, endorsing situated cognition is a philosophically motivated empirical bet. Cartesianism and situated cognition are both internally coherent, but they make opposite methodological demands about nearly all aspects of cognitive science.

Situated cognition eliminates the explanatory gap between the principles that describe life and the operating principles of cognition.[70] The situated framework requires the explanation of

---

[66] This move is the reason why situated cognition tends to deflate general-purpose reason. Situated cognitive agents are adaptive collections of variegated limited-domain capacities. Their appearance of general-purpose capacities is a consequence of proper integration of limited-purpose capacities.

[67] Details of this co-constitution relation are developed in Chapters 2 and 3 of this dissertation.

[68] One might read the emergence of situated cognition as a reason to hold that cognitive science is fragmenting and that cognition as an object of study is not a unitary phenomenon. See for example Dawson (2013). Evaluating this claim is beyond the scope of this project. Clearly, the present discussion assumes that cognition is at least minimally unified as a subject matter.

[69] As a specific example, the longstanding debate between representational and anti-representational views of cognition turn on prior commitments to C1-C8 or S1-S8. See Verdejo (2014), Steiner (2014), Ramsey (2017), Clark (2015), Sebastian (2017), Downey (2018)

[70] Whereas on the Cartesian view the primary explanatory demand is to get the architecture of cognition right, and then to address the secondary demand of fitting that architecture with the world, on the situated view the primary

cognition to incorporate rich temporality, bioeconomic properties, and the ecological embeddedness of the cognitive agent. Such principles are also essential for explanations of organisms in their environments. This entails that claiming that natural agency is only one possible implementation of cognition does not negate the importance to cognition of the explanatory frameworks that are appropriate to natural agency.

1.3 Heterodoxy 3: Complexity Science

The thesis of this section is that grade III complex systems, also known as adaptive autopoietic (AA) systems, are the physical realizers of natural agency. Complex systems come in three grades of complexity, each with a characteristically deeper set of capacities for the whole system to determine the dynamics of their parts.

Simply put, grade I complex systems are systems that realize of work-constraint cycles. The work-constraint cycle is a pattern of material and energy flow that results in the persistence of ordered structures. Such systems harness material and energy flow in such a way they maintain ordered structures that would otherwise be quite improbable. Grade II complex systems are systems that achieve a closure of constraints. This is a network of constraint whereby the system's organization itself persists as an ordered structure. Grade III complex systems are systems that achieve, in addition to a closure of constraints, the capacity of adaptivity. Adaptivity allows such systems to robustly and plastically pursue states that counteract not only the disintegration of the system, but *tendencies* toward the disintegration of the system.

This section unpacks the conceptual moves and terminology used to characterize complex systems.

Complexity science is the explanatory framework that links general physical principles and natural agency. But this link is not immediate. The connection between goal-directed agency and situated cognition runs through grade III complex systems.

---

explanatory demand is to get the agent-world fit right, and then secondarily to explain the agential capacities that support that fit.

Complexity science is a theoretical and methodological framework that exists as an alternative to the strategy of understanding systems by decomposition to their parts.[71] Complexity science is premised on the claim that even if component parts of complex systems are well-understood, explaining some characteristics of the whole system is a substantive scientific task over and above the explanation of its components.[72] Unlike decompositional analysis, where explaining a system means reducing its dynamics to the intrinsic dynamics of its parts, analysis in complexity science explains using whole-system properties. These properties include; global relations between the system's parts, network topologies, chaotic dynamics, and general tendencies of the whole system not present in any of its parts.[73]

The structure of this section is as follows. Section 1.3.1 sketches two fundamental structural motifs of all complex systems: (i) distance from thermodynamic equilibrium and (ii) the realization of work-constraint cycles. Section 1.3.2 gives a general definition of complex systems as systems that realize an essential tension between integration and differentiation. Section 1.3.3 sketches whole-to-part ("downward") determinative influence, which is the explanatory bedrock of explanation in complexity science. Section 1.3.4 sketches the three grades of downward determinative influence, and locates agency at grade III.

---

[71] Complexity science and its associated methods are especially popular in the so-called special, sciences, in particular systems biology (Bechtel & Abrahamsen 2011), ecology (Odenbaugh 2011), and sciences where decomposition strategies are hamstrung by complex or unclear ontologies of their main objects of study, such as anthropology (Lansing & Downey 2011). However, even in the more hard sciences complexity science methods are important, particularly in studies of chaotic dynamics (Hooker 2011).

[72] In this, complexity science's methodological rival is the mechanistic strategy goes back to the start of the scientific revolution. Recently, much attention has been given to so-called New Mechanism in philosophy of science (Machamer et al 2000, Woodward 2011, 2013, Glennan 2017). Mechanism will be explored in more detail in Chapter 2 of this dissertation.

[73] For a detailed overview of complexity science in see Hooker (2011). For a discussion of ontological challenges, see Bishop (2011), Bickhard (2011). For a discussion specifically related to cognitive science, see Froese & Ziemke (2009), Froese (2010), Stewart et al (2010) and Moreno et al (2011). For historical background on General Systems Theory the progenitor of modern complexity science, see Hofkichner & Schafranek (2011). For discussion of complexity in biology, see Bechtel & Abrahamsen (2011) and Newman (2011), Hesp et al (2019). For a simple but reasonably comprehensive introduction, see Mitchell (2009).

1.3.1 Dissipative Systems and Work-Constraint Cycles

The thesis of this section is that the objects of study of complexity science all share two characteristics: (i) being far from thermodynamic equilibrium, and (ii) enacting work-constraint cycles.

The fundamental shared feature of all complex systems is that they exhibit some form of macro order or organization as a result of being far from thermodynamic equilibrium.[74,75] All such systems contain flows of matter and energy that spontaneously produce ordered structures.[76]

For example, the Earth's weather system is essentially a complex mechanism that dissipates the heat differential between equatorial regions and polar regions. In doing this, different types of large-scale weather systems form because such large-scale organization are more efficient at dissipating the heat than heat transfer without such macro order.[77]

This is an instance of the general principle that when a system is driven far from thermodynamic equilibrium by an energy gradient in the environment, the steady state that is eventually reached is the one that maximizes entropy production—that is, a state where dissipation of the driving

---

[74] Thermodynamics, in its most general form, is the science of energy and its transformations. Energy is an abstract quantity that is conserved and drives change. It may transform into different types of energy. The formalisms and laws of thermodynamics have been enormously successful in describing and explaining the behaviour of many sorts of systems, and as such they are an indispensable component of any general systems theory. The transformations of energy in thermodynamics are described by two fundamental quantities: work and heat. To put it simply, work is present in energy transformations that order energy, whereas heat refers to transformations that make energy disordered or unusable. Thermodynamically, order and disorder are understood in terms of the fundamental quantity of *entropy*, the measure of the disorder present in a system. Work is thus defined relative to heat as a transformation of energy that produces less entropy than it might otherwise; work *channels* energy release into forms other than heat (Prigogine 1980, Moreno & Mossio 2015). For example, the breakdown of glucose in a fire is a reaction that releases energy as heat. The same breakdown of glucose in metabolism generates biochemical work.

[75] Macro order sometimes refers to order at a higher level of organization, and at other times refers to order that is obvious to unaided observation. My use of the term is neutral with respect to ease of observation.

[76] It appears to be a deep principle of nature that when an open system is pushed far from equilibrium (i.e., has a non-random distribution of some form of matter or energy) it will produce internal organization in order to dissipate the disequilibrium more efficiently. This underlies the widespread existence of dissipative systems in nature. The classic example is that of Rayleigh-Benard convection, wherein for a sufficiently large difference between the temperatures between the top and bottom of a fluid layer, the fluid will organize into macroscopically visible convection cells so as to dissipate the excess heat more effectively. The general rule is that whenever there is a thermodynamic gradient, self-organization emerges spontaneously (Thompson 2007).

[77] See Ozawa et al (2003).

energy gradient into the system's environment is most efficient. Such a state is ordered so as to maximize entropy production. Such structures are called dissipative structures.

This production of dissipative structures in far-from-equilibrium circumstances is the basic principle that connects complexity science to well-understood physics.[78] To gloss the issue, once a dissipative structure is set up, order comes for free in some macroscopic domain.[79]

A fundamental motif in far-from-equilibrium systems is the work-constraint cycle.[80] This is a general description of how matter and energy flow to establish macroscopic order. The coupling of work and constraint generates persisting macro ordered structures such as weather systems. Moreno and Mossio (2015) gloss the idea in the following way:

> [C]onstraints are required to harness the flow of energy [...] so that the system can generate work and not merely heat (due to the dispersion of energy). [...] [T]he system needs to use the work generated by the constraints in order to generate those very constraints, by establishing a mutual relationship between the constraints and the work. (Moreno & Mossio 2015, 10)

All complex systems realize such coupling of work and constraint. Constraints (i) limit the degrees of freedom of a system's parts, and (ii) modify the possibilities of events within systems happening, such that unlikely events or combinations of dynamics become more likely. The former sort of constraint is a limiting constraint; the latter sort of constraint is an enabling constraint.[81]

Constraints can be externally imposed on a system or internally generated. In an engine the work is produced as a result of the external constraints imposed during the manufacture of the engine.

---

[78] This idea is sometimes referred to as Prigogine's principle. The theoretical and analytic work establishing this principle can be found in Prigogine & Nicolis (1971), Nicolis & Prigogine (1977), Prigogine (1980). For an application to biological order, in particular metabolic systems, see England (2015).

[79] See Kauffmann (1996).

[80] See Kauffman (2000), Mossio & Bich (2017).

[81] See Juarrero (1999).

In contrast, for far-from-equilibrium systems constraint is maintained by the system itself. Some of the energy flow though the system is channeled for regenerating the constraint.

In summary, all complex macro order comes from the general thermodynamic properties of systems that exist far from thermodynamic equilibrium. The work-constraint cycle is the fundamental organizational motif that channels the flow of matter and energy through the complex system and supports complexity of all sorts.

1.3.2 Integration, Differentiation, and Metastability

This section sketches the many typical features of complex systems. The thesis of this section is that a complex system is a system that realizes a dynamic tension of integration and differentiation. Integration refers to a set of system-level properties that organize components into persistent, ordered structures. Differentiation is a set of system-level properties that refer to sensitivity to and amplification of changes both within and outside the system.

For example, the human body is integrated insofar as it can resist perturbations to its structure across a range of internal and external changes; it is differentiated insofar as it contains structures that pick up on subtle changes in its internal and external environment and can sometimes amplify such changes to affect perceptual faculties or immune responses. The dynamic tension of integration and differentiation is scale-invariant, showing up across multiple levels of analysis: organelles, cells, tissues, organs, organ systems, and the whole body.[82]

Hooker (2011) also offers an account of complex systems in terms of five dimensions:

> To approach [the characterization of complexity], first omit all epistemic notions as ultimately derivative considerations, e.g. those appealing to intelligibility and surprise, and all 'external' notions like controllability. Then at the least complexity has, I suggest, five quasi-independent dimensions to it: *cardinality* (component numbers), *non-linearity* (of interaction dynamics), *disorderedness* (algorithmic incompressibility), *nested*

[82] Discussion of biological organization presuppose a dynamical tension of integration and differentiation. See Rosslenbroich (2014) and Mossio & Moreno (2015) for extended discussions of biological organization based on this line of thought.

> *organisation* (organisational depth) and *global organisation*. (Hooker 2011, 40, my emphasis).

This five-dimension model can be reduced to the claim that a complex system is one that handles the dynamic tension between integration and differentiation. Integration means that a complex system is a distinguishable unity of many components, one having *global organization*, *cardinality*, and *organizational depth*. Differentiation means that its global dynamics are not derivable from the dynamics of its bottom-level parts, which affords the possibility of amplification and adaptation, and a degree of unpredictability. The global or higher-level dynamics are not compressible due to *disorderedness*. These high-level dynamics are also prone to chaotic behaviour due to the *non-linearity* of the interactions of their component parts. In this way, the dynamic tension between integration and differentiation encompasses the five dimensions of complexity.

The tension between integration and differentiation is essential to the phenomenon of metastability that all complex systems exhibit to some extent. Metastability is

> [...] the successive expression of different transient dynamics with stereotyped temporal patterns being continuously created and destroyed and re-emerging again [...]. (Friston 2000, 238)

The basic idea here is that metastable systems operate at the “edge of chaos”; they exhibit distinguishable order at all times, but the order is mutable and difficult to describe or compress succinctly. Systems with metastable dynamics are chaotic, especially at smaller scales, but more predictable at macro scales and longer timescales. This feature is a barrier to the strategy of decomposition into well-delineated parts. Metastability has the consequence that the very stability of the parts depends on global parameters of the system.

In highly integrated systems, the components are constrained to act in largely stereotyped, context-insensitive ways. This makes them susceptible to the mechanist strategy of decomposition to their parts. Conversely, a system that is highly differentiated stretches the meaning of the word "system" since citing the internal dynamics of such a loosely organized

entity tends to have little explanatory payoff. Complex systems exist in the middle ground: they are identifiable systems that nevertheless resist decomposition.[83]

1.3.3 Emergence

The thesis of this section is that emergence is the core ontological commitment of complexity science. The work-constraint cycles, macro order, integration, differentiation, and metastability all result from the presence of emergent properties.

Emergent properties are properties of the whole system that are not contained by any of the parts of the system, and which are also not straightforwardly derivable from the properties of the parts.[84] Emergent processes result from spontaneous global self-organization that constrain certain parts of the system, but do not belong to any of the parts so constrained.[85] Explanation of complex systems must cite emergent properties because of the centrality of the influence from the system as a whole to the behaviour of its parts. If complex systems, as wholes, did not have some sort of influence on their components, the explanatory strategy of decomposition would suffice.

It is constitutive of a complex system that it has emergent properties that exert determinative influence on the parts of the system. Thompson (2007) explicitly connects emergence, downward determinative influence, and complex systems:

> [a] network, N, of interrelated components exhibits an emergent process, E, with emergent properties P, if and only if:
>
> (E1) E is a global process that instantiates P, and arises from the coupling of N's components and the nonlinear dynamics, D, of their local interactions.

---

[83] For an extended discussion of this point, see Strevens (2017).

[84] Thompson (2007, 60) summarizes the matter thus: "In complex systems theory, an emergent process is one that results from collective self-organization. An emergent process belongs to an ensemble or network of elements, arises spontaneously or self-organizes from the locally defined and globally constrained or controlled interactions of those elements, and does not belong to any single element."

[85] Any general characterization of emergent properties is tendentious. For overviews of emergence, see O'Connor & Wong (2015), Huneman (2016), and Humphreys (2017). For an overview of emergence as used here, see Thompson (2007, p. 60). For a clear skeptical view, see Kim (1999, 2006).

> (E2) E and P have a global-to-local ("downward") determinative influence on the dynamics D of the components of N.
>
> And possibly,
>
> (E3) E and P are not exhaustively determined by the intrinsic properties of the components of N, that is, they exhibit "relational holism". (Thompson 2007, 418)

The precise nature of the relation of downward determination is contentious. A natural reading of downward determination is that it is a causal relation between the whole system and its parts. On this reading, systems with emergent properties have "downward causation", and complexity science studies systems with various sorts of downward causation.

There is an influential objection to emergence as a thesis about causality, articulated by Kim (1999, 2006). Roughly, the worry is that we intuitively think of objects as exercising their causal powers only if they *already* have them—the "causal power actuality principle". In the case of E2, if the emergent process E gets its causal powers from its constituents N and their dynamics D, but at the same time bestows causal power on its constituents N, there is a violation of causal power actuality. Either the causal powers of E are superfluous, or it gets them from somewhere other than D and N. If E gets its causal powers from anywhere other than D and N, this violates the causal closure of the physical, making emergentism anti-naturalistic.

There are four considerations against this skeptical line of thought.

The first response to this objection is that the causal power actuality principle begs the question against emergentism. To gloss the matter, the objection requires an additional assumption that the lower level of organization is the more real one. The causal closure of the physical is not violated if entities at different levels of organization have different actually, already existing causal powers. Without the assumption that lower levels are more real, the phenomenon of downward determinative influence is better understood as a situation where two entities exert their respective actually existing causal powers. In some systems the lower-level causal powers are most relevant for explaining their activity, and in others the higher-level causal powers are

most relevant.[86] The latter sorts of systems are best approached from the perspective of complexity science, particularly for systems at grades II and III of downward determination.[87] Systems at grade I of downward determination exist in a gray area where the explanatory payoff of whole-system dynamics is less apparent.

Second, for present purposes, only diachronic claims about emergence need to be entertained. Instead of the parts and wholes exercising synchronic causal powers both "upward" and "downward", parts and wholes' determining powers depend on dynamics at somewhat different spatiotemporal scales. The powers of the whole to determine the parts generate boundary conditions in the form of constraints on the parts' intrinsic dynamics, and the aggregate activity of parts generates constraints on the determinative powers of wholes. In this situation, there is no inconsistency in holding that determinative influence proceeds both "upward" and "downward". The “downward” aspect of this determinative influence can be understood as a system as a whole biasing the dynamics of its parts.[88]

Third, in this sense of emergence, the notion of level of analysis is relative to various sorts of spatial, temporal, and other criteria. Emergence in this sense outstrips mereological analysis of parts and wholes. Scientific practice gives many, sometimes incompatible, criteria for what constitutes a level of analysis.[89] On this view, there are local hierarchies of levels, but no ultimately preferred level. This view contrasts with views of levels of analysis that bottom out with fundamental physics.

Fourth, the objection from causal power actuality operates from a deep presupposition about causality. It presupposes that causal relations require transference of energy, which is a generalized version of the billiard-ball model of causality. There are alternative models of

---

[86] A similar line of thought, emphasizing the whole-organism and mechanical parts levels as autonomous explanatory schemes, is developed in Walsh (2012).

[87] This response depends on an ontology of levels of analysis. Levels of analysis and their reality are hotly debated. In such debates it is not particularly clear where ontology ends and methodology begins. The details of such ontology are beyond the scope of the present discussion. But see Wimsatt (1994) for the details of level-attribution. Also see Glennan (2017) for a non-foundationalist take on levels of analysis.

[88] See Walsh (2012) for a discussion of the way in which whole systems can bias their parts. See also Chapter 3 of this dissertation.

[89] See Wimsatt (1994) for discussion of epistemic criteria of levels of analysis. See also Glennan (2017).

causality, in particular models on which causal relations are relations of counterfactual dependence between events.[90]

Dependence theories of causality weaken the objection from causal power actuality. On a dependence view, the causal relation is a counterfactual relation between events. Such relations are not restricted to the transfer of conserved quantities between entities. As such, there is a sense in which whole-system properties can cause changes in the dynamics of their parts, because there are plausibly robust relations of counterfactual dependence between the whole-system property and the dynamical possibilities of some of the system's components. In short, dependence theories allow for robust causal relations across levels. The limitation to "upward" causal determination is an assumption born from transference theories of causality.

Taken together, the above lines of thought deflate the initial objection to emergent properties being a problem for a naturalistic account.

Emergent properties entail that wholes exert determinative influence in the form of constraints on the activities of lower-level constituents. Constraints can be restrictive or enabling. In systems with emergent properties the context that a system's lower-level constituent parts find themselves is indispensable for explaining their activity.[91]

Methodologically, emergent properties entail that there are global parameters that are indispensable for explaining the behaviour of the parts of a suitably complex system. Typically, systems with emergent properties are understood in terms of order parameters and control parameters. Order parameters are parameters that describe system-level properties. Control parameters are global parameters that make a difference to the dynamical regime the system finds itself in. In systems with emergent properties, order and control parameters are not compressible into information about the dynamics of the parts of the system.[92] That is, it is

---

[90] For a discussion of the main distinction between transference and dependence theories of causation, see Hall (2004). See also Woodward (2003).

[91] Juarrero (1999) and Moreno & Mossio (2015) take this approach to emergence.

[92] Such claims can be found in Bedau (2002). Similar claims can be found in Huneman (2016) and Humphreys (2017).

explanatorily indispensable that at least some order and control parameters take the form of whole-system dynamics.

So, complex systems are fundamentally characterized by the presence of emergent properties with powers of downward determination. Downward determination is a phenomenon where whole-system properties modify the dynamical possibilities of a system's parts. The precise nature of downward determinative influence can be categorized into three grades. At each level the extent and scope of downward determinative influence undergoes a qualitative change.

1.3.4 Three Grades of Downward Determination

This section synthesizes the ideas of constraint, complexity, and downward determination presented in the previous three sections. The thesis of this section is that there are three grades of downward determinative influence, corresponding to three grades of complex systems. Agency is realized only in systems with grade III downward determinative influence.

Grade I is characterized by the material realization of constraints. Recall that constraints are modifications of the dynamical possibilities of entities or processes. The material realizers of constraints are states of the system that are required for material or energetic transformation that are not used up by the transformation (see figure 1). Grade II is characterized by the closure of constraints. This means that the persistence of constraints depends on a closed network of mutually supporting constraints. Grade III is characterized by an adaptive closure of constraints, that is a closure of constraints that has the additional capacity to resist the tendency of the network of constraints to degrade and run down.[93]

Each grade transition deepens the scope and strength of downward determinative influence. The higher the grade, the more comprehensively the degrees of freedom of the respective parts can be modified. Grade I complex systems are aggregates or ensembles of constraints. Grade II systems are the first instance of genuine whole-system downward determination. Grade III systems

[93] The key concept of adaptivity is developed in depth in Di Paolo (2005), Thompson (2007), Barandiaran & Moreno (2008), and Skewes & Hooker (2009).

represent the first instances of more complex self-driven activity.[94] Grade III downward determinative influence is the theoretical contribution of complexity science to the project of naturalized agency.

Grade I downward determinative influence consists of the presence of work-constraint cycles in parts of the system. Work here means a controlled release of energy. Constraint here means the energy so released modifies the dynamics of some other entities. Typically, this means the flow of matter and energy is harnessed to maintain ordered structures, typically at the macro scale. Such downward determination is present in all complex systems. It is responsible for the very constitution of complex systems as distinguishable unities.

For example, the persistent ordered structures present in weather phenomena exert grade I downward determinative influence (see figure 1). In all such situations there is a gross phenomenon of persistent macroscopic order, for example a convection cell. Energy within such systems is channeled in a constrained manner so as to produce order. The effect of the constraint generated is to limit the degrees of freedom of some of the constituent parts of the macroscopic phenomenon. Whereas air molecules are theoretically free to move in all directions, the convection cell biases their motion in the direction of the roll of the convection cell.[95]

Grade I downward determination gives rise to the various aspects of complex systems such as their differentiation, integration, and metastability. The extent of the depth, richness, and causal power of the constraints operating at grade I varies depending on the interactive domain of the system. Typically, constraints in the physical domain are the simplest, and domains such as population dynamics and social interaction presuppose various and convoluted constraints. For example, physical constraints can be understood as attractors in the phase space of some system's dynamics. Chemical constraints are typically understood in terms of catalyst-like phenomena. Biological and more complex constraints typically show up as thickly nested feedback mechanisms that resist decomposition.

---

[94] Grade III downward determination is necessary for adaptive behaviour, the first instance of "outward" determination from system to world, which is essential for our pre-theoretical idea of agency.

[95] See Thompson (2007) for an extended discussion of the distinction between grade I and grade II downward determinative influence. Thompson terms these two grades "self-organization" and "autonomy", respectively.

Persistent ordered structures are the simplest consequence of downward determinative influence. But iterations of this basic motif give rise to systems that seem to pursue stable end-states in increasingly sophisticated ways. Whereas in the simplest case a constraint works to channel matter and energy, in more sophisticated cases there are chains of such constraints. Constrains can be organized into positive (amplifying) or negative (steady-state attaining) cycles.

The two paradigmatic cases of systems at the more sophisticated end of grade I are (i) systems whose activities seem to flexibly pursue goal states across a range of activities, and (ii) systems that maintain a stable internal milieu despite the vagaries of external circumstances. For example, many complex systems display taxis-like behaviour, robustly and persistently attaining some state. And many systems harness energy to resist disintegrative influences of all sorts. The less sophisticated end of grade I is exemplified by convection-like phenomena. Less sophisticated grade I systems seem somewhat mechanical or limited in their capacities. The more sophisticated grade I systems seem more fluid and adaptive in the pursuit of their target states.[96]

Grade II downward determinative influence consists of the achievement of a closure of constraints.[97] Achieving this form of closure makes is possible to attribute aspects of the system's organization to the system itself in a global way. This idea amounts to the phenomenon of autopoiesis, in which the system continuously constitutes itself.[98] At grade II, the attainment of a stable end state allowed by grade I is developed into the stable attainment of the organization of the system itself.

A closure of constraints means that the various constraints operating within a system form a network where each constraint depends on at least one other constraint in the network for its persistence (see figure 2). Roughly, the constraints in operation form a knot-like structure of mutual dependence. Chapter 2 develops the formal details of this notion. With respect to material

---

[96] The cybernetic research program could be seen as dealing with systems that have such sophisticated grade I downward determinative influence. See Cannon (1929) for a discussion of homeostatic systems, Wiener (1948) for a discussion of feedback systems more generally, Bechtel (2007) for a general discussion of cybernetic principles in systems biology, and Bechtel & Abrahamsen (2011) for a discussion of cyclic organizations of constraints in complex systems.

[97] The idea of closure of constraints is developed formally in Montevil & Mossio (2015). See also Chapter 2 of this dissertation.

[98] See Thompson (2007).

realization, closure of constraints manifests as a diachronic dependence of different constrained flows of matter and energy on each other. The key difference between tightly regulated feedback loops at grade I and grade II is that grade II licenses attribution of downward determination to the whole system, not just to local parts that act as constraints.[99] Grade II builds on the notion of macro order and adds the physical grounding of actively maintenance of macro order by the system as a whole.

Closure of constraints is a second-order phenomenon from the perspective of constraints. Whereas individual constraints channel specific flows of matter and energy, the closure of constraints is a constraint on some set of constraints operating within the system. A set of constraints that is constrained in this way defines an autonomous unity in some particular domain.

The paradigmatic example of grade II downward determination is an autopoietic system.[100] An autopoietic system is one where:

> ...the constituent processes (i) recursively depend on each other for their generation and their realization as a network, (ii) constitute the system as a unity in whatever domain they exist, and (iii) determine a domain of possible interactions with the environment (Thompson 2007, 44).

The free-living cell, considered as abstracted away from its environment, is the simplest example of an autopoietic system. The free-living cell achieves autonomy in the biochemical realm. In this domain, the cell is a persistent structure of constraints that takes in matter and energy and expels degraded material and energy as a relatively unified whole.[101]

---

[99] See Moreno & Mossio (2015).

[100] There is considerable terminological slippage in the definition of autopoiesis. Slightly different definitions are offered in Maturana & Varela (1980), Varela et al (1991), and Thompson (2007). The definition cited does enough justice to the nuances of autopoiesis for present purposes. The emphasis on the closure of constraints here departs from the more high-level formal definitions offered by Maturana and Varela, but this achieves a clear link to physical principles that the abstract definition tends to downplay.

[101] A more formalized version of the autopoietic idea can be found in Rosen's (1991) Metabolism-Repair (M,R) System theory. M,R Systems introduce a formal analogue of the Aristotelian distinction between efficient and formal causes. M,R systems are systems where the effective causation that leads to the production of some component is only possible with the intervention of some further component that acts so as to make the conditions

Systems that realize a closure of constraints reproduce the basic logic of living systems in some domain. Whereas grade I systems stably attain particular states, systems at grade II stably attain a particular closure of constraints. This is what it is for a system to be autopoietic, i.e., to reliably constitute itself.

Grade III downward determinative influence is realized by adding adaptivity to grade II downward determinative influence. Adaptivity is the capacity to anticipate whether the system is approaching or receding from the conditions that would allow it to maintain its characteristic closure of constraints and to take appropriate measures to keep itself out of the regions of its state space that tend towards its disintegration. Adaptivity is lost, for example, when a system enters a dynamical regime where it is on an irreversible course towards determination. For a time, the system may maintain its capacity for grade II downward determination, but disintegration is inevitable. Grade III downward determinative influence constrains the system's resources towards maintaining its organization in an anticipative way.

Systems with grade III downward determinative influence are adaptive autopoietic (AA) systems. Whereas grade II is achieved with the static organizational criterion of closure of constraints, grade III also achieves a dynamic criterion of adaptivity in some domain of the system in question.[102]

AA systems are grade II systems that have an additional capacity to regulate their boundary conditions.[103] These are the conditions that influence the flow of matter and energy through the system. The regulation of boundary conditions is the downward determination in AA systems of

---

fortuitous for the production of the original component. Crucially, a system can produce the component that acts as the enabling formal cause through some other process in the system. The formalism allows one to qualitatively capture the relative depth and distance from thermodynamic equilibrium of an M,R system based on the number of layers in this structure (Hofmeyr 2007).

[102] Adaptivity as a dynamical notion addresses concerns with the overly abstract nature of the autopoietic criterion. Such concerns can be found in Boden (1999, 2000), and Bechtel (2007). In a way, adaptivity brings the autopoietic criterion back to concreteness.

[103] See Di Paolo (2005), Barandiaran & Moreno (2008), Barandiaran et al (2009), Barandiaran & Egbert (2014), Moreno & Mossio (2015).

the parameters that govern their coupling to their environment. This idea is formalized as a set of differential equations:

(1) $dS/dt=F_Q(S, E)$

(2) $dE/dt=G_Q(S, E)$

(3) $\Delta p=H_T(S)$

In this system of equations, S describes the state of the system, E describes the state of the environment. Equations (1) and (2) jointly describe a system-environment coupling. In (1), the system S evolves over time as a function of some function F which takes as its inputs states of the system and the environment E. Similarly, in (2), the environment E evolves as another function G of states of S and E. The parameter Q represents the set of conditions and constraints on the coupling, and the parameter p represents a subset of the constraints Q. Equation (3) describes the fact that some subset of the parameters governing the coupling depends on the state of the agent system over some period of time. Adaptivity is the system's capacity to affect changes in p and thereby Q as a function H of its own states at least some of the time. In essence, (1)-(3) constitute the necessary and sufficient conditions for a system to exert grade III downward determinative influence.

In simpler terms, grade III complex systems marshal their resources to select which interactions with the environment to undergo. This is why it makes sense to say that the agent acts: it modulates its coupling with respect to the environment in accordance with its own internal processes. In this way, it realizes a certain degree of skillful coping with its environment.

Natural agency requires grade III downward determination. The mere presence of constraints, i.e. grade I downward determinative influence, is not enough. Grade I downward determinative influence is not sufficient to establish autonomy in a domain, which means that analyzing a system's dynamics in terms of inputs and general conditions could be a sound empirical strategy. Grade II downward determinative influence suffices for autonomy in an abstract sense, but is not enough to establish a system's active involvement with its environment. Such a system is distinguishable as a unity but is not necessarily active in its self-constitution. Grade III downward determinative influence adds the element of whole-system influence on its own

boundary conditions. Such grade III downward determination the necessary set of complex system properties out of which agents are constituted.

In sum, the systems studied by complexity science are those that offer up spontaneous order as a result of fundamental thermodynamic principles. Spontaneous order manifests as a wide variety of system features and capacities. The most reliable indicator of complexity is the presence of clear differentiation but also clear integration. Complex systems instantiate emergent properties that result in downward determination of parts by wholes. There are three distinct grades of downward determination. In grade I the complex system realizes constraints. In grade II the complex system realizes a closure of constraints and thereby can regenerate its organizational form. In grade III complex systems there is the capacity to resist the disintegration of their organizational form proactively and thereby engage in goal-directed behaviour.

Systems at grade III of downward determination, i.e. adaptive autopoietic systems, are a necessary component of a naturalistic theory of agency.

1.4 Heterodoxy 4: Situated Darwinism

Put simply, situated Darwinism is the view that organisms enact evolution. The evolutionary process is driven by the responses of organisms, as agents, to the affordances in their environments.[104] This stands in stark contrast to the orthodox view of evolution, the Modern Synthesis, on which evolution is the imposition of selective constraints by the environment on the differential survival of replicators.

The thesis of this section is that situated Darwinism illustrates the applicability of the framework constituted by goal-centered agency, situatedness, and complexity science.

Many recent phenomena in development and evolutionary biology make sense under the situated Darwinist framework whereas the Modern Synthesis struggles to accommodate them. As such, taking an empirical bet on situated Darwinism gives a broad empirical grounding to (1) the view of agency as goal-directedness, (2) situated cognition, and (3) complexity science. In other words, the situated Darwinism makes the framework sketched in 1.1-1.3 plausible, quite

[104] This view is presented in Thompson (2007), Walsh (2012a, 2015).

independent of its value for explaining cognition. This discussion is relevant to naturalized agency because Situated Darwinism suggests that agency cannot be dissolved by appeal to evolution; if anything, evolution requires an appeal to agency at its core.

The plan for the section is as follows. 1.4.1 will sketch the Modern Synthesis. 1.4.2 sketches situated Darwinism, connecting it also to the earlier discussion of situated cognition.

1.4.1 Orthodoxy: The Modern Synthesis

Evolutionary theory aims to explain the fit and diversity of organic form. To do so, it must account for four processes: (1) inheritance, (2) development, (3) adaptation, and (4) the generation of novelty. Inheritance is the phenomenon of the existence of patterns of resemblance and difference between parents and offspring. Development is the process whereby organisms grow in complexity over the course of their life cycles. Adaptation is the phenomenon of populations coming to be comprised of individuals that are well suited to the conditions of life. The generation of novelty is the source of variation within populations on which differential survival and reproduction depends.

The orthodox view of evolution is the Modern Synthesis.[105] The Modern Synthesis is a theory of adaptive population change brought about by the differential survival of replicators.[106] Evolution by natural selection is the change in the relative frequencies of replicators within populations. The Modern Synthesis gives theoretical pride of place to populations and replicators. Organisms occupy a middle ground between two more theoretically important levels of analysis: the level of replicators and the level of populations.

The Modern Synthesis way of conceptualizing evolution gives corresponding accounts of the four component processes of evolution by natural selection. Inheritance is the intergenerational transmission of replicators from organism to organism. Development is the unfolding of a

---

[105] The description of the Modern Synthesis in this section closely follows Walsh (2015).

[106] The term "replicator" is often used interchangeably with the term "gene". As such, the modern synthesis offers a gene's-eye-view of evolution. Here I use the term replicator because it is more neutral with respect to substrate and implementation. All that is required to be a replicator is that a process is reliably reproducible across generations, and possess the right kind of longevity. Replicators refer to sub-organismic processes exclusively. Organisms replicate, but are not replicators in this sense.

program realized by a particular subset of replicators. Adaptation is change in the relative frequencies of replicator types within a population. The generation of novelties is brought about by chance mutation of replicators. In sum, replicators are the units of inheritance, the units of adaptation, and the units of control for development (Walsh 2015, Uller & Helentera 2019).

As a consequence of this conceptualization of the component processes of evolution, the Modern Synthesis has a fractionated view of evolution. Fractionation means that the four component processes are quasi-independent of each other (Walsh 2006, 2015). Some of these fractionated evolutionary processes occur are the level of replicators (inheritance, development, the generation of novelty) and some occur at the level of populations (adaptation). In sum:

> [T]he evolutionary biology of the late twentieth century was a suborganismal phenomenon comprising quasi-independent processes of inheritance, development, the origin of form and adaptive population change. Each process had its own theory; and in each theory the canonical unit of biological organisation was the replicator or gene. (Walsh 2015, 86)

On this view of evolution, organisms are mere aggregates of replicators, and mere a components of populations.

On the Modern Synthesis view, the insights of complexity science are relevant only to the process of development, which is peripheral to the explanatory apparatus. As such, complexity science is relevant only for the proximate explanations of biological forms, not the ultimate, evolutionary explanations.[107]

1.4.2 Situated Darwinism

Situated Darwinism is a thesis about the metaphysics of evolution. On this view, organisms respond to their environment as agents. That is, they respond to the affordances of the environment. Evolution by natural selection is a predictable and inevitable higher-order effect of the activities of all the organisms within a population. Just as the temperature of a gas is a

[107] The distinction between ultimate and proximate explanation is from Mayr (1993). See Laland et al (2009, 2012) for a critical take on the distinction.

higher-order effect of the mean kinetic energy of the molecules making up the gas, so natural selection is a higher-order effect of the life processes of organisms as agents going about their lives. In both cases, the former phenomenon is a higher-order effect of the latter (Walsh 2012, 2015).

This claim about the metaphysics of evolution entails the de-fractionation of the processes of inheritance, development, adaptation and novelty-generation that have been fractionated by the Modern Synthesis.[108] This is because, on a situated Darwinist understanding, the four component processes of evolution arise from the activities of organisms as they go about the struggle for life.

Inheritance is not simply the transmission of replicators from generation to generation. It is a broader phenomenon which secures the patterns of difference and resemblance across generations. Inheritance relies at least on epigenetic and environmental systems in addition to genomic systems.[109] This view presupposes that an organism's has adaptive autopoietic capacities. In inheritance, the organism's goal is to reliably reproduce a pattern of difference and resemblance to its offspring. And organisms are constantly marshaling their capacities to both construct and maintain the inheritance pattern.

Development is not merely the unfolding of a developmental program encoded in some portions of an organism's genome.[110] Development is more properly understood as a process of complexification orchestrated by organisms as they respond in sophisticated ways to the affordances in their environments. Genetic resources are one amongst a number of factors in securing the stability of form within an organism's lifetime. Again, the capacity of an organism

---

[108] Situated Darwinism is compatible with a cluster of research programs in evolutionary and developmental biology that aim for an Extended Evolutionary Synthesis (EES). The EES is most prominently motivated by the need to integrate developmental biology into the modern synthesis (West-Eberhard 2003, Walsh 2015). Other concerns relate to the Modern Synthesis treatment of ecological relations (Odling-Smee, Laland & Feldman 2003, Laland et al (2012, 2015), its oversimplified conception of inheritance (West-Eberhard 2003, Jablonka & Lamb 2008), and its parochial definition of the source of evolutionary novelty. Situated Darwinism provides a plausible general framework for integrating these somewhat disparate critiques of the Modern Synthesis.

[109] See West-Eberhard (2003), Odling-Smee et al (2003), Thompson (2007), Jablonka & Lamb (2008, 2010), Walsh (2015).

[110] See Carroll (2008), Pigliucci (2009), Pigliucci & Muller (2010).

to perform sophisticated orchestration of its development presupposes adaptive autopoietic capacities. In this instance, the goal is the creation and maintenance of complex structures.

Adaptation is not simply a population-level process whereby populations come to respond to environmental problems. The situated Darwinist understanding of adaptation is of a process based on the responses of organisms to their affordances. The population-level change is a consequence of sophisticated, plastic, adaptive organism-level activity.[111] This understanding of adaptation presupposes that organism and environment are co-constituted, which relies on both adaptive autopoiesis and the situatedness of organisms in their environment.

Finally, the generation of novelty is not just based on random processes of mutation. Organisms have, in virtue of adaptive autopoiesis, a surprisingly robust capacity to modulate the effects of mutation on themselves. This capacity has been conformed both empirically and theoretically.[112]

So, situated Darwinism de-fractionates inheritance, development, adaptation, and the generation of novelty by citing organisms' capacities as adaptive autopoietic systems, that is as systems with capacities for grade III downward determination.

Situated Darwinism integrates critiques of the Modern Synthesis in evolutionary biology. The theoretical and empirical contest between the two frameworks is far from settled, but it is plausible to bet on the empirical staying power of this situate Darwinism.[113]

Regardless of the long-term theoretical trend in evolutionary biology, situated Darwinism illustrates the integrative potential of grade III downward determination. The adaptive autopoiesis of organisms is a prerequisite for the evolutionary process on the situated Darwinist view. Situated Darwinism also illustrates well the processes behind the relations of intimate situatedness in H2. Situated cognition is simply the situated Darwinist framework applied to one kind of sophisticated capacity of certain organisms.

---

[111] See Lewontin (2001), Odling-Smee (2003).

[112] See Gerhart & Kirschner (2007), Wagner (2011), Walsh (2015).

[113] See Laland et al (2014) for an overview of the empirical side of the debate between the Modern Synthesis and views affiliated with situated Darwinism.

1.5 Conclusion

This chapter has surveyed four heterodox frameworks: (H1) the framework of agency as goal-directedness, (H2) situated cognition, (H3) complexity science, and (H4) situated Darwinism.

H1-H3 form a powerfully mutually supportive set of views. Complexity science vindicates the naturalness of adaptive autopoietic systems. Adaptive autopoietic systems vindicate the guiding intuitions of both the goal-directedness view of agency and situated cognition.

Adaptive autopoietic systems have their own goals and influence the way they are coupled to their worlds. This is the naturalistic starting point for thinking about how agents are goal-directed. The activities of adaptive autopoietic systems occupy a middle ground between systems determined by external input and systems that are responsive to rational norms. Adaptive autopoietic systems are not reason-responsive, but they respond to norms in the wider sense. They anticipate the world and their own internal states, which manifests as the capacity to pursue goals in their environments robustly and plastically.

Adaptive autopoietic systems are also intimately situated. They show, in basic outline, the relational profile of any situated system, and so offer the natural grounding of situated cognition. Of course, cognitive systems are far more sophisticated than the simplest adaptive autopoietic systems, but that may be a matter of implementation details. The essential link is that adaptive autopoietic systems have internal norms and respond to affordances, and the marshaling of an adaptive autopoietic system's capacities is a general instance of the deep structure of situated cognition. Adaptive autopoietic systems cope with their worlds as affordances for action first and foremost.

Situated Darwinism illustrates the application of this set of views in evolutionary biology. It serves as a platform from which to develop an account of cognition compatible with the situatedness of organisms. The agency of organisms suggests that inheritance, development, adaptation, and the response to novelty are best approached as subsets of organismic capacities, as goal-directed sub-systems of the whole organism. And given continuity between life and mind, situated Darwinism suggests that highly sophisticated forms of agency that respond to norms other than mere survival and reproduction can evolve from the sorts of simple agents characterized by AA systems.

The four heterodoxies constitute a powerfully mutually supportive set of views. They are internally coherent and increase each other's plausibility.

Thus, a suitable naturalistic account of agency requires all four heterodoxies. Given that a naturalistic account of agency is desirable, there is *prima facie* reason for adopting revisionary views of agency, cognition, explanatory practice, and evolutionary explanation. The ecosystem of ideas established by these four heterodoxies stand and fall together.

Together, H1-H4 open the conceptual space in which pursuing the assimilation of cognitive agency to natural agency looks possible with the explanatory tools of current science. But even if H1-H4 are right, further lacunae must be filled in this account.

First, the complexity science account of natural agency must be given. Chapter 2 does this, pointing out as it does possibilities and pitfalls in connecting agency as goal-directedness with the best existing complexity science accounts of natural agency. Chapter 3 develops the conceptual tools that will be necessary for making goal-directedness a naturalistically acceptable explanatory framework. Finally, Chapter 4 extends the account of natural agency to cognitive agency.

2 The Organizational Account of Agency and its Limits

This chapter examines the most prominent existing account about the natural realization of agency within the framework of complexity science. As discussed in the previous chapter, agency requires complex systems that realize adaptive autopoiesis. Such systems must be situated in the appropriate environments to enable their capacities.

The most developed account of the natural realization of agency is what I will call the Organizational Account of Agency (OAA). On this account, an agent is:

(OAA1) A system that has:

(1.1) A *core* that realizes

(1.1.1) a *closure of constraints*,

(1.1.2) *constitutive precariousness* such that it has being-by-doing

(1.2) A *periphery* that realizes

(1.2.1) at least one *thermodynamic gradient*

(1.2.2) a suitable flow of *free energy* exploitable by the core,

(1.2.3) the capacity to be a sink for the *entropy* generated as the inevitable byproduct of the maintenance of the core

(OAA2) Where core and periphery jointly realize:

(2.1) *negentropic* character,

(2.2) the comparatively greater *organizational depth* of the core compared to the periphery

(2.3) the *adaptive autopoiesis* of the core.

This view is an uncontroversial amalgam of the most prominent accounts of natural agency (Barandiaran et al 2009, Moreno & Mossio 2015, Montevil & Mossio 2015, Di Paolo et al 2017).[114] The Organizational Account of Agency integrates the insights of complexity science,

[114] The Organizational Account of Agency integrates several theoretical traditions. The attempts to account for the key features of agential systems form a complex and branching historical lineage. This intellectual background traces its roots back to Kant (2000/1793) who first proposed that organisms are self-organizing. One strand has attempted to characterize the key features of agency by appeal to various aspects of self-organization found in nature

system theory, and far-from-equilibrium thermodynamics into an account of the causal, organizational, and system-theoretic patterns that pick out natural agents.

The thesis of this chapter is that doing justice to agents' natural normativity and goal-directedness requires a supplementing the OAA. The OAA, properly construed, is not a theory of agency *per se*. It is, however, essential for naturalizing agency because it is an account of how physical systems can be organized so as to realize agency. This view is the Ecological Account of Agency (EAA), on which agents are fundamentally relational patterns between a system's repertoire and its affordances. An ecological account requires an explanatory framework that is complementary to the physical, causal, and systems-theoretic explanatory framework that is the basis of the OAA. In contrast to the OAA view that an agent is a causal, thermodynamic, and constraint-based pattern, the ecological framework takes agency to be a pattern of an agent's response to its affordances.[115] The OAA faces explanatory limitations regarding sophisticated normativity and goal-directedness, and the conceptual tools for addressing that gap require the EAA.

The argument strategy of the chapter is as follows. Section 2.1 sketches the theoretical background of the Organizational Account of Agency, establishing that it is the best account for the locus of real natural agency. Section 2.2 develops the key features of the agential core, particularly the key ideas of closure of constraints and constitutive precariousness. Section 2.3 develops the key features of the agential periphery. Section 2.4 develops the key properties that are realized jointly by the core and periphery. Section 2.5 develops the OAA of the normativity and goal-directedness of agents. Section 2.6 argues that the OAA fails to allow for autonomous goal sets—capacities that are autonomous from the agent's basic goal of persistence. Section 2.7

(Bernard 1865, 1878; Cannon 1929; Wiener 1948; Ashby 1956; Von Bertalanffy 1969; Varela et al 1974; Maturana & Varela 1980, Rosen 1991). Another important strand has been the organicist tradition in biology (Oyama 1985; Sarkar & Gilbert 2000; Ruiz-Mirazo et al 2000; Oyama et al 2001; Ruiz-Mirazo & Moreno 2004, 2012; Thompson 2007; Nicholson 2013, 2014, 2018). Another strand has given more explicitly philosophical critique of conceptions of self-organization (Somerhoff 1950; Jonas 1953, 1966, 1968; Ayala 1970; Boden 2000; Weber & Varela 2002; Di Paolo 2005; Thompson 2007; Barandiaran et al 2009; Moreno & Mossio 2015; Di Paolo et al 2017).

[115] The ecological account of agency is developed in Fulda (2016, 2017) and Walsh (2015, 2018). Chapter 3 of this dissertation discusses the view at length.

offers a way out of the tension at the price of price of taking on a more fully ecological explanatory framework. Section 2.8 concludes with a sketch of the way in which the OAA and the Ecological Account of Agency could be made complementary.

2.1 The Intellectual Background of the Organizational Account of Agency

The thesis of this section is that the OAA is the most developed phase of a long intellectual tradition that aims to naturalistically explain agency by appealing to self-organization.[116] It is an account of systems that demand explanation in terms of their own goals and purposes.

An initial objection must be addressed. There is a deep intuition that agency and agents are entities *constituted* by explanatory practice.[117] If this is right, there is no agency in nature, instead there is an *explanatory stance* that one can take towards very complex natural patterns such as organisms, animals, or other human beings in order to predict and understand them. On this instrumental view of agency, such systems are so complex that we must attribute a web of concepts such as wants, feelings, movements, goals, and so on to such systems for the sake of gaining any epistemic purchase on these systems at all.

Dennett's (1989) intentional stance is an example of stance-taking practice with regard to cognition. Taking the intentional stance involves projecting a conceptual web onto a system of interest in order to make its behaviour intelligible.

> …first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and

---

[116] The traditions and authors discussed in this section are quite heterogeneous. I use agency to refer to what has variously been called natural purposiveness, organismality, life, self-organization, self-movement, the maintenance of an internal milieu, and many other terms besides. In this section I use naturalism in a broad sense because the demands of naturalistic explanation have changed over the centuries.

[117] In cognitive science, Dennett (1989) articulates the intuition well. See Jacob (2019) for context. In biology, the intuition that denies real status to agents is discussed by Walsh (2008, 2015).

> desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do. (Dennett 1989, 18)

Generalizing this view, agency in general might be seen as the projection of an *animistic conceptual web*.[118] This involves the projection of concepts like *wanting*, *caring*, *feeling*, *expecting*, and *moving intentionally* to describe some complex entities. As with the intentional stance in the realm of cognition, one might hold that agents and agency arise only from the deployment of the animistic stance. Taking an explanatory stance does not entail commitment to a natural pattern characteristic of cognitive agency. And more strongly, the attempt to characterize agency outside explanatory practice is a non-starter.

Such "stancy" intuitions are orthogonal to the present concerns. While instrumentalism about natural purpose may serve certain explanatory functions, the tradition of which the OAA is a part aims at providing independent metaphysical warrant to our attributions of natural purpose. Some attributions of natural purpose might align with the OAA criteria. Others may not. For the cases where the OAA criteria and instrumental criteria do not align, the OAA criteria are a reason to hold that the instrumental view is overextended.[119]

The intellectual background of the OAA has offered a battery of criteria by which one can understand systems as intrinsically purposive.

The attempt to characterize the distinctive naturalistic profile of agency in modern philosophy goes back to Kant's *Critique of Judgment* (Kant 2000, Walsh 2006, Thompson 2007, Mossio & Bich 2017). Kant pointed out that living things had a peculiar circularity to their organization in two senses: (1) their parts were necessary to the whole, and the whole was necessary for the parts, and (2) the development of organisms depended on previously existing organisms. For Kant, the natural purposiveness inherent in such self-organization of living things could not be

---

[118] This term is due to Steward (2012).

[119] Jonas (1966) offers a discussion of instrumentalism's challenge and its limitations to any account of natural purpose. While a degree of anthropomorphism is inescapable in all attributions of agency, he argues that this problem is not as deadly as is commonly assumed. It is indisputable that *some* of what is going on in agent-based explanations is projection, or anthropomorphism, but organizational criteria deflate the fatality of this objection.

accounted for by the pervasive mechanistic explanatory framework of his time, generating an antinomy of teleological judgment (Walsh 2006):

> The first maxim of the power of judgment is the thesis: All generation of material things and their forms must be judged as possible in accordance with merely mechanical laws.
>
> The second maxim is the antithesis: Some products of material nature cannot be judged as possible according to merely mechanical laws (judging them requires an entirely different law of causality, namely that of final causes). (Kant 2000, 387)

Ultimately, Kant stepped back from endorsing the conceptual framework of natural purposiveness as an explanation of how the world works, relegating it to a regulative principle of judgment. Nevertheless, the problem he posed generated a long tradition of attempting to find acceptable naturalistic grounding for natural purposiveness, thereby addressing the antinomy of teleological judgment in a way that carries ontological commitment (Walsh 2006, Thompson 2007).

Bernard (1865, 1878) introduced the idea that agents preserve an internal milieu, a contingent and local set of properties that serve as the target for biological self-organization. Agents maintain a stable internal milieu in order to make conditions suitable for their existence.

The idea of maintaining the internal milieu as a target for agents is thoroughly uncontroversial. Cannon (1929) developed this idea into the notion of homeostasis, the capacity of compensating for perturbations through coordinated action. Homeostatic capabilities are unquestionably essential for biological organization of all sorts. The idea of homeostatic capability was theorized more abstractly by the cybernetic tradition in the form of feedback systems (Wiener 1948, Ashby 1956, Von Bertalanffy 1969). This idea is central to systems biology and has received considerable theoretical development (Kitano 2004, Whitacre 2012).

However, the maintenance of an internal milieu as the basis of biological self-organization is theoretically insufficient. Jonas (1953, 1966, 1968) argued that cybernetic analyses could account only for extrinsic purposiveness, the goal-directedness that is inherent to artefacts. This is because cybernetic accounts failed to account for the difference between the purposiveness of artifacts and the purposiveness of natural agents. The cybernetic account did not offer a

characterization of the constitution of the agent system, namely the way in which the components of the system depend on each other.

The tradition of second-order cybernetics took up this conceptual challenge. This constitutive dimension of biological self-organization has been theorized as organizational closure (Varela et al 1974, Maturana & Varela 1980). On this view, agents are constituted by a closure of processes that support each other, and the capacities to maintain a stable internal milieu are a function of this more basic constitutive aspect (Weber & Varela 2002). The criteria of what natural purpose consists of have been clarified into three aspects: (i) asymmetry, (ii) normativity, and (iii) individuality (Di Paolo 2005, Barandiaran et al 2009, Di Paolo et al 2017). The basic idea is that agents constitute themselves as individuals over and against their world, they hold themselves to self-generated norms of behaviour.

In addition, theoretical work has explicitly connected the organizational account of agency to physical processes. This strand of theorizing has introduced the capacities and requirements of far-from-equilibrium systems into the characterization of self-organization (Boden 2000, Kauffman 2000, Moreno & Mossio 2015, Mossio & Bich 2017). The needs for self-constitution and compensation arise from basic physical facts on this view.

As such, the tradition offers a thermodynamically grounded account of nuanced features of agents, and it grounds these features in the intrinsic properties of certain kinds of complex systems.

The OAA is a formalization of this theoretical tradition. The next three sections develop its key features. First, the agential core is a far-from-equilibrium system that is, in virtue of this, precarious, and achieves persistence against this precariousness by realizing a closure of constraints. Second, the agential periphery is a source of thermodynamic (material and energetic) flow sufficient to sustain the closure of constraints, as well as a sink for degraded matter and energy. Third, the core and periphery jointly realize the properties of negentropy and adaptive autopoiesis.

2.2 The Agential Core

Agency fundamentally requires that the agent be the source of its actions, not an entity that only undergoes changes. The agential core is where the action comes from. In order to produce actions, the core must realize closure: a distinctness from the processes that surround it. Closure cannot be absolute, however. It must in some way be balanced by openness to the agent's environment.

The thesis of this section is that the agential core realizes two related features (i) a closure of constraints, and (ii) constitutive precariousness. These two features jointly provide the best existing naturalistic basis for the qualitative features of closure and openness characteristic of agents of all sorts.

Closure is a property of a network of interdependent processes. In a network that instantiates closure, every component of the closed network depends on at least one other component of the network and is itself a necessary condition for the operation of at least one other component in the network. This definition does not preclude some processes being affected by processes that are not themselves conditioned by the processes within the network, nor does it preclude the processes that constitute the closed network affecting other processes outside the network.

Closure has been essential for defining the architecture of systems that are taken to genuinely act on their environments. Paradigmatically, living things are taken to be the basic case of entities that act. The autopoietic definition of life entails that living systems are operationally closed while being thermodynamically open.

> [Varela (1979, 55-60)] defines an autonomous system as a system that has organizational closure (later called operational closure). Here closure does not mean that the system is materially and energetically closed to the outside world (which of course is impossible). On the contrary, autonomous systems must be thermodynamically far-from-equilibrium systems, which incessantly exchange matter and energy with their surroundings. Organizational closure refers to the self-referential (circular and recursive) network of relations that defines the system as a unity, and operational closure to the reentrant and recurrent dynamics of such a system. (Thompson 2007, 44-45)

The closure of a system of processes is necessary, though not sufficient, for the system to actively establish itself as a distinct entity in a particular domain. It is the feature of biological systems that underlies their active persistence through the incessant flux of material and energetic transformation.

However, there is an important limitation of this way of thinking about closure. If closure is explicated by appealing to a network of processes, then the assignment of closure applies to any system where a causal loop can be identified (Moreno & Mossio 2015). For example, the hydrological cycle realizes a causal loop, and so by the definition above achieves causal closure. If closure applies only at the level of processes, autonomous dynamics are too prevalent in nature to do the explanatory work demanded by the agential core.

In order to clarify the notion of closure, one must distinguish two causal regimes: that of processes, and that of constraints (Juarrero 1999, Moreno & Mossio 2015, Montevil & Mossio 2015, Di Paolo et al 2017). This distinction allows for an account of self-organization without the need to commit to synchronic downward determination.[120] The basic idea is that processes and constraints mutually determine each other at different timescales.

A process ($A \rightarrow B$) is any series of events that result in transformation of some entity A into entity B at timescale τ. A constraint C is a local, contingent entity or process that modifies the possibilities inherent in another process $A \rightarrow B$ and exhibits symmetries with respect to timescale τ on which $A \rightarrow B$ operates.[121] This means that C can be treated as a conserved, unchanged entity at the particular timescale (see figure 1).

> In contrast to physical fundamental equations, constraints are local and contingent causes, exerted by specific structures or processes, which reduce the degrees of freedom of the system on which they act[.] As additional causes, they simplify (or change) the description of the system, contributing to providing an adequate explanation of its

[120] A problem whose seriousness has been forcefully defended by Kim (1999, 2006). See also Walsh (2012) for a discussion of ways in which to avoid the problem for naturally purposive systems.

[121] The term constraint should be understood in both a limiting and enabling sense. The topology of a car engine limits gas diffusion within the engine. But it also enables the concentration of fuel and gas that brings about internal combustion.

> behaviour, which might otherwise be under-determined or wrongly determined. (Moreno & Mossio 2015, 5)

A constraint is an entity that makes a difference to A→B without thereby being used up in the process. At the relevant timescale τ, the constraint can be treated as an invariant entity that modifies the possibilities inherent in A→B. For example, an enzyme is a constraint on a particular chemical transformation: it enables the chemical reaction to happen at biologically relevant timescales without itself being affected at the timescale of the reactions it catalyzes. Similarly, the semipermeable membrane boundary of a cell is a constraint that enables the metabolic network to function as it does and can be treated as (approximately) invariant from the timescale of the metabolic transformations.

Montevil & Mossio (2015) offer a more formal definition:

> Given a process A→B (A becomes B), C is a constraint on A→B, at a specific timescale τ, if and only if the following two conditions are fulfilled:
>
> (1) The situations A→B and A_C→B_C (i.e. A→B under the influence of C) are not, as far as B is concerned, symmetric at a time scale τ.
>
> (2) A temporal symmetry is associated with all aspects of $C_{A\rightarrow B}$[122] with respect to the process A_C→B_C, at time scale τ. (Montevil & Mossio 2015, 182)

In short, constraints are processes that, at a relevant timescale, harness or canalize energy and material flows while remaining invariant at the relevant timescale. A causal process under the influence of a constraint is capable of bringing about states that are unlikely for the process. The work-constraint cycles that are fundamental for all complex organization depend on the presence of constraints (Kauffman 2000, Moreno & Mossio 2015).

Closure of constraints is a property of a network of constraints such that each constraint depends on at least one other constraint in the network and also enables (generates) at least one other

---

[122] $C_{A\rightarrow B}$ is those aspects of C which play a role in the above asymmetry between A→B and A_C→B_C at timescale τ.

constraint in the network (see figure 2). The closure of constraints means that a network of constraints and processes as a whole realizes self-constraint.

More formally:

> A set of constraints *C* realises [closure] if, for each constraint C_i belonging to *C*:
>
> > (1) C_i depends directly on at least one other constraint belonging to C (C_i is dependent)
> >
> > (2) There is at least one other constraint C_j belonging to C which depends on C_i (C_i is generative). (Adapted from Montevil & Mossio 2015, 186)

The dependence of a constraint C on another constraint C* means that that, at a certain timescale, C is the product of a causal process A→C that is constrained by C*. Dependence can be fast or slow. Slow dependence is the situation where C* exists at a longer timescale and so from the specific timescale of A→C it can be treated as an unchanging entity. Fast dependence a is the situation where C* has faster intrinsic dynamics than A→C, and so its involvement in enabling the process that produces C entails having some dependable, recurring aspects that allow for effective constraint.

A constraint C being generative of another constraint C* means that C constrains the process A→C*.

The definition of self-organization in terms of closure of constraints makes self-organization non-trivial. The closure of constraints is the specification of a set of (locally, contingently) invariant process into a higher-order invariant (closure). Because constraints are defined as being invariant with respect to the local processes, in every case of constraint, there is not an exchange of conserved quantities between the process and the constraint (Montevil & Mossio 2015). A closure of constraints is thus not achieved through efficient causal cycles, which involve exchange of conserved quantities. In a real sense, closures of constraints stand over and above the causal processes they constrain, while being causal processes themselves from a different analytic perspective.

The closure of constraints is a higher-order structure of invariants that is indispensable for explaining the specific character, occurrence, and dynamics of the network of processes so constrained. The mutual dependence of constraints explains the way in which the agential core differentiates itself from all other processes. Namely, the core realizes a set of locally invariant structures for channeling thermodynamic flow. To the extent the agent persists, the closure of constraints persists.

The other essential feature of the agential core is constitutive precariousness. A system is constitutively precarious if achieving thermodynamic equilibrium implies the system's disintegration. In other words, without input the whole structure that realizes the closure of constraints would run down. Constitutive precariousness is a consequence of the agential core being a closure of constraints. Constitutive precariousness gives a gross behavioural characterization of an entity that exists as a closure of constraints.

All dissipative systems are constitutively precarious. Such structures are eddies in a thermodynamic flow that tends towards equilibrium.[123] Closures of constraints are precarious in a different way than for dissipative systems. Closure of constraints are a network of dissipative systems that keep each other running by channeling material and energetic flows.

The difference between the precariousness of dissipative systems and the constitutive precariousness of closures of constraints was noted by Jonas (1966) in his characterization of such systems as having needful freedom.

Needful freedom is a "dialectical" relation between the agential core and its world. A needfully free system is free in that it is a source of action, but it is needful because the capacity for such free action requires machinery which is energetically expensive to maintain. For Jonas, agential cores "transcend" the necessity of physics and chemistry, but at the cost of having to exploit such necessities to maintain themselves.

> Opposing in its internal rule the entropy of general causality, [life] is yet subject to it: free, yet under the whip of necessity; isolated, yet in indispensable contact; seeking

[123] Whereas dissipative structures are "eddies" in the thermodynamic flow, more static ordered structures, like crystals, are not constitutively precarious because they are near-equilibrium structures.

> contact, yet in danger of being destroyed by it, and threatened no less by its want: imperiled thus from both sides, by importunity and aloofness of the world, and balanced on the narrow ridge between the two; in its process, which must not cease, liable to interference; in the straining of its temporality always facing the imminent no-more: *thus does the living form carry on its separatist existence in matter.* (Jonas 1966, 5, my emphasis)

In the present context, emancipation from necessity can be understood in terms of the closure of constraints. The most relevant explanatory principles for a closure of constraints are the local, contingent, and invariant constraints and their interdependence. Closure of constraints are in a sense underdetermined by the overarching law-like generalizations that apply to physicochemical processes. At the same time, since constraints are energetically expensive to maintain flow of energy and matter through the system is required to maintain the whole system. Needful freedom is a property of the agential core that generates an inherent tension in the individuation conditions of agents.

In Jonas' terminology, the first glimmer of "needful freedom" occurs for metabolic systems, and the realm of needful freedom increases in a manner roughly corresponding to the major transitions in evolution (Jonas 1966). Each major transition intensifies the inherent trade-off between being a simple, efficient closure of constraints and being a complex, more precarious closure of constraints.

The kind of being that an agential core has is *being-by-doing*. This is in contrast to being-by-being—the criterion used for individuating objects of scientific theorizing like atoms or chemical compounds (Jonas 1968). Entities that have being-by-being have intrinsic criteria of individuation.[124] Entities characterized by the closure of constraints presuppose a constant dynamism; in a closure of constraints, failing to act on material and energetic flows entails the

[124] This is somewhat of a gloss. There are different ways of being-by-being. For some entities, such as atoms and chemical compounds, they exist due to universally invariant law-like processes. For some other entities that have being-by-being, like a rock in a field, their individuation criteria do not advert to law-like generalizations, but they are intrinsic to the entity. What makes *that* rock is the set of internal relations between its constituent particles, and this would be true regardless of context.

eventual disintegration of the system.[125] Systems that have being-by-being are tricky to characterize with the apparatus of intrinsic properties and externally imposed individuation conditions. This is because such systems individuate themselves quite independently of the conceptual frameworks we apply to them. There is an inherent residuum in such systems that decompositional analysis cannot explain.

In sum, action is always attributed to the agential core. The agential core is the source of action. This entity is always a constitutively precarious closure of constraints. This idea grounds the way in which an agential core stands against material and energetic flux while being nothing over and above it. The more qualitative characterization of the core, which does justice to both the essential closure and essential openness of the agential core is constitutive precariousness. This idea grounds the necessity of the active nature of the agential core—its being-by-doing.

2.3 The Agential Periphery

The agential core is always situated. The thesis of this section is that given its constitutive precariousness, the agential core requires a periphery with at least one thermodynamic gradient, which acts as a source of free energy for the self-constructive and self-maintaining activity of the core, and also a sink for the entropy that the core generates.

The agential core channels material and energetic flow in the environment into building and repairing the structures necessary for its own existence. In concrete physical terms, this is done through the coupling of spontaneous reactions that release hear, and channeling part of that heat release to drive nonspontaneous reactions, such as the building of complex molecules from simple precursors. The closure of constraints in part instantiates the metabolic channels by which the energy is harnessed into work (Moreno & Mossio 2015).

The agential core is of necessity thermodynamically grounded:

> […] thermodynamic grounding means assigning a key role to the physical magnitude energy, and specifically to two basic types of energy transformations (within the system and between the system and its environment): work and heat. […] [W]ork is generated as

[125] For an extended discussion of this difference, see Nicholson (2013, 2018) and Walsh (2018).

> a result of endergonic-exergonic couplings[126], which are not spontaneous and absorb and store energy, whereas heat is related to exergonic transformations, which are spontaneous and release energy. (Moreno & Mossio 2015, 8)

Any environment in which the agential core can persist must contain at least one thermodynamic gradient. The environment must contain constantly renewed sources of energy. Thermodynamic gradients can take many forms, from the predictable and exploitable solar energy irradiating Earth's plants, to a constant flow of sulfides in the hydrothermal vents in the deep oceans. Agential cores harness such thermodynamic flow in order to maintain themselves far from thermodynamic equilibrium.[127]

Agential cores take in energy and matter from the environment, channel the release of energy from the source into building and maintaining their structure, and release degraded energy and matter (heat and waste) back into the environment. In this way, organisms act as "eddies" in the general flow of ordered energy into disordered energy. They are, as Schrodinger (1944) put it, *negentropic*:

> …a living organism continually increases its entropy … and thus tends to approach the dangerous state of maximum entropy, which is death. It can only keep aloof from it, i.e. alive, by continually drawing from its environment negative entropy … Or to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive. (Schrödinger 1992/1944, 70-71)

Thus, any agential periphery is constrained by the general requirement that it can sustain an agential core. Any agential periphery has at least one thermodynamic gradient: a source, in

---

[126] Endergonic processes are those that require the absorption of energy to proceed. Exergonic processes release energy. An exergonic-endergonic coupling is an arrangement whereby the release from one process is used to drive another process.

[127] For sophisticated forms of biological organization (organisms at higher trophic levels, or multicellular organisms) multiple kinds of thermodynamic gradients may be exploited, corresponding roughly with the complexity of the organism in question. Thermodynamic gradients tend to be exploited more *mediately*, that is, more objects in the environment are exploited as proxies for the basic energy gradient that fuels the organism. Predators do not track thermodynamic gradients; they track prey. Thermodynamic gradients themselves may not always be salient to complex creatures.

whatever form, of exploitable matter and/or energy that can be harnessed by the closure of constraints to constitute and develop itself. The agential periphery is a fundamentally thermodynamic concept, in contradistinction to the typical physicochemical characterization of an agent's situatedness.

2.4 Negentropy and Adaptive Autopoiesis

The thesis of this section is that, given the way in which the agential core and periphery are defined, two key notions can be characterized as a set of relations between the core and the periphery. They are: (i) negentropy and (ii) adaptive autopoiesis. Both are jointly realized by the core and periphery.

The negentropic character of an agent is a relation between a core and a periphery. The core is more ordered than the periphery. Such ordering requires the harnessing/canalizing of energy and matter to generate enough work to maintain the inherently precarious core. The harnessing of matter and energy produces degraded matter and energy that is released back into the periphery. This relational profile between core and periphery is a observable signature of agency.

Adaptive autopoiesis is the fundamental property of a natural agent’s core-periphery relationship. Recall, from Chapter 1, that autopoiesis is an organizational form where:

> ...the constituent processes (i) recursively depend on each other for their generation and their realization as a network, (ii) constitute the system as a unity in whatever domain they exist, and (iii) determine a domain of possible interactions with the environment (Thompson 2007, 44).

And adaptivity is:

> A system’s capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,
>
> > (1) tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,

> (2) tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity. (Di Paolo 2005, p. 438)

If one thinks of all the possible configurations of the states of the agential core, the region of viability is the set of states in which the core's s regenerative capacities work. Such a system is said to be in a *viable* state. At the "edges" of this region is the set of states in which the system is *threatened*, where the tendency is towards disintegration unless something changes. And finally, beyond the threatened region is the *futile* region where the system's dynamics cannot possibly impact the tendency to disintegration. An adaptive autopoietic system has ways of compensating for some trajectories in its region of viability (Barandiaran & Egbert 2014). Adaptivity is a dynamic, anticipative capacity that confers robustness on the self-organization of the agential core.[128]

Adaptive autopoiesis unpacks into a relation between the closure of constraints and its context. Adaptivity is a relational capacity between the way an agential core is constituted and the particular thermodynamic gradients in the agential periphery. The energetic requirements of the agential core constitute the demands of viability. The particulars of the agential periphery constitute the resources by which the agential core actively maintains itself. The energetic demands of the core and the energetic profile of the periphery jointly determine the region of viability.

Taken together, the Organizational Account of Agency says that realizing an agent in the causal structure of the world involves realizing a system with a core-periphery structure. The core achieves the formal criterion of the closure of constraints, and the more qualitative criterion of constitutive precariousness. The closure of constraints is what makes the system a unity even though each of its components is always in flux. The closure of constraints is energetically expensive; it requires constant, unceasing activity to maintain. In order for the constant, unceasing activity to be at least sometimes effective, the system must be situated in a suitable thermodynamic gradient. A thermodynamic gradient is a context with at least one source of

---

[128] Also see Skewes & Hooker (2009). The idea of anticipative potentiation is a criterion for agency that bears a very close resemblance to the notion of adaptivity.

energy and at least one sink for degraded energy. The relations of core and periphery jointly realize the negentropic character of agents, their relative organizational depth, and adaptive autopoiesis. Thus, on the OAA view, an agent is a constitutively precarious closure of constraints that exploits at least one thermodynamic gradient, thereby using its negentropic organizational depth to achieve adaptive autopoiesis.

2.5 The Organizational Account of Normativity and Goal-Directedness

The thesis of this section is that the OAA grounds agents' normativity and goal-directedness in the constitutive properties of the agential core. According to the OAA, a constitutively precarious closure of constraints situated in a thermodynamic gradient which is negentropic, organizationally deeper than the periphery, and adaptively autopoietic is the natural basis for agency. Such a system is taken to constitute its own norms, and the states that the system attains in a robust and plastic way are its own goals. In this way, normativity and goal-directedness are grounded in the organizational properties of the situated agential core as developed in sections 2.2-2.4.

The OAA holds that agency results from the fact that some subset of the closure of constraints functions to regulate the boundary conditions of the agential core (Barandiaran et al 2009, Moreno & Mossio 2015, Bich & Mossio 2017). In this way, the core not only undergoes coupling with the periphery, but *determines* which sorts of couplings happen. This is how the core is the causal source of action.

Constitutive functions constitute the system as an agent of a certain kind. For example, the nature of a core's metabolic machinery sets up the sorts of material and energetic requirements for self-maintenance.

Interactive functions are constraints that regulate boundary conditions (Barandiaran et al 2009, Di Paolo et al 2017). The regulation of boundary conditions is the influence of agent systems on the parameters that govern their coupling to their environment. Put more formally, regulation involves the following dynamics:

(1) $dS/dt=F_Q(S, E)$

(2) $dE/dt=G_Q(S, E)$

(3) $\Delta p=H_T(S)$

Here, S describes the state of the agential core, E describes the state of the environment. Equations (1) and (2) jointly describe a system-environment coupling. The parameter Q represents the set of conditions and constraints on the coupling, and the parameter p represents a subset of the constraints Q. Equation (3) describes the fact that some subset of the parameters governing the coupling depends on the state of the agent system over some period of time.

Regulation of boundary conditions typically takes the form of goal-directed behaviours (Moreno & Mossio 2015, Mossio & Bich 2017). For example, the coordinated beating of flagella to move a paramecium changes the boundary conditions for the paramecium in moving it to different material and energetic contexts. Given these interactive functions, the paramecium not only suffers interactions, it influences the course of its interactions.[129]

More generally, a system S is an agent for a particular coupling C with an environment E iff:

(1) S is a closure of constraints system in an environment E, meaning that:

(1.1) Among a set of processes a system S can be distinguished as a closure of constraints;

(1.2) The set of processes (not constrained by S) that can affect S and are affected by S defines S's environment (E); and

(1.3) S depends on certain conditions (specified by S) that in turn depend on E

(2) S modulates the coupling C in an adaptive manner.

(2.1) Where modulation indicates an alteration (dependent on S) in the set of constraints that determine the coupling between S and E;

[129] The appeal to functions in this context relies on the organizational account of functions (Moreno & Mossio 2015).

> (2.2) Adaptive means that the change in the coupling C contributes to the maintenance of some of the constraints that constitute S. (Adapted from Barandiaran et al 2009, 8)[130] (see figure 3)

That is, some portion of the closure of constraints is dedicated to adaptive modulation of coupling with the environment. Adaptive modulation involves not just harnessing energy and matter from the environment but active pursuit of states of affairs that are desirable for the agent.

The constitutive constraints establish what the system needs in order to persist as a closed, negentropic, autopoietic system. The system's needs generate a range of viable circumstances. The interactive constraints regulate the boundary conditions of the system so as to keep the system within the range of viable circumstances. The states robustly and plastically attained by the operation of the interactive constraints are the system's goals. This is how the OAA accounts for normativity and goal-directedness.

In other words, performing actions matters for the agent insofar as actions contribute to the end of persistence. Proponents of the OAA broadly agree on this point:

> [S]elf-maintenance grounds normativity. The activity of a self-maintaining system has an intrinsic relevance for itself, to the extent that its very existence depends on the constraints exerted through its own activity. Such intrinsic relevance generates a naturalised criterion for determining what norms the system is supposed to follow: the system must behave in a specific way, otherwise it would cease to exist. Accordingly, the activity of the system becomes its own norm or, more precisely, its conditions of existence are the intrinsic and naturalised norms of its own activity. (Mossio & Moreno 2015, 70-71)

> [W]e need to justify [intrinsic] normativity based on the agent's own nature. For example, with the self-individuation requirement in mind, certain asymmetrical modulations performed by the system support the processes that distinguish it from its environment. Other modulations, in contrast, may interfere with these processes and threaten to break the system down. At least one source of intrinsic norms thus originates

[130] Clause 1 has been modified to reflect the difference between processes and constraints, which was not a distinction in the original account. The substance of the account is not modified with this change.

> in the very organization of the system that maintains itself as a self-distinct, self-producing entity. (Di Paolo et al 2017, 121)

In sum, for the OAA normativity and goal-directedness arise from how constitutive and interactive constraints hang together. A particular closure of constraints is a sophisticated causal regime far from thermodynamic equilibrium; the kind of material and energetic inputs required to keep the closure of constraints structurally sound generates the basic norm of persistence. Functional subsystems and capacities for goal-directedness are evaluated by the agent in accordance to whether they contribute to the core's persistence. What makes the goal-directedness genuine goal-directedness is the normativity established by the needs of the agential core.

2.6 The Organizational Account of Autonomous Goal Sets and its Limits

The thesis of this section is that while the OAA succeeds in establishing the ultimate goal of an agent's persistence and that certain actions are normatively required by that goal, it forecloses on the possibility of agents having goals that normatively require actions that are neutral for persistence or deleterious to it.[131] I call such neutral or deleterious goals *autonomous goal sets*.[132] Autonomous goal sets capture the fact that even a deleterious or neutral goal can have the capacity to mobilize agential resources towards its attainment, which indicates a kind of autonomy to the set of goals.[133]

---

[131] There are important subtleties to the relation of A being normatively required for B. Such subtleties are developed in Broome (1999). Crucially for current purposes, the relation of normative requirement holds between a means and a goal regardless of whether the goal is one the system ought to have in a broader sense. For example, if you want to get drunk, you are normatively required to drink the tequila on the table. The normative requirement to drink tequila says nothing about the attainment of your goods. This relation is discussed in more detail in Chapter 3 of this dissertation.

[132] Autonomous goal sets operate in two types of cases. First, when agents can robustly and plastically attain many states that are deleterious for the agent. Addiction in humans satisfies that description. Second, when agents can robustly and plastically attain many states that are orthogonal to persistence. Most sentient animals appear to move or play for their own sake. In both these cases, the autonomous goal set mobilizes considerable agential resources towards the attainment of the deleterious or neutral end.

[133] Depending on the sophistication of the agent, goals of sensorimotor interaction, behaviour, affect, development, learning, and so on, are all candidate processes that realize autonomous goal sets. The norms pertinent to autonomous goal sets appear to be ends in themselves in situations where an appropriate dynamic or organizational closure is achieved. Such norms relate to the concept of open-ended complexification of agency (Moreno & Mossio

The OAA establishes two things. First it establishes the plausibility of physical systems pursuing goals, in this case the goal of their continued persistence as negentropic, dynamically organized systems. Second, it establishes that where a system is capable of pursuing goals, normative evaluations of its actions apply. We can judge whether its activities are appropriate given its goals. So, goals induce norms, and physical systems are capable of responding to norms *by* pursuing goals. This is an essential piece of giving a naturalistic account of agency. But the OAA has limitations as well.

On the OAA view, the concept of a goal is of a state which is good for the system. From this, the activities of a system are normatively evaluable only insofar as they are good for the persistence of the system as a whole.

The phenomenon of autonomous goal sets is a problem for the OAA. The formal reading of the OAA, taking the account on its own terms only, accounts for an agent's goal of persistence, and shows how the means to that goal are normatively required by that goal. But the formal reading is too restrictive on what can count as an agent's ultimate goal. Reading the OAA more loosely—i.e., less formally—can account for autonomous goal sets, but this would undermine the OAA's detailed, formal specificity, which is its greatest strength.

This tension motivates the search for an alternative, complementary account of natural goals. In contrast to the OAA's grounding of goals in what is good for the system, I suggest that an account of autonomous goal sets is not available unless the connection between goals and goods is severed. I suggest that the normative force of goals comes not from goods for the system but from a hypothetical sense of normativity on which goals and goods for the system do not line up neatly.

This section proceeds in three steps. First, I give the OAA account of norms and goals. Second, I make plausible that the OAA is committed to accounting for autonomous goal sets. Third, I

2015), and the idea of sensorimotor agency (Di Paolo et al 2017). Rosslenbroich (2014) offers an extended discussion of the empirical details of open-ended complexification with respect to the major transitions in evolution.

discuss how detaching goals from goods for the agent opens the possibility of accounting for autonomous goal sets.

As per the OAA, an adaptive autopoietic system S realizing a particular closure of constraints CC_1 has a set of norms N_1 and goals G_m whose characteristics are grounded by the specifics of CC_1. More precisely, since CC_1 is constitutively precarious, an appropriately functioning agent must continuously realize this CC_1. Constitutive precariousness thus generates a basic goal of self-maintenance G_m and fundamental evaluative standard N_1. The content of N_1 is, roughly, that if S has the capacity to plastically and reliably attain certain states (G_1, G_2, G_3, ...), the appropriateness of attaining a particular state like G_1 depends on how conducive it is to G_m. S should do the thing most conducive to G_m, all else equal.

An agent has the following generic normative structure:

$$S\text{----}>G_1\text{----}>G_2\text{----}>[G_m(N_1)(CC_1)]$$

The terms in the square brackets indicate the tight inter-definition of a closure of constraints, the goal of persistence for that particular closure, and the norm applicable to all the conducive sub-goals. The arrows indicate the relation of being a normative requirement. G_i is normatively required by G_j just in case S has adopted G_j as an end and G_i is conducive to the attainment of G_j. Arbitrarily long telic chains eventuate in the basic norm and goal of persistence.

For example, a bacterium (S) has a metabolism (CC_1) that includes, among other features, pathways for breaking down sucrose into usable energy. For this bacterium it is good for it to detect and pursue sucrose (N_1), which sets up an associated goal of being in a sucrose-rich environment (G_m). Chemotaxing to a sucrose rich area (G_m) requires a chemotaxing appropriately (G_2), that is conductively G_m. So G_2 is a subgoal of G_m. The sub-goal has its own sub-subgoals. Appropriate use of the bacterium's functional subsystems that enable the sort of sensory and motor coordination necessary for chemotaxis constitute a related capacity to attain a sub-subgoal (G_1) that is conducive to G_2. G_1 is normatively required by G_2, and G_2 is normatively required by G_m. CC_1 generates N_1 and thereby grounds G_m as a good for the bacterium.

On the OAA, either a goal is tied directly to the maintenance of the closure of constraints which constitutes the agent, or a particular goal is conducive to that goal through a chain of sub-goals. S's adaptivity grounds the possibility of such chains. What is good for the persistence of the closure of constraints is good for the agent. The means to this persistence are good for the agent in virtue of their contribution to persistence.

Bedau (1992) grounds the goal-directedness of agents with the following analysis:

> A Bs in order to C iff A Bs because [A's Bing contributes to Cing and Cing is good]

This analysis makes the condition of genuine goal-directedness that the consequences of doing B are beneficial non-accidentally for the system. Having a CC fulfills this requirement: conducing to persistence is beneficial to the system, and it is beneficial to the system because of the way that a particular closure of constraints is constituted.[134]

This analysis is limited in accounting for autonomous goal sets. Attaining a state G is only a goal for the system if it is good for the system's persistence, or if is good for attaining a state that is good for the system. Otherwise, they do not fit into the analysis.

The OAA is committed to the possibility that autonomous goal sets comprise a significant aspect of agency. For example, Thompson (2007) approvingly cites the following description of agential life:

> The animate form of our living body is [...] the place of intersection for numerous emergent patterns of selfhood and coupling. Whether cellular, somatic, sensorimotor, or neurocognitive, these patterns derive [...] from distributed networks with operational closure. In Varela's image, our organism is a meshwork of "selfless selves," and we are and live this meshwork [...]. (Thompson 2007, 49)

---

[134] Bedau (1992) usefully distinguishes the above analysis of goal-directedness from a similar, but inadequate analysis of teleological reasoning:

> A Bs in order to C iff [A Bs because A's Bing contributes to Cing] and Cing is good.

On this analysis, the good of Cing for A can be accidental. A Bs and this just happens to be good for A. The goal of persistence must be non-accidentally good for a particular closure of constraints.

The metaphor of the meshwork of selves implies that cellular, somatic, sensorimotor, and neurocognitive closures are meaningfully autonomous from each other significant degree. They key point here is that each self is a distinct closure of constraints that pursues basic goals independent of the other selves.

Similarly, for Jonas (1966), the sophisticated agential capacities associated with each major transition in evolution—motility, perception, emotion, cognition—appear to realize autonomous goal sets.

> ...such "means" of survival as perception and emotion are never to be judged as means merely, but also as qualities of the life to be preserved and therefore as aspects of the end. It is one of the paradoxes of life that it employs means which modify the end and themselves become part of it. The feeling animal strives to preserve itself as a feeling, not just a metabolizing, entity, i.e., it strives to continue the very activity of feeling: the perceiving animal strives to preserve itself as a perceiving entity--and so on. Without these faculties there would be much less to preserve, and this *less* of what is to be preserved is the same as the *less* wherewith it is preserved. (Jonas 1966, 106)

Similarly, Di Paolo et al (2017) hold that sophisticated agents instantiate sensorimotor schemes, which are distinctive closures of constraints sufficient to ground an autonomous goal set.

> The self-individuation of a network of sensorimotor schemes is related with the network of biochemical processes that constitutes the organism, but it is in fact a different kind of system. [...] [T]hese two operationally closed entities have different environments, as well as different ways of self-producing and self-distinguishing. [...] [T]he norms that emerge in each case will be also related, but different. In general ... we can postulate a relation of dependence between the closed sensorimotor network and the organismic body. Processes in the organism (metabolic, physiological, neuromuscular, etc.) enable and constrain all of the sensorimotor schemes in a network individually and in terms of how they relate to each other. However, there remains certain indeterminacy in this enabling relation, which is to be expected if the sensorimotor network can truly achieve its own autonomy. (Di Paolo et al 2017, 154)

Taken together, the three quotes above suggest that the OAA needs to account for autonomous goal sets. Sensorimotor and other sorts of "selves" realize autonomous goal sets. The capacities for moving and feeling are more than instruments of more effective persistence. The means towards movement and feeling are also normatively required by the goals of movement and feeling. The autonomous "selves" are normatively important persisting structures in their own right.

The OAA can accommodate autonomous goal sets using the naturalization scheme developed for the goal of an agent's persistence. An agent with an autonomous goal set realizes a distinct closure of constraints (CC_2) with its own set of norms (N_2) and its own fundamental goal (G_n) of the persistence of CC_2.[135] The logic is identical to the case of CC_1. CC_2 generates G_n and to normatively requires the means (G_1, G_2, G_3...) to G_n. Goals that conduce to G_n are normatively evaluable in light of G_n, N_2 and CC_2. This yields the following autonomous goal set:

$$S\text{----}>G_1\text{----}>G_2\text{----}>[G_n(CC_2)(N_2)]$$

Di Paolo et al (2017) offer a sketch of development and organization of autonomous goal sets in the sensorimotor domain. Agents have recurrent dynamics of all sorts. As certain capacities are exercised, they come together with other capacities to form self-reinforcing habits. Such habits proceed, over developmental time, to self-reinforce to the point that they become entrenched, autonomous, closures within an agent. At their most self-reinforced, such habits come to marshal significant portions of the system's resources into maintaining its characteristic closure.[136] The self-reinforcement habits display how, on Jonas' view, "[life] employs means which modify the end and themselves become part of it".[137]

---

[135] As developed in section 2.2, this is quite a stringent demand. It is more than the demand of identifying a closed loop of causal processes. There must be a meaningful distinction between a constraint and a process, otherwise there are no constraints of which there can be a closure. And without a closure of constraints there is no distinct good of persistence for the system that comes to normatively require the means to it.

[136] The account is somewhat more nuanced regarding the intermediate stages between a set of means to the goal of persistence and an autonomous goal set. See Di Paolo et al (2017), 170.

[137] This scheme is also discussed by Moreno & Mossio (2015) under the term "dynamical decoupling" of autonomous systems. The scheme is similar to that described by Di Paolo et al (2017). The differences in these accounts are not relevant for present purposes.

Briefly put, each autonomous goal set on the OAA view requires grounding in a separate closure of constraints.

But there is a crucial problem with this scheme.

> We may ask where exactly this network of [sensorimotor] schemes resides. Can we identify a physical boundary that confines it, like the membrane in the case of the cell? Is it located in the brain or perhaps in the biological body? The short answer must be negative. While we may be inclined to point to an organism's body as the locus of sensorimotor agency, it is important to stress that sensorimotor schemes, and networks of these, constitutively involve both the organic body and its environment. [...] As such, it makes no sense to try to identify their physical boundaries. We may inspect the anatomical and physiological properties of a human body and at best, we will be able to risk a very general guess as to what kind of sensorimotor agency it contributes to [instantiating]. (Di Paolo et al 2017, 152)

The problem with this account is that the discussion of the self-reinforcing pattern of sensorimotor habits describes a closure of processes, not of constraints. Without an account of the process/constraint distinction there can be no adequate grounding of CC_2. Accounting for CC_2 is crucial for grounding any autonomous goal set.[138]

The problem is that the resources for making the process/constraint distinction rest on the system-theoretic and thermodynamic criteria outlined in 2.2-2.5.[139] There seems to be an explanatory gap in the sensorimotor domain, in that there is no domain-appropriate theoretical language for articulating the process/constraint distinction pertaining to sensorimotor schemes. This problem generalizes to any other autonomous goal set: emotion, cognition, and so on.

In sum, the OAA is committed to accounting for autonomous goal sets. This grounding requires the OAA to characterize a closure of constraints in order to ground the autonomy of each goal

---

[138] Di Paolo et al (2017) discuss agency in terms of closure. But, given the argument in section 2.2., the steelmanned version of their account must involve appeal to a closure of constraints, which is developed in Moreno & Mossio (2015) and Montevil & Mossio (2015).

[139] See Montevil & Mossio (2015) for the detailed unpacking of criteria by which something can count as a constraint in a context.

set. The problem is, the way to do this is only clear in the case of the goal of persistence. In contrast, the nature of the closure of constraints that grounds goal of sensorimotor activity is obscure. Unlike with CC_1, which has a detailed accounting, no account exists of CC_2 that can sustain the constraint/process distinction in a way that ca account for autonomous goal sets.

The way to resolve the tension is to give an account of an autonomous goal set that does not derive its goal-directedness from a closure of constraints. But the OAA appears committed to the following reasoning:

(P1) Autonomous goal sets require grounding in an ultimate goal.

(P2) Ultimate goals are the persistence of a particular closure of constraints.

(P3) The persistence of a closure of constraints is a good.

(P4) Goals induce norms.

(P5) Norms require goods.

(P6) So, by P2-P5, the persistence of closures of constraints are the only goods.

(P7) So, by P1 and P6, autonomous goal sets require autonomous closures of constraints.

(P8) Closures of constraints are physical systems.[140] So,

(C) A goal is either the persistence of a closure of constraints, or a means to the persistence of a closure of constraints. For any two non-identical autonomous goal sets, there must be two non-identical closures of constraints such that each of the closures grounds a distinct autonomous goal set.

But this argument is flawed. In particular, (P5) is the weak premise.

[140] Physical here is meant in a broad sense: a pattern of organization and thermodynamic flow explained by the law-like generalizations of complexity science.

While it is natural to think that norms must be grounded in goods of some sort, a more nuanced analysis this relation undercuts the whole argument, and opens the way to an account that can do justice to autonomous goal sets.

The concept of a goal need not be tied to either (i) the persistence of any physical system, or (ii) the good of the state attained by the goal. This is because norms do not require goods *simpliciter*. Some goals are clearly grounded in what is good for a system. But there is a distinction between a particular state (say, G_2) being something the system S *ought* to produce given the goal G_n, and G_2 being a state S is *normatively required* to produce. S is normatively required to bring G_2 about, regardless whether G_n is good. Normative requirement is a sort of normativity, one that requires no commitment to the ultimate goal being intrinsically good.[141]

Normative requirement is a relation between goals and their means. Goals normatively require theis mains. We can evaluate means by their appropriateness to normative requirements. But normative requirement as a relation is neutral on appropriateness of goals. This relation allows for autonomous goal sets to have normative force without requiring the identification of a good. Without needing to identify a good, the inference that an autonomous goal set requires an autonomous closure of constraints is blocked. On this view, goals and goal sets are cheap, in the sense that there is no need to ground them in the good of a system.[142]

The normativity of agents, on this view, is hypothetical normativity. This means that for a system that has a given goal, those goals hypothetically necessitate their means, and the means are normatively required for their goals.[143] Hypothetical normativity governs most agential activities, and this is most apparent in the operation of autonomous goal sets.

---

[141] Normative requirement was identified and analyzed by Broome (1999) and applied to natural teleological reasoning by Walsh (2008).

[142] Clearly, the goal of persistence serves as a general norm that winnows out excessively deleterious goal sets. But this does not mean that the goal of persistence has anything to do with determining the positive content of autonomous goal sets.

[143] Fulda (2016, 12) puts it thus: "The normativity of teleological explanations is grounded in the relation of hypothetical necessity between goals and means. The idea is that goals require their means and means are appropriate for or good for their goals given the circumstances. So the fact that S has goal G and doing x is the means to attain G under circumstances C imposes a normative requirement on S to bring x about. On this basis we can evaluate whether x was appropriate or good for attaining G in C."

Detaching the requirements of being a goal for S from the requirements of being a good for S allows for an account of autonomous goal sets.[144] However, detaching goals from goods requires a different account of what a goal is. The alternative view makes goals much cheaper to account for. On this view, a goal is a discernible stable state of S to which other states are a means. That is, many states that S robustly and plastically attains can count among an agent's goals. A means is just a state that robustly and plastically produces the goal state.[145]

This is a crucial move. If goals are no longer grounded by their relation to persistence as a good for S, they seem to escape the OAA explanatory apparatus. But there is another perfectly naturalistic way to think about goals, one implicit in much of the OAA framework. Goals can be determined ecologically. The next section addresses this issue.

2.7 Ecological Agency and Autonomous Goal Sets

The thesis of this section is that accounting for autonomous goal sets requires an ecological conception of goals. A goal, on the ecological view, is constituted as a relation between repertoire and affordances, which are inherently relational properties that cross-cut the core-periphery boundary.

The ecological conception of goals is implicit in much of the literature around the OAA, and is occasionally explicit, though de-emphasized (Barandiaran et al 2009, Moreno & Mossio 2015, Bich & Mossio 2017). For instance, Di Paolo et al (2017) hold that:

> While we may be inclined to point to an organism's body as the locus of sensorimotor agency, it is important to stress that sensorimotor schemes, and networks of these, constitutively involve both the organic body and its environment. (Di Paolo 2017, 152)

The idea that agents are fundamentally and inextricably situated is not controversial for the OAA. However, the full implications of that idea are not always made explicit. The ecological

---

[144] To be clear, detaching goals from goods only means that *some* goals are not the attainment of goods. I take the persistence of a closure of constraints to be a genuine good for the closure in question.

[145] This idea will be fully developed in Chapter 3 of this dissertation.

view takes situatedness as a point of departure, emphasizing gross behavioural aspects of agency over the particulars of how such behaviours may be realized.

The ecological account of agency (EAA) holds that an agent is an ecologically embedded goal-directed system (Walsh 2015, 2018; Fulda 2016). On this view, the OAA is an account of how agency is realized, but is not a definition of agency. Agents are characterized by the framework of *goals*, *repertoire*, and *affordances*:

> (1) A state of affairs is a goal only if there is a purposive system that tends to attain and maintain the state by marshaling its repertoire in response to its affordances.
>
> (2) A repertoire is a biased range of potential behaviours that enables a purposive system to realize its goals in response to its affordances.
>
> (3) An affordance is a property of an organism-environment system that impedes or promotes the deployment of a purposive system's repertoire in pursuit of its goals. (Fulda 2016, 88)[146]

The ecological account of agency can secure the explanatory independence of autonomous goal sets. Since goals are by definition aspects of the gross behaviour of an agent in a context—a repertoire's response to affordances—there is no requirement for providing an underlying closure of constraints.

An ecological understanding subtly modifies the understanding of goals and norms. Understanding agents ecologically means that normativity and goal-directedness are not intrinsic to the agential core. Having a goal is not an intrinsic property of a system, no matter how sophisticated the intrinsic property is. Instead, having a goal is a property of a system in an ecological setting. More specifically, having goal X is partially constituted by the system's repertoire and the affordances that exist in a situation. Norms also depend on the system in its ecological setting.

---

[146] A few clarifications are necessary here. Purposive systems are to be understood along the lines already outlined in this chapter. They are adaptive autopoietic systems. Repertoires are modal structures: roughly, the potential inherent in a system at a time. Affordances are relational properties, roughly constituted as relations between agential capacities (pieces of repertoire) and extra-agential features of a situation. For details on the ontology of affordances, see Gibson (1979), Chemero (2003), Stoffregen (2003), Walsh (2013, 2018).

The ecological conception of agency takes the norms implied by the constitution of the agential core as context-dependent. For example, a sucrose-metabolizing bacterium has a norm of (ceteris paribus) "sucrose-responsiveness" (N_1) not only because its metabolism is constituted thus-and-so, but because it is constituted thus-and-so in its context.

Furthermore, whereas the OAA holds the bacterium to have the goal of obtaining sucrose because it plastically and robustly attains sucrose-rich states, the ecological understanding of the same goal would be that the goal is really X-in-C, where C is the context. That is, the goal of sucrose is a relation between the bacterium's repertoire and the affordances that exist in a particular situation.

This foregrounding of context makes hypothetical normativity the default assumption for agents.[147] Agents typically do things not because the action is conducive to a good for the agent; they do things because of context-sensitive responses to affordances. Particular contexts induce a set of normative requirements on the agent. As such, no closures of constraints are needed to ground autonomous goal sets.

On the ecological view, goals are constituted by the gross behaviour of an agent. We explain the production of a means to that goal by citing the goal for which it is normatively required. This explanatory structure is independent of any closure of constraints.[148]

Autonomous goal sets are built into the definition of norms and goals. The definition of norms and goals are inherently contextual makes even the most basic norms explanatorily independent from their causal realization. The ecological view can make good on the observation that the way an agent is constituted underdetermines its behaviour.

---

[147] On an ecological understanding, the goal of self-maintenance is actually self-maintenance of agent S in its ecological context C. The ecological context C is avariable ser of affordances. Attaining a goal is a matter of negotiating a system's ecological setting. The agent must ameliorate those conditions that pose a threat, and exploit those conditions that promote the attainment of its goals. Persisting, for an agent, is never just persisting per se. It is persisting by actively responding to the features of its ecological setting. An ecological context with any variability at all generates an abundance of means to the goal of self-maintenance.

[148] The details of the explanatory structure of ecological agency is developed in detail in Chaapter 3 of this dissertation.

This answer to the problem of autonomous goal sets motivates adopting an ecological view of agency. However, the definition of agency on the ecological view is silent about the realization of agency. Agency exists in the relation between an agent's repertoire and its affordances. The OAA can account for how repertoire can be realized, but it is insufficient as an account of the normative dimension because it presupposes the agential core is more important to normativity than the periphery. The ecological view takes the core and periphery as on par for explanatory purposes.

The problem for the OAA is that the dissolution of the problem of autonomous goal sets requires the importation of robust teleological explanatory machinery. While the OAA is meant to be an account of how the stable end-points of adaptive autonomous systems are teleological, doing justice to normativity requires abandoning the simple grounding of goals and norms in the organizational features of the agential core. The details of the ecological framework will be developed in Chapter 3.

The important feature of the ecological approach is that the agent is responding to relations of conduciveness of its actions *as* relations of conduciveness, not as relations explainable in terms of causality or constraints. We have already seen that closures within an agent that are not governed by self-maintenance norms are vastly heterogenous from the perspective of the physical, thermodynamic, and systems-theoretic explanatory framework of the OAA. What is heterogenous from the perspective of the causal and thermodynamic ground of the agential core can nevertheless be made intelligible from the perspective of the ecological analysis of agency. Autonomous goal sets are individuated as a particular pattern of an agent's repertoire responding to its affordances.

In sum, taking the ecological view that an agent's norms arise from the relationship of the agent's repertoire to affordances, not the intrinsic qualities (complex and sophisticated as they are) of the agential core dissolves the tension inherent in the problem of autonomous goal sets. All norms and goals are causally dependent on the closure of constraints, but they are explanatorily independent. This means that agency does not rest in the agential core's capacity to regulate its boundary conditions; that is an aspect of the agent's repertoire, but that does not determine the goals and norms operating in the context. Agency rests on a situated pattern of responsiveness between repertoire and affordances.

2.8 Conclusion

The foregoing discussion suggests that that the OAA is the best naturalistic account of how to implement agency. But the OAA is an inadequate analysis of the nature of norms and goal-directedness. It faces a theoretical bind between (1) reducing norms and goals to the persistence of a closure of constraints in the agential core, or (2) allowing for autonomous goal sets. The demands of descriptive justice require the OAA to subscribe to (2), but it is not clear how to understand autonomous goal sets from this perspective.

The OAA conflates two questions: (1) How is goal-directedness realized?, and (2) What is it to have a goal? The OAA is the best naturalistic answer to (1). But the OAA analysis of (2) is too closely tied to the norms of persistence of a closure of constraints. The OAA does address (2) in the special case of persistence goals, but not in the general case. To have the basic goal of persistence really is a matter of being constituted as a particular closure of constraints. But this answer does not work for autonomous goal sets. Autonomous goal sets can only be made sense of if one detaches the ultimate goal of a system from a system attaining a particular good for itself.

If this is right, then the explanatory framework of the OAA is impoverished. Understanding agents' norms and goals as deriving from the capacity of a closure of constraints to persist uses causal, thermodynamic, and system-theoretic explanatory tools. Such tools are inadequate for characterizing more complex forms of agency, because autonomous goal sets such as sensorimotor schemes, and other essential components of motility and sentience, are not simply decomposable to causality, thermodynamics, and systems theory.

The EAA accounts for autonomous goal sets by holding that norms and goals are not intrinsic to the agential core. The ecological framework views an agent's goal as constituted by the relation of a repertoire to affordances. This characterization is descriptively adequate for understanding autonomous goal sets, but it goes beyond the physical and systems-theoretic explanatory tools of the OAA. It requires a firmer commitment to teleological explanation.

Nevertheless, it is likely that the two explanatory frameworks are complementary.[149] The OAA offers the best existing framework for how agents are at all possible in the physical world, and of how one implements agency. The ecological analysis offers a framework for theorizing the telic and normative dimension of agents in its fullness. The OAA emphasizes the causal dependence of norms and goals, whereas the ecological account shores up the explanatory independence and open-endedness required for sophisticated forms of agency.

[149] The complementarity between the OAA and the ecological analysis motivates distinguishing two sense of "environment": the environment as *life-world*, and the environment as *thermodynamic predicament*. Making this distinction offers a three-layer theoretical analysis of an agent's environment, with each layer grounding the next. At the first layer, we have the thermodynamic predicament of the agential core, a description of the material energetic gradients surrounding the agent formulated in the language of thermodynamics and systems theory. The second layer is a characterization of a system with some degree of organizational depth: a system that given a thermodynamic predicament persists for longer than expected. The first layer constrains the possible organizations that can be agent systems in its particular context C. And finally, the third layer takes the environment to be the affordance landscape of the agent. The affordance landscape overlays onto the gradients in the environment a set of affordances. The affordance landscape both requires and enables open-ended refinements of the adaptiveness of agency, allowing for autonomous teleological and normative regimes.

3 Natural Teleological Explanation and the Frame Problem

The previous chapter has shown that the Organizational Account of Agency (OAA) is not enough to account for the rich sorts of goal-directedness observed in natural agents. In particular, it accounts for agents' autonomous goal sets in an implausible way, by positing independent closures of constraints for each autonomous goal set.

The way to fill the lacuna, I suggest, is to adopt an Ecological Account of Agency (EAA). On the EAA, goals are ontologically inexpensive, and so autonomous goal sets are a natural consequence of agents' repertoires responding to their affordances.

Taking on the EAA generates expensive explanatory commitments. Most importantly, on the EAA, teleological explanation is essential to agents. As agents cannot be reduced to their realization as complex systems, teleological explanation cannot be reduced to a species of physical explanation. So ecological agents require unreduced, natural teleology.

The thesis of this chapter is that the unreduced teleology inherent to the EAA is an acceptable commitment for an account of natural agency. There are three broad reasons for this. First, unreduced teleology is not as anti-naturalistic as it appears. Second, unreduced teleology is compatible and complementary with mechanistic explanation. Third, accepting the EAA and unreduced teleology allows for partial progress on addressing the frame problem, thereby overcoming one deep conceptual inadequacy of Cartesian views in cognitive science.

The argument will unfold as follows. Section 3.1 unpacks the EAA. Section 3.2 argues that unreduced teleology is not, in principle, anti-naturalistic. Section 3.3 argues that unreduced teleology is compatible with mechanist views that every event has a full mechanistic explanation. Section 3.4 pushes the claim of 3.3 further: not only is teleological explanation compatible with mechanistic explanation, it is indispensable in certain cases. As such, 3.4 revisits the question of what types of systems require teleological explanation, arguing that systems that realize adaptive autopoiesis require teleology indispensably, whereas systems with lower grades of downward determination only require ersatz-teleological explaantions. Section 3.5 argues that the EAA helps solve the frame problem by dissolving the need for agents to transduce inputs from the

world, which weakens the problem of how to establish context-sensitive relevance. Section 3.6 concludes.

3.1 Ecological Agency

The thesis of this section is that agency is covered in a descriptively adequate way by the ecological account of agency (EAA). The ecological approach to agency has been developed by Walsh (2008, 2012, 2013, 2015, 2018), and Fulda (2016, 2017). The exposition of the EAA and the motivation of natural teleology related to the EAA follow these sources closely.

On the EAA, agency is a gross behavioural property of a goal-directed system embedded in its ecological setting. Agency manifests as macro-level behaviours that unfold predictably over time.[150] Agency is defined by the three interdependent theoretical concepts: goals, repertoire, and affordances.

Fulda (2016, 88) offers a compact statement of the three key EAA concepts:

> (1) A state of affairs is a *goal* only if there is a purposive system that tends to attain and maintain the state by marshaling its repertoire in response to its affordances.
>
> (2) A *repertoire* is a biased range of potential behaviours that enables a purposive system to realize its goals in response to its affordances.
>
> (3) An *affordance* is a property of an organism-environment system that impedes or promotes the deployment of a purposive system's repertoire in pursuit of its goals.

An agent is a goal-directed system that marshals its repertoire in response to affordances. The capacity to be goal-directed is the capacity to bias repertoire appropriately, in response to affordances found in the agent's ecological setting.

From the basic concepts of repertoire, affordance, and goal, one can define additional features of agents: (i) teleology, (ii) means, (iii) reciprocal constitution, (iv) hypothetical normativity, (v) normative requirement, and (vi) salience/superaffordance (Walsh 2018).

---

[150] This is true of typical and paradigmatic cases of agency such as the actions of people and animals. But there are some manifestations of agency such as in bacteria that require special instruments to observe.

The presence of agents underwrites teleology. Recall from Chapter 1 that a teleological explanation is one where the presence, occurrence, or character of some process is explained by the end or goal that it serves. The linguistic markers of teleological reasoning are phrases like "x is for y", or "this entity did x in order to y", or "the function of x is to y" (Taylor 1964, Ruse 1971, Nagel 1977, Walsh 2008). In a teleological explanation, the explanans must involve such goal-based expressions. Teleological explanations cite goals as a way of making sense of an occurring action. In agents, explaining why a particular piece of repertoire was marshaled in response to particular affordances cites the goals of the agent. Teleological explanation is indispensable for understanding agents.[151]

Means are particular instances of an agent's repertoire deployment that relate to goals by the relation of conduciveness. Conduciveness is importantly distinct from the relation of causal production. Section 3.3 develops this basic idea.

Reciprocal constitution is the phenomenon whereby agents' activities construct their environments and the environment in turn constructs the agents. This is entailed by the actions of agents being responses to affordances. Affordances are jointly constituted by the agent's repertoire and the "bare" features of the environment.

Hypothetical normativity is the modal relation that holds between a means and a goal. It is th relation that captures the flexibility and adaptiveness of the relationship between goal and means (within a range of conditions).

Normative requirement is the relation that holds between an agent and a piece of its repertoire, namely, that an agent is normatively required to bring about the means to its goal. In this way, an agent's goals provide an evaluative standard for its actions.[152]

Superaffordances are the salient affordances for agents. Whereas affordances are all the potential options for action in a given situation, superaffordances are the subset of affordances that are especially relevant to the pursuit of an agent's goals.[153]

---

[151] Teleological explanation is discussed extensively in Taylor (1964), Bedau (1991, 1992), Juarrero (1999), Walsh (2008, 2012), and Fulda (2016).

[152] For developments of the concepts that follow from the ecological definition of agency, see Walsh (2018).

Explanations of actions typically cite superaffordances rather than affordances. How superaffordances are related to "bare" affordances is a theoretical problem that is especially relevant for how natural agency scales from minimal instances to sophisticated instances.[154] All agents face the problem of dealing with the overwhelming plurality of affordances.

However, natural agents do not seem to face the stupendous difficulties that beset attempts to construct artificial agents. Natural agents seem to deal with superaffordances fluidly and, for the most part, effectively. This fact stands in need of explanation. This will be the task of section 3.5.

In sum, the EEA is descriptively adequate to the phenomenon of agency. Agency is the marshalling of a system's repertoire in the pursuit of goals. Goals are pursued by response to affordances. The phenomenon of agency requires teleological explanation. Agents reciprocally constitute their worlds. The normative standards inherent in agency involve hypothetical necessity and normative requirements. These two phenomena constitute genuine rules that agents follow, not just rules that describe the system in question.

3.2 Naturalistic Explanation and Teleological Explanation

The thesis of this section is that teleological explanation is, in principle, compatible with naturalistic explanatory commitments.

The strategy for defending this thesis is as follows: first, to offer a general account of explanation, and then of naturalistic explanation. With this in place, I will outline the commitments of teleological explanation, and will then examine the standard battery of arguments purporting to show that teleological explanation's commitments are incompatible with

---

[153] The distinction between affordances and superaffordances has been marked in various ways in the literature that bears on this question. Affordances can be distinguished from saliences (Walsh 2018). Saliences are those affordances that are especially conducive to actions. Phenomenologists often distinguish affordances and solicitations, where solicitations have an action-guiding (or sometimes action-eliciting) aspect (Dreyfus 2007). Similarly, there is a distinction between the landscape of affordances and the field of affordances, where the latter is constrained by goals, actions, and other sorts of agent-end optimization processes (Rietveld & Kiverstein 2014, Withagen et al 2017, Weichold 2017, Kiverstein et al 2019).

[154] Relevant work on this can be found in Wu (2011).

naturalism. I conclude that the anti-teleological arguments are premised on a parochial reading of teleological explanation.

An explanation, at its most general, is the provision of a set of phenomena (the *explanans*) that make sense of a target phenomenon (the *explanandum*).[155] Making sense involves situating the presence, occurrence, or character of the explanandum in a wider pattern. There are two components to any explanation.

First, the relation between the explanans and explanandum must involve a change-relating invariance. This is a relation between the explanans and explanandum whereby variation of aspects of the explanans modifies aspects of the explanandum in a systematic way across a suitable range of conditions. Change-relating invariance must be present in order for explanations to situate the explanandum in a context, thereby making its presence, occurrence, or character intelligible.[156]

Second, the explanans must involve an elucidative description; that is, the explanans must be described in such a way that it enhances understanding of the explanandum.[157] Whereas change-relating invariance concerns the facts that are provided in an explanation, the elucidative description concerns the extent to which an explanation is informative.[158]

---

[155] The specific constraints on what can be included in the explanans, and the relations that exist between explanans and explanandum are subject to a variety of views in the philosophy of science. See Woodward (2014) for an overview. The general characterization here is compatible with all more specific commitments. It is a generalization of the invariance account of explanation in Woodward (2002).

[156] A classic version of change-relating invariance is deductive necessity: showing that an explanandum follows deductively from initial conditions and laws of nature cited in the explanans. However, the Deductive-Nomological model of explanation has failed under the weight of its many anomalies. Current philosophy of science holds that explanation involves the subsumption of phenomena under phenomena which may be law-like in their degree of generality, but in discipline-specific ways. This consensus is strongest in the philosophy of the special sciences, especially biology and cognitive science, where there are no obvious universal generalizations to proceed from (Woodward 2014).

[157] The formulation of explanation as an invariance and an elucidative description is due to Walsh (2012, 2013).

[158] Elucidation is partly determined by field-dependent scientific practice, with all the ambiguity that entails.

A naturalistic explanation is one that is compatible with the ontological commitments of science.[159] For present purposes, the means explanation needs to be compatible with taking reality to be a closed space of causes. There are three ways to conceive this requirements that naturalism puts on explanation, which correspond to three grades of increasing stringency for some phenomenon P to be natural.

> Grade 1: for P to be natural is for P to be causally realized.
>
> Grade 2: for P to be natural is for P to be causally realized and for P to play an indispensable role in scientific theory.
>
> Grade 3: for P to be natural is for P to be causally realized, for P to play an indispensable role in scientific theory, and for the indispensable role to be specified in terms of its realizer. (Fulda 2016, 47-48)

Grade 1 naturalism is too weak of a constraint. All manner of phenomena are natural in this manner but not thereby explanatory. Grade 3 naturalism is too strict. It forecloses on high-level constructs in the special sciences such as thermodynamics and population genetics, which rely on macrostates that are largely insensitive to the details of the microstates that realize them. As such, grade 2 naturalism is the appropriate constraint for what it is for an explanation to be naturalistic (Fulda 2016).

In sum, a naturalistic explanation is the provision of an explanans that involves a change-relating invariance relation under an elucidative description that constrains itself to phenomena that are causally realized and indispensable for our best scientific theories.

Teleological explanations elucidate their target phenomena in four cases: (i) artifacts, (ii) entities that have functions, (iii) gross behaviour of organisms, (iv) the gross behaviour of intentional agents.[160] For example: (i) the sharpness of a knife's edge is explained in part by its function of

---

[159] There are many ways of conceiving of naturalism. See Lewens (2012) for an overview of various naturalisms in philosophy of science. The description of naturalism here focuses on the methodological aspect of naturalism, with emphasis on mechanistic explanation as a constraint on naturalism.

[160] Aristotle and the explanatory frameworks derived from him clearly also used teleological reasoning to explain physical events. However, the scientific revolution's mechanistic approach greatly restricted the domain of applicability of teleological explanation. I take it that teleological explanation really is otiose for physical

slicing, (ii) hearts have valves because hearts are for pumping blood, (iii) a gazelle is leaping in order to avoid a predator (iv) Susan is opening the fridge because she wants to get some food.

All four cases appeal to purpose. It is important to distinguish original and derived purpose. Organisms and intentional agents have original purposes. Artifacts and functional entities have derived purposes. Artifacts inherit their purposes from the intentions of designers, whereas functional parts inherit their purposes from the purposes of the organisms to which they belong.

The EAA is an account of original purpose.[161] Extending the account to the purposes of artifacts and functional sub-systems is beyond the scope of this discussion.

In general, where S is an originally purposive system, ϕ is the means, ψ is the goal, and "in order to" is the telic connective, the canonical form of teleological explanation is:

S did/does ϕ in order to ψ

There are various ways of elaborating this basic form. Some teleological explanations explicitly include the intentions of agents.[162] Functional explanations pertain to the parts of S that must function in order for S to attain its goals. Teleological explanation also carries an evaluative and normative component. If ϕ contributes to attaining ψ, and S has the goal of ψ, then ceteris paribus S should ϕ, and ϕ-ing is appropriate in the situation (Fulda 2016).

The relation between an agent's goals and the means to those goals is captured in the relation of hypothetical invariance. Hypothetical invariance is a modal profile that satisfies the following two counterfactuals:

---

phenomena qua physical phenomena. It is also spurious to use teleology to explain the course and products of natural selection such as the trait structure of populations (Ayala 1970, Walsh 2008). However, such spurious uses of teleology do not entail that all uses of teleological explanation are spurious.

[161] Original purpose is analogous to Haugeland's (1990, 2002) distinction between original and derived intentionality. The problems of original intentionality apply, *mutatis mutandis*, to the present context. The difference is that there are plausible accounts of original purpose, whereas original intentionality remains obscure.

[162] The canonical form of psychological explanation has the form: S did ϕ because S desires or intends to ψ and S believes or thinks that doing ϕ is the/a means to ψ.

> (1) If the fulfillment of ψ had required Θ-ing (rather than ϕ–ing) then (ceteris paribus) S would have Θ-d.
>
> (2) If the goal of S had been ψ* rather than ψ then (ceteris paribus) S would have ϕ*-ed. (Fulda 2016, 5)

In other words, agents pursue their goals robustly and plastically. Robustness indicates that if the situation changes such that the appropriate means to ψ change, the agent's behaviour will change from ϕ to Θ. Plasticity indicates that if the agent's goal changes, the marshalling of the agent's repertoire will typically change appropriately.

There is a strong intuition that teleological explanation is insufficiently naturalistic.[163] This intuition can be sharpened into a battery of three arguments: (i) the argument from non-actuality, (ii) the argument from intentionality, and (iii) the argument from normativity. The discussion here follows closely from Walsh (2008) and Fulda (2016).

The battery of arguments presupposes a Platonic understanding of teleology, which treats it as inherently intentional, and transcendent. In contrast, an Aristotelian understanding of teleology emphasizes goal-directedness over intention, and immanence over transcendence. The Aristotelian understanding of teleological explanation is in principle compatible with naturalism. The discussion of the battery of arguments sharpens this claim.

The argument from non-actuality says that teleological explanation works by explaining actual states of affairs (S's ϕ-ing) by goals (ψ), which are typically unactualized while S is ϕ-ing. Explanation by non-actualia is not naturalistic (Walsh 2008).

The argument from non-actuality fails because it assumes that it is goals *qua* unactualized goals that do the explaining in teleological reasoning. It is correct that unactualized states of affairs cannot explain anything. But teleological explanation does not advert only to the unactualized state of affairs. What explains an instance of ϕ-ing is the relation of hypothetical invariance between ψ and ϕ. Unactualized goal states enter into this relation, but they do not define it fully.

---

[163] Expressions of this attitude can be found in Kant (2000), who famously discusses the antinomy between teleological and mechanistic reasoning. Skepticism about teleology in the biological sciences is also rife. See Hull (1969), Mayr (1983), and Ghiselin (1994).

The argument from intentionality states that goal-directedness presupposes S being designed. The capacity to achieve ψ depends on S being able to represent ψ, and this capacity depends in turn on the design of S.

On a naturalistic understanding of the world, the order manifest in nature is not the product of design. So, if teleology depends on design, it is a non-starter.

The problem with the argument from intentionality is that it illegitimately conflates derived and original purpose. Derived purpose relies on the purposes of either a designer or an organism. But derived purpose does not exhaust all kinds of natural purpose. Original purpose need not rely on the intentions of a designer. So, teleology does not presuppose design.

In addition, the relation of goal-directedness is not an intentional relation. Goal-directed systems are sometimes, but not always, intentional systems. Goal-directedness as a capacity is more primitive and pervasive in nature than the mind-like representation of a world by S. So teleological explanation is not committed to representation of goals by all agents.

The argument from normativity states that teleological explanations carry the implication that S ought to, ceteris paribus, do ϕ if ϕ is a means to ψ. But ψ can only carry the normative implication if ψ is good (Bedau 1991, 1992). This entails that teleological explanations rest on making states of affairs in the world intrinsically normatively evaluable. Naturalism does not admit intrinsically evaluable states of affairs, so the normative commitments of teleology are an expensive, anti-naturalistic commitment.

The response to the argument from normativity is to distinguish two types of normative relations between means and ends. Broome (1999) calls the two normative relations "ought" and "normative requirement".

“Ought” is a detaching relation. This means that if Z is the goal, and Y is the means to the goal, one can conclude that Z ought to be attained. Normativity based on “ought” entails that goal-directed systems can only have normativity if the goals are good.

“Normative requirement” is a non-detaching relation. The relation of normative requirement between means and ends entails only that *if* Z is the goal, the means Y should be produced.

Normativity based on "normative requirement" holds the norms to be indexed to the goal-directed system.[164]

So, the relation of normative requirement holds between means and goals regardless of the intrinsic goodness of a goal. If S has a particular goal, S is normatively required to bring about the means to that goal, but this does not entail that the goal itself is good (Broome 1999). To derive the goodness of the goal requires the stronger, detaching relation of "ought".

The normativity of natural purposiveness is the normativity in the non-detaching sense. It makes no evaluative claims about the goal.[165] This way of thinking about the normative entailments of S having a goal does not require intrinsically evaluable states of affairs. So teleological reasoning does not breach that commitment of naturalism.

In sum, the three arguments against naturalistic teleological explanation arise from assimilating all teleological explanation to the Platonic model of teleology. Platonic teleology works for intentional action explanation, and for explanation in cases of derived purposiveness, but it does not exhaust all teleological explanations. In contrast, Aristotelian teleology is immanent to the originally purposive systems that realize it. An organism's matter and its goal-directedness are both inherent in its nature.[166] The activities and capacities of goal-directed systems are not explained by non-actualia, or by design. Further, the normativity that applies to goal-directed systems inheres in what a goal-directed system ought to do given a particular goal, not in the intrinsic goodness of states of the world.

So, teleology is compatible with naturalism. Goal-directedness is a complex system-level property that is in principle analyzable by complexity science.[167] And the normativity of teleological explanation does not entail intrinsically evaluable states of affairs.

---

[164] This point was first made by Walsh (2008).

[165] As discussed in Chapter 2, goals are ontologically cheap. And there is no fact of the matter about whether the robust and plastic pursuit of certain states by agents is good outside the context established by the agent's constitution.

[166] This particular kind of immanent goal-directedness is realized in systems constituted as an ecologically situated closure of constraints discussed in detail in Chapter 2 of this dissertation. The functional parts of such systems inherit their purposiveness from the natural purposiveness of the whole.

[167] And in fact inheres in systems that attain the capacity for grade III downward determination.

This discussion establishes that teleology is compatible with naturalism, but it does not yet establish that teleological explanations play an indispensable theoretical role. The next section does this.

3.3 The Mutual Autonomy of Teleological and Mechanistic Explanation

The thesis of this section is that teleological explanation explains patterns in the world that are not captured by mechanistic explanation. This makes teleological explanation autonomous from mechanistic explanation and therefore theoretically indispensable.

Just as mechanistic explanations explain by invoking certain kinds of invariance relations, so teleology explains by invoking a different sort of invariance relation, namely hypothetical invariance. The intuition that teleological explanation is excluded by mechanistic explanation is misleading.[168]

Mechanistic explanation is an explanatory mode where the presence, occurrence or character of some phenomenon is explained by citing the mechanism that produces it. Mechanistic explanations deal only in "productive" relations. The linguistic markers of mechanistic reasoning are terms such as "binding", "bending", "stretching", "pushing", all terms indicating the transmission of force (Machamer, Darden & Craver 2000, Darden 2008). Machamer, Darden, and Craver (2000) offer an influential definition:

> Mechanisms are activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (Machamer, Darden, and Craver 2000)

In mechanistic explanation, to explain some phenomenon E, one cites a set of phenomena $c_1 \ldots c_n$ that suffices to bring about E, where $c_1 \ldots c_n$ is a set of "bottom-out" activities of a mechanism. The set $c_1 \ldots c_n$ of bottom-out phenomena is a suitable description of initial conditions and the

[168] The argument in this section is drawn from Walsh (2012).

thick causal powers of the components of the explanatory mechanism—the way that they bend, stretch, push, or otherwise influence each other.[169]

Citing bottom-out phenomena is an elucidative description. In a particular explanatory context, the bottom-out phenomena have simple intrinsic propensities. The various ways that $c_1...c_n$ combine and influence each other constitutes the mechanism of interest (Walsh 2012).

All explanations cite change-relating invariances. For mechanistic explanation the change-relating invariance holds between the components of the mechanism $c_1…c_n$ and the effect to be explained E. Call this $<c_1…c_n,E>$. This change-relating invariance supports the counterfactual that systematic changes to some component c_i lead to systematic changes in aspects E across a suitably wide range of circumstances. $<c_1…c_n,E>$ is characteristic of the modal profile of mechanistic explanation.

For example, in explaining why a block on an inclined plane slides down the plane (or alternately fails to slide), we cite the fact that the block is subject to two forces: the pull of gravity down the block and the force of friction. The force of friction depends on the magnitude of the normal force and the incline, whereas the pull of gravity depends on the magnitude of the incline, the mass, and the force of gravity. The pull of gravity is a resultant force that depends on the incline of the plane. The frictional force can be manipulated by intervening on the normal force or the angle of the incline, and the force pulling down the block can be manipulated by intervening on gravity, mass or the angle of the incline. If the pull of gravity on the block is greater than the frictional force, the block slides down. The specification of the normal force, the angle of the incline, the mass of the block, the coefficient of friction, and the strength of gravity, taken together, are the $c_1...c_n$. Modifying any of those components would affect the presence, absence, and occurrence of the block's sliding E.[170]

There is a powerful intuition that each phenomenon that is putatively explained teleologically can be explained mechanistically. Because of this, mechanistic explanation supersedes or

---

[169] Here thick causal concepts are in contrast with thin causal concepts. For example, bending and pushing are thick causal concepts; the transmission of conserved quantities is a thin causal process. See Cartwright (2004).

[170] This example is from Woodward (2002).

undercuts teleological explanation, since all goal-directed systems are causally realized, and each even involving the system or its parts has a complete mechanistic explanation.

This intuition has been developed into the argument from explanatory exclusion (Kim 1989, 1999, 2006). The argument runs as follows: (1) every event has a complete mechanistic cause; (2) every event has a complete mechanistic explanation; (3) every event has at most one complete explanation; (4) mechanistic explanation is complete and exhaustive; (5) for there to be a role for goals in science they would have to cause phenomena that mechanism cannot (Fulda 2016, 9-10).

If this argument is right, the universality and pervasiveness of mechanistic explanation in science excludes teleological explanation from serving a useful role.

However, premises (3) and (5) are too strong. Goal-directedness does not cause phenomena that mechanisms cannot. To deny this would be anti-naturalistic. But goal-directedness does explain phenomena that mechanism cannot. Teleological explanations capture a modal profile that goes beyond the mechanistic modal profile. Teleological explanations capture the *plasticity* and *robustness* of the relationship of means to goals.[171] This idea is captured by the modal profile of hypothetical invariance, as seen in the two counterfactuals already seen above:

> (1) If the fulfillment of ψ had required Θ-ing (rather than ϕ–ing) then (ceteris paribus) S would have Θ-d.
>
> (2) If the goal of S had been ψ* rather than ψ then (ceteris paribus) S would have ϕ*-ed. (Fulda 2016, 5)[172]

Here, the hypothetical invariance relation invokes these two counterfactuals to relate the means M and goals N in a systematic way, <M,N>. While every instance of ϕ-ing (or Θ-ing) has a complete mechanistic explanation, the modal profile described by <M,N> does not.

---

[171] To be more precise, explanations citing original purposiveness appeal to plasticity, robustness, and adaptiveness. Explanations citing derived purposiveness typically leave the original purposiveness on which they depend implicit.

[172] To be clear, mechanistic reasoning can explain particular instances where the behaviour of a system fits these counterfactuals. In fact, in constructed artifacts, we have more or less complete mechanistic accounts of their counterfactual responsiveness. In the case of agent systems, there is no general account of how all agents attain such counterfactual responsiveness.

The occurrence of ϕ and its variants can be explained in at least two different ways. The fact that teleological, but not mechanistic, explanation capture the modal profile developed above means that teleological and mechanistic explanations of the same sequence of events are not reducible to one another. For any causal relation <C,E> there are two elucidating descriptions that allow for explanation.

First, <C,E> can be described under productive relations. This description is obviously elucidatory. It tells us how C produces E, invoking whatever thick causal concepts are required. It supports counterfactuals about what would happen if C were intervened on.

Second, <C,E> can be described under relations of conduciveness. This description tells us that E is a goal to which C is an appropriate means. Such a description is elucidatory but in a distinct way from the productive description. It tells us why C is occurring, and it supports the two counterfactuals above that give a modal profile describing a robust and adaptive means-ends relationship (Walsh 2012).

For example, where C is the triggering of a shiver reflex and E is the body temperature being at 37 degrees Celsius, the mechanistic description tells us how C produces E, thereby elucidating the phenomenon. The teleological explanation tells us why C was triggered, namely because the body being at 37 degrees Celsius is a goal of the thermoregulatory system. The triggering of the shiver reflex is elucidated as an appropriate means to the attainment of the set point, i.e., its conduciveness to the set point.

This situation suggests that mechanistic and teleological explanations are autonomous, and complementary. Mechanistic explanation explains how phenomena occur, leaving unelucidated why they occur. Teleological explanations explain why phenomena occur leaving unelucidated how they occur (Walsh 2012).

So, the universality and completeness of mechanistic explanation does not undermine teleological explanation in the case of goal-directed systems. In such systems, the change-relating invariance <M,N> adds a further layer of elucidation, making clear why, given a system's goals, a particular phenomenon occurs, and why that phenomenon’s presence is counterfactually robust to changes in the goal-directed system’s situation.

3.4 The Realization of Goal-Directedness

The thesis of this section is that teleological explanation is only indispensable for systems with grade III downward determination.[173] Such systems offer a sufficient causal realizer of goal-directedness. Explaining processes within grade III systems calls for teleological explanation because such systems have the resources to implement goal-directedness.

To sharpen this claim, I will contrast teleological explanation to two cases of explanation that are also putatively non-mechanistic, and involve change-relating invariances with non-mechanistic modal profiles, but which are not teleological.

The three cases for comparison will be:

> (i) Explaining why a marble dropped in a bowl comes to rest at the bottom of the bowl for a wide range of initial conditions.
>
> (ii) Explaining why a fire tends to consume the log it was lit on with considerable plasticity and robustness in the way it does this.
>
> (iii) Explaining why a bacterium tends to bias its tumbling so as to approach nutrients and avoid toxins.

All three cases have features that resist straightforward mechanistic explanation, but only (iii) requires genuine teleological explanation. (i) does not need appeal to teleology. (ii), despite being strongly analogous to (iii), is only *superficially* teleological. The essential difference is on of modal profile; the counterfactuals supported by (ii) are ones that express causal necessity, whereas the counterfactuals supported by (iii) express hypothetical necessity.[174] Both (ii) and (iii) entail the autonomy of explanation at the macro level, but in (iii) there are also invariances that are characteristic of the modal profile of goal-directedness.

---

[173] This view entails that teleological explanation is dispensible to artifacts. To be clear, teleology is dispensible in these cases only in that the purposes of an artifact are inherited from the purposes of some systems that have intrinsic teleology. The intrinsic part of this telic chain is indispensible. The artifact part of this telic chain is, in principle, redescribable so as not to need teleological terminology.

[174] Causal necessity is the necessity of productive relations unfolding as they do. Hypothetical necessity is the necessity that exists between goals and their means. The details of the distinction are developed in detail in Walsh (2012, 2018) and Fulda (2016).

In all three cases there appears to be two distinct types of explanation operating, each with their own characteristic change-relating invariances and elucidative descriptions. In case (i), in addition to productive, mechanistic relations, there are also equilibrium relations. In case (ii), mechanism is supplemented by "far-from-equilibrium" relations, and in case (iii) there are teleological relations of hypothetical invariance operating as well. Below is a table summarizing the shorthand for this section:

Case	**$c_1 \ldots c_n$**	**E**	**C**	**R**	**Type of change-relating invariance <C,R>**
(i) Marble in a bowl	Position and initial velocity of marble relative to bowl.	Trajectory of marble in bowl from start to rest.	The shape of the bowl and the operation of gravity.	Marble rests at bottom of bowl across variation in set-up conditions	Equilibrium
(ii) Fire consuming a log	Micro description of matter and energy at time of lighting of fire.	Fine-grained chemical events composing the flame over time.	Relevant thermodynamics, causal feedback, contextual factors, macroscopic order of flame.	Log will turn to ash across a variety of set-up conditions.	Far from equilibrium
(iii) Bacterium chemotaxing up a sucrose gradient	Bottom-out description of state of chemotaxis machinery at setup.	Trajectory of states of chemotaxis machinery in bacterium.	Context of chemotaxis machinery within a closed metabolic network.	Swimming up-gradient attained across a variety of set-up conditions.	Teleological

3.4.1 Equilibrium Explanation

Let $c_1 \ldots c_n$ be the bottom-out set-up conditions--the position, initial velocity, and trajectory of the marble, the properties of the bowl).

Let E be the trajectory of the marble between the starting point and when it comes to rest.

Let R be the regularity instantiated by E—the fact that a suitably wide range of initial states (particular ways $c_1 \ldots c_n$ could be) eventuate in a marble resting at the bottom of the bowl. R expresses the fact that there is a modal profile that needs explaining that is over and above the modal profile realized by $c_1 ... c_n$. Specifically, R is explained by citing a different set of change-relating invariances.

Clearly, $c_1 \ldots c_n$ are the difference-makers for the presence, occurrence, and character of E. Changes in any of $c_1 \ldots c_n$ entail predictable changes in the details of the marble's trajectory before the end-state is reached. In contrast, the difference-makers for R involves the set of change-relating invariances under a macro description C which cites the shape of the bowl and the operation of gravity.

The explanation of E by citing $c_1 ... c_n$ is the straightforward mechanistic explanation that cites the relevant productive relations. The explanation of R by C is different. While a particular E realizes R, C will involve the following sort of counterfactual:

> If $c_1 ... c_n$ had had been $c_1 ... c_n$* within a set of parameters P, E* would still realize R.

This counterfactual is not particularly important theoretically, but it expresses the limits of mechanistic reasoning. Woodward (2013) articulates the trouble well:

> …to the extent that there is reason to think that some system would behave in the same way (for some behaviour of interest) even if the spatio-temporal organization of the components were changed considerably, or even if some or many of the components were replaced with others with different causal properties—then the motivation for constructing explanations that focus on them is correspondingly diminished, at least on a conception of explanation according to which this involves the exhibition of difference-making factors (Woodward 2013, 57)

Equilibrium explanations capture the fact that the end-state is insensitive to differences in initial conditions. Nevertheless, the insensitivity of macro states to micro realization is not enough to establish that the explanatory autonomy of the macro states because each macro-state has a micro-state realizer, even though it is multiply realizable.

Taking all this together, in equilibrium explanation the existence of a macro-scale dynamical attractor makes the explanans robust to changes in micro-mechanical detail within a certain range. It is enough to explain R to cite the macro detail in C, which contains different change-relating invariances and descriptions.

Of course, the endpoint to which the marble tends is not a goal, because the marble-bowl system does not realize goal-directedness. The macro-state attractor which the marble-bowl system realizes allows for a degree of *robustness* in how the marble attains the end state, but not *plasticity*.

3.4.2 Far-from-equilibrium Explanation

In case (ii) let $c_1 \ldots c_n$ be a description of the bottom-out events and activities making up a flame. $c_1 ... c_n$ include the temperature, the rate of combustion, the presence of chemicals of a certain variety, the shape of the fire, and so on.

Let E be the sequence of events as the fire consumes a particular log specified in terms of the productive relations among $c_1 ... c_n$.

Let R be the fact that a suitably wide range of initial conditions would have eventuated in that log turning to ash. For example, the fire might have started at the other end of the log, or the air temperature might have been different within suitable parameters.

Let C be the facts that make R the case.

As in case (i), the explanation of E by citing $c_1 ... c_n$ is unproblematically mechanistic, citing as it does the sufficient bottom-out activities for explaining the dynamical evolution of the flame and the log.

The explanation of R by citing C is richer than in case (i). R indicates that fires have a degree of plasticity and robustness in how they attain an end-point. The explanation of R by C adverts to

the fact that fires are systems with either the higher end of grade I or low-end grade II downward determinative capacity.[175]

To explain the robustness and plasticity of the fire's log-consuming behaviour one must cite a change-relating invariance <C,R> between the fire's constitution as a system and the fuel-consumption that is necessitated by the fire's being far from thermodynamic equilibrium. Being so constituted, fires realize their endpoints in ways that are more robust and plastic than marbles in bowls.

Fires are constituted as systems in more complicated ways than the marble-bowl system. As with all dissipative systems, a fire's existence depends on material and energetic flow through the fire. As such the invariance <C,R> supports a modal profile where the fire's dynamics have a degree of robustness and plasticity in how they achieve their end-point.

The explanation that cites the <C,R> change-relating invariance in case (ii) is close to that used in teleological explanation, but not quite it. The crucial difference between fires and goal-directed systems is in the modal profile supported by the change-relating invariance. Two types of necessity are involved here: causal and hypothetical (Fulda 2016, 17-19). Consumption of fuel is a causal necessity for a fire to constitute itself, and the fire as a dissipative system makes fuel-consumption a plastic and robust process. The explanans must cite, among other things, the following counterfactual:

> If the fulfillment of self-maintenance had required Θ-ing (rather than ϕ–ing) then (ceteris paribus) the fire would have Θ-d.

This captures the robust and plastic attainment of fires' endpoints. But this is not yet the hypothetical necessity of goal-directedness. Fires don't realize the second aspect of hypothetical necessity for the simple reason that their endpoints are fixed as a matter of nomological

---

[175] There is ambiguity about what grade of downward determination is reached by fires. This reflects the practical difficulties of thinking about closures of constraints (Moreno & Mossio 2015). A raging fire certainly seems to individuate itself against its conditions. But to what extent this is a consequence of being a dissipative system, and to what extent the constraints realized by the fire are self-reinforcing is an open question. Certainly there is causal feedback. But it is less certain that there is a work-constraint cycle in operation. Fires are a genuine gray area in the definition of grade II downward determination.

necessity. This is because fires must consume fuel to self-maintain. There is insufficient modal breadth with respect to fires' endpoints. As such, the counterfactual

> If the goal of the fire had been ψ* rather than ψ then (ceteris paribus) the fire would have ϕ*-ed.

holds in only a degenerate sense. Fires have a range of endpoints, but these endpoints aren't goals. The end-points are too closely, and causally, dependent on the fire's constitution as a dissipative system. The relation between fires' attractor states and the paths to such states do not have the necessary flexibility to attain the modal profile of hypothetical invariance. As such, <C,R> does not need to cite hypothetical necessity, only a sophisticated, non-mechanistic causal necessity.

3.4.3 Teleological Explanation

In the case of a bacterium chemotaxing up a sucrose gradient, let $c_1 \ldots c_n$ be the bottom-out description of the activities of the molecular biological network that mediates the "sucrose" signal.[176]

Let E be the trajectory of the bacterium's biased tumbling towards the source point of the sucrose gradient.

Let R the fact that bacteria of this type reliably move towards the gradient's source-point.

Let C be the facts that make R true.

An essential change-relating invariance in the explanation of the bacterium's R, <C,R>, is teleological in nature. It cites the fact that the presence of $c_1 \ldots c_n$ happens because it is a means to the bacterium's goal of obtaining sucrose. $c_1 \ldots c_n$ happens because it is conducive to sucrose-

[176] The chemotaxis mechanism has been worked out in considerable detail. Our understanding of how chemical gradients in the external world are detected through the properties, arrangement and interactions of proteins embedded in the cell membrane, and how such signals are transduced into differential activation of flagellar motors is a triumph of mechanistic explanation (Barkai & Liebler 1997; Alon et al 1999, Bi & Sourjik 2018). Showing that teleological explanation has an autonomous and informative role here is not merely an argument from ignorance of mechanistic detail.

attainment. And furthermore, $c_1...c_n$ is hypothetically necessary for appropriate nutrition. The two counterfactuals characteristic of teleological explanation both hold.

Why does the bacterium's chemotactic machinery satisfy the modal profile of hypothetical invariance? This is because the bacterium, as a whole, satisfies grade III downward determination. As an adaptive autopoietic system it has a rich set of potential behaviours that is potentially biased by various sorts of goal-pursuit. The difference between the bacterium case and the fire case comes down to this biased set of potential behaviours—the bacterium's repertoire. A fire's repertoire is comparably impoverished, and appealing to it is less elucidative. In contrast, the bacterium's extensive capabilities allow for various types of biasing.[177]

Put another way, what makes case (iii) a genuine teleological explanation is the integration of the mechanism of chemotaxis into the whole system. Stripped of context, the operation of a bacterium's chemotaxis machinery is just a more complicated instance of the unfolding of the causal necessity inherent in any far-from-equilibrium system. Such a decontextualized explanation would be exactly the same in kind as the explanation of a fire's activity. But in the context of the whole bacterium, the process of chemotaxis E conduces to sucrose, and thereby feeding.

If we take the chemotaxis machinery out of context, we can account for its dynamical trajectory in the same way as in the case of fire. Namely, the chemotaxis machinery is a far-from-equilibrium system that, with the appropriate triggers, can plastically and robustly attain an end-state across a variety of set-up conditions. However, in full context, viewed as a sub-system of the whole bacterial system, the chemotaxis machinery is not just a far-from-equilibrium system with robust and plastic causal potential. It is hypothetically necessary for attaining sucrose.

What makes <C,R> teleological in this case is the constitution of the bacterium as a whole organism. The bacterium is a way of naturally realizing goal-directedness. The bacterium is constituted as a whole in a more robust way than a fire. This is because bacteria, but not fires, have three features: organizational depth, metabolism, and adaptive autopoiesis.

---

[177] Because elucidation is relative to scientific practice, the usefulness of teleological explanation exists on a continuous spectrum. Nevertheless, different sorts of objectively identifiable classes of systems make an almost categorical difference in how useful it is.

Organizational depth tracks the extent, thickness, and sophistication of the feedback interactions interaction that constitute a self-organizing system. A system that transforms A into B where the rate of B's production is modulated by some other process F is more organizationally shallow than a system where—in addition to A, B and F—there is a process G that modulates the production or presence of F. F is a constraint on the process of turning A into B, and G is a constraint on the production of F.

In the context of living systems, organizational depth tracks the sophistication of regulatory interactions. In case (ii), fire has some degree of organizational depth in that there is feedback from the activities of the fire into maintaining its macroscopic order; that is, the burning of A to B produces C (released energy) that facilitates the contact of A and B at sufficient energy to burn A into B. Most dissipative systems have non-zero organizational depth.

Organizational depth is a measure that scales continuously but that generates qualitatively different regulatory regimes.[178] Fires and bacteria represent different sides of one such qualitative shift.

Metabolism is the way that a bacterium is a system that attains grade III downward determination. Both the fire and the bacterium persist based on the energy released from the breakdown of chemical compounds. The crucial difference is that for bacteria the energy released by the breakdown of nutrients is channeled not merely through release of thermal energy, but of free energy—energy available to do work. This release of free energy is used by its biochemical machinery to drive nonspontaneous processes and generate ordered structure (Moreno & Mossio 2015). The sophistication with which the energy released is harnessed is a function of the organizational depth of the bacterium.

Metabolism is the primary driver of an organism's negentropic character. In contrast to the case of fire, a metabolic network produces organization not just at the level of macroscopic order, but in its activities are precisely organized so as to reproduce the metabolic network across a wide

---

[178] These are marked out by the major transitions in evolution. See Maynard Smith & Szathmary (1995), Rosslenbroich (2014), Moreno & Mossio (2015), Montevil & Mossio (2015). Metazoans are organizationally deeper than non-metazoans. Creatures with nervous systems are organizationally deeper than those without. Creatures with many-layered nervous systems are organizationally deeper than those with fewer layers. And similarly for other sorts of regulatory interactions, such as the presence and depth of homeostatic or immune mechanisms.

range of circumstances. A metabolic system is afforded the opportunity to persist in a more open-ended way than a merely dissipative system.

Importantly, the processes that realize organizational depth in the bacterium require their own persisting organizations through the flux of molecules and energy in a living cell. This is achieved via autopoietic organization. The bacterium is a realization of autopoiesis in the chemical domain. It satisfies three conditions:

> (AP1) It is an interdependent network of processes such that each of the processes, in isolation, would tend to run down.
>
> **(**AP2) It has a boundary that separates the network from what is "outside" the network.
>
> (AP3) It has interdependence between the boundary and the network (the network and boundary are each necessary for the other).[179]

A bacterium is the minimal instance of autopoiesis in the chemical realm. Fires satisfy AP1, but not AP2 and AP3.

Autopoiesis, metabolism, and organizational depth make the bacterium an integrated enough system so that the goal associated with R elucidates the presence of the means to that goal. In contrast, (ii) is self-organizing and has some robustness of activity, but it lacks autopoietic organization and relative organizational depth. In that case, the presence of the fine-grained events is not really elucidated by the pseudo-goal that is robustly and (somewhat) plastically attained. For marbles in bowls and fires, there is one direction of counterfactual dependence that suffices to account for their dynamics. For bacteria and more complex goal-directed entities, there is an additional, teleological, layer of explanation. Teleological explanation depends on the fact that the precarious individual processes that constitute the bacterium hang together only because persistence (for some time) of the complex organization is attained. Metabolism and autopoiesis are necessary for the indispensability of this mode of explanation.

---

[179] This definition is a rather generic definition of autopoiesis. For subtleties, see Thompson (2007), Varela et al (1991), Di Paolo (2005) and Wheeler (2011).

So, the explanation in case (iii) is teleological. It is teleological because of the nature of the organism that the chemotaxis machinery finds itself in; viewing the machinery in isolation--i.e., giving an explanation only in terms of productive relations--misses the obvious modal profile realized by the bacterium, namely that the bacterium's goal of feeding hypothetically requires this particular means in this particular circumstance, and other means in a range of other circumstances.

In other words, the explanation in (iii) adverts to an end-state that is actually a goal because the system actively regulates the state changes that lead to the end-state. The bacterium, unlike the marble or the fire, biases its repertoire in goal-conducive ways. Equilibrium explanations are not teleological because their explananda do not have repertoires. Far from equilibrium explanations are not teleological because the limited repertoires of dissipative systems are not actively regulated by the system itself. Bacteria have sufficient capacity to support the hypothetical necessity of the micro-states that are the means to its various goals; they are there because they conduce to goals, not just because they contribute causally to the realization of the goals.

3.5 The Burden of Superaffordances and Higher-Order Agency

The agent's world is made up of affordances, namely the agent-indexed features of the environment, to which the agent's repertoire can be deployed to attain the agent's goals, or those aspects of the environment that impede the satisfaction of the agent's goals. Complex systems with adaptive autopoietic organization realize a capacity to constitute worlds in the sense used here, and the only way to make sense of their activities is teleologically.

The ecological analysis of agency is useful for cognitive science because it offers a framework for dissolving the frame problem.

As discussed in Chapter 1, the frame problem is a deep epistemological issue for Cartesian views of agency. Shanahan (2016) summarizes it thus:

> How is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language-like representations of the sort used in classical AI? (Shanahan 2016, 5)

The problem is not limited to the context of propositional, language-like representations. It is a problem for any agent. Cartesian approaches to cognition are particularly vulnerable to it. This is because Cartesian approaches to cognition require a two-stage process: (1) the agent must *transduce* the raw inputs of the world into a usable format, and (2) *construct* significance over the worldly features formatted in that way. Both steps introduce combinatorial explosion.[180]

The frame problem is one of how an agent is to establish a frame--a processing context where, in general, only the relevant aspects of the world and actions need to be processed.

Dennett (1987) illustrates the problem as follows:

> [A]n intelligent agent must engage in swift information-sensitive 'planning' which has the effect of producing reliable but not foolproof expectations of the effects of its actions. That these expectations are normally in force in intelligent creatures is testified to by the startled reaction they exhibit when their expectations are thwarted. ... To be surprised you have to have expected something else, and in order to have expected the right something else, you have to have and use a lot of information about the things in the world. (Dennett 1987, 7)

The illustration by Dennett depends on a prior question: how can an agent navigate a complex world with reasonable expectations of the effects of its actions without having to perform the cycle of transducing inputs and then constructing significance at all times?

The task of constructing open, holistic, context-sensitive relevance appears insuperable because the Cartesian agent must accomplish two tasks: (1) transduce aspects of the world into a usable format for its processing, and then (2) determine which of the inputs or their combination are the relevant ones for action-guidance. Once this is done, the system can process inputs and related it to its action appropriately. The problem is that proper transduction depends on grasping which

---

[180] See Wheeler (2005, 2008), Dennett (1987), Vervaeke et al (2009), and Shanahan (2016). Historically, the frame problem was about predicting or accounting for changes in the environment caused by an agent's actions. This version of the frame problem has been solved at least in principle. The epistemological version is an open question (Shanahan 2016).

context applies, but so to figure out which context the agent is in, it must again solve a frame problem. A vicious regress is apparent.[181]

Ecological agents do not encounter the frame problem because to be an agent is to construct (or constitute) one's own frame.[182] Agents construct and respond to selected features of the environment as affordances. This starting point sidesteps task (1) in the previous paragraph. Natural agents exist primarily in a mode of absorbed coping. They do not begin as subjects over against a world, and so do not face a transduction problem. The relevant aspects of the unimaginably complex network of agent-environment relations do not need to be, as a matter of logical necessity, transduced into the agent. Rather, agents begin as inescapably embedded in a landscape of affordances partly of their own making.

The frame problem does not arise for natural agents because a natural agent's world is inherently agent-indexed. Worlds are agent-relative domains of possible action. Such worlds are usefully understood as affordance landscapes: structures of features that elicit actions of attraction, repulsion, or elicit only indifference. Affordances ultimately advert back to an agent's repertoire and goals, and in this they are bounded by the agent's capacities.

For instance, the status of particular molecules as nutrients depends on the particular organism's metabolic capacities; sucrose is attractive to a sucrose-metabolizing bacterium and indifferent to bacteria that lack sucrose metabolism. The bacterium in no sense needs to determine whether sucrose is a source of nutrition—it is a consequence of the bacterium's constitution that sucrose is a nutrient.

But this solution to the frame problem generates a *burden of superaffordances*—the ecological analogue of the frame problem. This problem arises because the sophistication of an agent's repertoire entails that the complexity of the resulting affordance landscape will outstrip the complexity of its repertoire. Hence, affordance landscapes are poor guides for action, simply because there are too many affordances in every situation. So, while natural agents do not

---

[181] See Shanahan (2016), and Wheeler (2005).

[182] This point is perhaps the central insight of situated robotics and its cognate research programs, which flourished in part as a response to the limitations of the standard approaches of the time to the frame problem (Dreyfus 2008, Wheeler 2008).

encounter the first aspect of the frame problem because they start with affordances, they encounter (2): the problem of determining which repertoire is relevant for a given affordance landscape.

This problem can be illustrated with a simplified example.[183] Consider an agent in a situation consisting of three affordances: A, B and C. For example, a squirrel is in a situation that affords (A) drinking water, (B) stuffing its cheeks with acorns, (C) preening itself. Suppose that the agent has four ways of combining the affordances present: it may be able to do some simultaneously; it may have to do some but not others, it may be able to sequence acting on the affordances, and some of the affordances may depend on the ability to have other affordances, in a constitutive, not temporal, sense. These four functions generate a combinatorial picture of the situation, summarized in the following table:

Affordances	Combinations	Temporal Sequencing	Conditionals (-->)	Temporal Conditionals
A	A and B and C	A then B then C	A-->B	(A then B)-->C
B	A and B and not C	A then C then B	A-->C	(B then A)-->C
C	A and C and not B	B then A then C	A-->B and C	(B then C)-->A
	B and C and not A	B then C then A	A and B-->C	(C then B)-->A
		C then A then B	A and C-->B	(A then C)-->B
		C then B then A	B-->A	(C then A)-->B
		A then B and not C	B-->C	
		B then A and not C	B-->A and C	
		B then C and not A	B and C-->A	
		C then B and not A	C-->A	
		C then A and not B	C-->B	
		A then C and not B	C-->A and B	

[183] In this example, the combinatorial possibilities are simplified for the sake of illustration.

In the table above, the leftmost column lists the affordances in the situation. The other four columns list the higher-order affordances (or superaffordances) of the situation. These are potential actions that can arise from the combinations of affordances. These higher order affordances depend on other pieces of the agent's repertoire.

The most obvious feature of this simple example is that a situation with a modest number of affordances actually presents the agent with considerable possibility regarding what to do. Here, with only three affordances, there are 34 possible ways to organize action over them, even under unrealistically simplifying assumptions. Adding a further affordance D would increase the complexity of the situation combinatorially.

The problem gets worse. Realistically, affordances and elements of repertoire do not map in a one-to-one fashion, so the combinatorial explosion compounds. Furthermore, under the more realistic view that affordances cannot be cleanly individuated, the picture explodes further, since repertoire will now be required to distinguish the saliences for some purposes but not for others. This implies further functions for individuating affordances in specific contexts. So, even in the simple case, the affordance landscape outpaces repertoire elements rapidly. To gloss the point, the more realistic the model of affordances, the more obvious the burden of superaffordances.

The upshot is that affordance landscapes have a depth to them; what is individuated as obvious is just the surface of any affordance landscape. Possible combinations of affordances are also affordances, and this generates the burden of superaffordances. Taking these considerations together, agents' affordance landscapes are far too rich to constitute anything the agent could act on. That is, the agent-relativity of the affordance landscape entails that affordances are superabundant for even very simple repertoires.

Under such conditions of superabundance the link between any given affordance and any given action is opaque. This is a problem because affordances are what frames the action of agents. If under very simple cases the combinatorial possibilities of affordances landscapes explode, then their action-guiding role seems impossible to fulfill. Put another way, the problem of superaffordances is the problem of how suitably integrated action, consisting of chains of

activities with complex dependencies, is possible if the affordances landscape is supersaturated in the way described above.

Here one might be tempted to appeal to background coping with the claim that the agent's capacity to do background coping is what keeps the superaffordances in check. This is undoubtedly correct. However, the salience landscape is supposed to be the starting point, or ground, of background coping. And if saliences are superabundant it becomes unclear how smooth background coping characteristic of natural agents gets off the ground. A vicious regress looms. Addressing the superabundance problem rescues the capacity of agents to engage in background coping.

However, unlike with the frame problem, the burden of superaffordances does have a solution. Or at the very least, the solution to the problem is immanent to the framework of natural agency. The way to address the superabundance problem is just more agency. Unlike with the frame problem, no new faculty needs to be introduced. This means that in the course of acting, agents generate affordances, just this time at a higher order.

Agents are in the same situation in the case of superaffordances as with affordances. The possibilities for action in the table are elements of the repertoire. So agents choose between elements of a higher-order repertoire.

In a slogan: just as agency entails the constitution of an affordance landscape, so a set of higher-order agential capacities entail the constitution of higher-order affordances. Sophisticated agents have the repertoire to make certain second-order affordances (as seen in the three right-side columns in the table above). Intuitively, this means that some of the huge set of higher-order affordances pop out as more live options. The agent's higher-order capacities manage the combinatorial explosion inherent in any and all situations.

The key difference between the frame problem and the burden of superaffordances is that the vicious combinatorial explosion common to both arises differently in the two contexts. In the frame problem, the combinatorial explosion is a function of the necessity of mediating between a Cartesian agent's capacities and its world. In the superabundance of salience problem, the combinatorial explosion is a function of the constitutive features of being an agent—namely of affordance-making. The different sources of the problem mean that avenues for addressing the

two are different. In the Cartesian case, appealing to higher-order capacities of the agent looks homuncular. In the case of natural agents, however, this is not the case. It is a condition of the possibility of action itself that agents' higher order capacities succeed in controlling the superabundance problem, at least enough of the time. It does this by making affordances the relations between affordances. This is not a homuncular explanation, since it simply appeals to the basic aspects of agency which is already natural enough.

The general strategy for addressing the burden of superaffordances is through second-order agency. Second-order agency is a domain-specific capacity to construct second-order affordances—that is, to pick out which affordances afford what.[184] Second-order agency is, to put it metaphorically, a filter applicable to a domain to reduce the burden of superaffordances and make action possible by allowing for smooth background coping against which particular actions are possible. By picking out the affordances that afford action, the burden of superaffordances is reduced because the superabundant affordance landscape acquires some contrast. With sufficient contrast amongst affordances, suitably integrated action can proceed.[185]

How does second-order agency construct salience over the affordance landscape? Consider a simple case where a bacterium is faced with two overlapping gradients. One is a nutrient gradient, the other is a toxin gradient. The two gradients attract and repulse the bacterium respectively. Now let us assume that the bacterium swims part of the way up the nutrient gradient and then at some critical point begins to oscillate around a point where the two gradients exert roughly the same, but opposite attraction and repulsion. The second-order salience-making

---

[184] The successful implementations of second-order agency are as plentiful as the multitude of adequate ways of life. That said work in cognitive science and systems biology that draws on the Free Energy Principle, predictive processing, and bioeconomics offers promising avenues for outlining the causal realizers of second-order salience-making. The Free Energy Principle is the view that complex biological and cognitive systems function to minimize the quantity of informational entropy inherent to their being systems in complex sensorimotor interaction with the world. This is a general principle that aims to unify accounts of perception, cognition, and action under one umbrella. It may also be a useful guide to implementation of agency, though it is not itself a theory of agency. Predictive processing is a more circumscribed application of the FEP to sensorimotor interaction in human cognition. The bioeconomic framework is concerned with the question of how, given finite resources and an inescapable energetic cost to every action, an agent is to dispose itself. In the present context, we can ask what sorts of "filtering" would best work with a given salience landscape. The filter is, of course, answerable to an agent's goals and repertoire, so the global norm on second-order agency is one of enabling more appropriate coping in situations. A general finding of the bioeconomic approach is that there probably is no universally applicable optimal strategy for modifying the salience landscape. So strategies are invariably local and context-dependent.

[185] The strategy is more properly one of higher-order agency. But for ease of exposition, second-order agency will be used. There is no limit to the orders of agency that can be implemented.

is the capacity that "decides" which of the gradients is more salient in a given situation. Both the gradients retain their attractive or repulsive valence, but the second-order salience is the additional feature that biases the bacterium's action tendencies towards approach or withdrawal. In the early phase of the example, second-order salience is aligned with the nutrient gradient's salience, and in the later phase of the example, the second-order salience is indifferent between the nutrient and toxin gradient. In the first phase, the second-order salience makes action possible. In the second phase, the utility of that particular second-order salience-making process has met its limits. The unintegrated action of the bacterium stuck between two equal but opposite gradients can be made more integrated by second-order salience construction of a different sort, or by going even more high-level.[186]

So, in sum, natural agents sidestep the Cartesian dead-end of the frame problem because the starting point of theorizing natural agency is different. Natural agents still face a superficially similar burden of superaffordances. This problem can be addressed in particular domains by the domain-specific operation of higher-order agency. Natural agents already have this capacity, constitutively. While the particulars of any optimization process will be complex and fraught with trade-offs, the basic coherence of a higher-order strategy does not face the insuperable obstacles to implementation that the frame problem does.

As such, a sophisticated agent's experience of its world is the result of multiple orders of salience-making applied to the affordances that exist in a situation.

3.6 Conclusion

This chapter has offered a conception of teleological explanation on which natural agents can be explained by two complementary and autonomous explanatory styles. Natural agents, as pieces of the causal structure of the world, are suitable for mechanistic explanation. But they are also suitable for teleological explanation. Teleological explanations enrich the naturalistic picture of the world by elucidating a set of modal invariances that exist between a means and its goal.

[186] Interesting empirical work on this and similar examples can be found in van Duijn (2017).

Agency is an ecological phenomenon that arises from the activities of goal-directed systems in their environmental contexts. Agents are characterized by the triad of repertoire, affordance, and goal. The activities of agents have a hypothetical invariance that in turn characterizes the norms to which such agents are subject.

Agents experience their environments as affordance landscapes. The connection between affordances and the appropriate activity of agents looks to be threatened by the burden of superaffordances. However, addressing this superaffordances problem requires only re-iteration (or re-iterations) of the basic agential motif of goal-directed biasing of repertoire. Sophisticated agents experience their worlds through multiple layers of biasing their repertoires in response to affordances.

4 From Ecological Agents to Cognitive Agents

The previous two chapters have argued that an ecological account of agency (EAA) is the best naturalistic approach to agency. On this view, agency is a gross behavioural phenomenon that is apparent when repertoire is marshalled to respond to affordances so as to attain goals. The EAA best deals with both the normative and teleological dimensions of agency. It also entails that agents are intimately situated since their environment are structures of affordances.[187]

The concern of this chapter is to extend the EAA to cognition. But here there is a problem. Cognition fits poorly with the EAA commitment to intimate situatedness because the workings of cognition seem to *desituate* agents from their immediate environment. Cognitive processes seem to distance the agent from the particulars of its situation. This affords acting on very abstract, removed, or otherwise distal affordances. Desituatedness marks the difference between the world-responsiveness of agents in general and the reason-responsiveness of cognitive agents.[188]

The task, then, is to make out a reasonable way in which the EAA can accommodate the difference between cognition and other forms of agency. Doing this would constitute a response to the intuition that cognition is, after all, best thought of in the Cartesian mode, as a phenomenon that is constituted by cognition's desituatedness. To underwrite an account of cognition, the EAA must give an account of the difference between world-responsiveness and reason-responsiveness.

The thesis of this chapter is that cognitive agency is the kind of agency that allows for desituatedness. Specifically, cognitive agency occurs on the mind-like end of the continuum between life-like and mind-like explanations. All the explanations along the continuum are

---

[187] The EAA is developed in Fulda (2016) and Walsh (2012, 2015, 2018).

[188] The world-responsiveness/reason-responsiveness distinction is important in controversies about the explanatory resources of situated (E4) cognitive science. Many argue that situated cognition cannot account for the reason-responsive aspects of cognitive life. This argument has been made in many different ways. For controversies in cognitive science, see Chemero (2009), Gomila et al (2012), Hutto & Myin (2013, 2017), Clark (2015), Steiner (2014), Salay (2016), Aizawa (2015), and Ramsey (2015), Schetz (2014), Roy (2015). For controversy about the nature of reason, see Schear (2013a, b), Dreyfus (2013), and McDowell (2013). Such discussion are ultimately orthogonal to the present project of offering an ecological take on the problem of reason-responsiveness, since they typically leave discussion of agency implicit.

teleological. The life-like end of the continuum explains the world-responsiveness of agents. The mind-like end of the continuum explains the reason-responsiveness of agents.

Reason-responsiveness is a very high-order analogue of a repertoire responding to affordances in order to attain goals. In reason-responsiveness, an agent taking the world to be a particular way is essential, and the sophistication of reason-responsiveness tracks the sophistication of capacities to take the world to be certain ways.

The burden of this chapter is to make reason-responsiveness a natural extension of the EAA. The argument will proceed as follows. Section 4.1 sketches the spectrum of life-like and mind-like explanation. Section 4.2 develops the difference formally into the difference between agential explanation and reason explanation. Section 4.3 develops the key role that the faculty of taking/making suppositions plays in reason explanations. Section 4.4 develops the two different failure modes of agential and reason explanation. Failing in agential explanation entails that the agent did something wrong. Failing in reason explanation allows for failing without doing something wrong. Section 4.5 contrasts reason explanation in this EAA mold with reason explanation as conceived of by both Davidson and Anscombe. Finally, section 4.6 discusses predictive processing as a plausible way of implementing the faculty of taking, and thereby mind-like explanation in general.

4.1 Life-Like and Mind-Like Explanation

The thesis of this section is that life-like and mind-like explanation are extremes on a spectrum of teleological explanations.

The ambit of natural agency ranges across vast differences of sophistication, complexity, suppleness, and adaptivity. The simplest agents are capable of plastically and robustly attaining enough of their goals. More complex agents enhance that basic capacity in many sorts of ways.[189]

---

[189] Complex agents typically have some combination of: (i) capacities to sustain long and adaptive chains of actions, (ii) the capacity to progress towards goals that are distal in time and space. They may also have meta-capacities that refine the regulatory capacities, such as (iii) adaptive memory storage and use, (iv) learning capacities, and (v) optimizing capacities. Development of these capacities is to some extent independent of the others. For simplicity, I lump all these into the more complex end of telic activity.

This vast ambit can be understood as scaling between two extremes. At the simple end of this range is life-like explanation, a form of teleological explanation that invokes repertoire, affordances, and goals and the agent's responsiveness to a world.

For example, explaining the actions of a free-living, sucrose-metabolizing bacterium in response to a sucrose gradient is a clear instance of life-like explanation. The bacterium marshals its repertoire in order to attain its goal of being in a sucrose-rich place, all else equal. The sucrose gradient and the bacterium's various capacities constitute the affordances in the situation. Even though its world is co-constituted, the bacterium's response is to a worldly state of affairs.

The key feature of life-like explanation is world-responsiveness. In world-responsiveness there is a comparably direct relationship between an affordance and the corresponding biasing of repertoire in response to it. Even sophisticated agents engage in world-responsiveness on occasion.

At the complex end of this range is mind-like explanation, a form of teleological explanation that invokes higher-order aspects of repertoire, affordances, and goals. Mind-like explanation emphasizes an agents' responsiveness to reasons, not its responsiveness to its world. Reasons, in this context, unpack into higher-order features of repertoire, affordances, and goals. Mind-like explanations advert to an agent *taking* the world to be such-and-such a way whereas life-like explanations invoke the world *being* in such-and-such a way. The phenomenon of an agent taking the world to be a particular way is neutral on how epistemically serious such a taking is. A number of attitudes can be held to a taking. The fact that taking depends on very high order affordances suggests that such affordances are invisible to immediate perception. However, invisibility to perception does not entail the very high order affordances are disconnected from lower-order, world-responsive processes. To the extent that faculties of taking are invoked is the extent to which mind-like explanations approach commonsense belief-desire psychology.

For example, explaining the actions of person who is pursuing an education adverts to very, very high-level goals, repertoire, and affordances. The goal is open-ended such that no particular state of the world amounts to being educated. Many actions make sense as responses to very, very high-order affordances. The explanation of the vast majority of the actions that conduce to getting an education are responses to reasons, not to the world. For example, choosing to take course X is conducive to getting an education though a long, convoluted telic chain. Whereas

world-responsiveness will depend on more or less immediate affordances, repertoire, and goals, reason-responsiveness is the management of very, very high-order affordances, repertoires, and goals, and the agent's navigation of sophisticated telic chains.

On this sketch of the spectrum, life-like and mind-like explanations do not have a sharp cutoff. As agents become more sophisticated, the likelihood of needing mind-like explanation increases. A considerable amount of the actions of sophisticated agents fall into a gray area between world- and reason-responsiveness.[190] The vast majority of agents might be usefully approached by the two kinds of attitudes in tandem, for different purposes and contexts. The view here is neutral on the question of where there may be a cutoff, or on even whether there is a cutoff. All that is needed for present purposes is for the two extremes to be intuitively quite different.

4.2 Agential and Reason Explanation

Both life-like and mind-like explanations are teleological. The thesis of this section is that mind-like explanation is a specific, unpacked form of teleological explanation that foregrounds the agent's reason-responsiveness. This contrasts with the generic form of teleological explanation, which foregrounds world-responsiveness.

As such, two forms of teleological explanation can be distinguished: agential explanation and reason explanation.[191] Reason explanation is a type of agential explanation.

Where X is an agent, Y is an action, and Z is a goal, agential explanation has the generic form:

(1) X did Y in order to Z

Making the teleological relations and the relationship of the agent to the world S more explicit[192], agential explanation unpacks to:

---

[190] Hanna & Maiese (2009) offer an unusually sophisticated articulation of the embodied agency that exists in this gray area between world- and reason-responsiveness.

[191] In this, reason explanation is not rationalizing explanation as discussed in section 1.1.1., although it is compatible with an Anscombian understanding of providing a reason for an action.

[192] To be clear, S cannot, on the EEA, be the relationship of an agent to a pre-existing world. It is more accurate to say that S is a way that the agent's affordance landscape is. S is an aspect of the totality of an agent's affordances in some situation, or set of situations.

(1') Z is a goal of X, and Y is conducive to attaining Z, given the state of the world S[193]

Teleological explanation is contrastive, meaning that (1) also expands to:

(1'') X did Y_1 instead of other actions in the contrast class $\{Y_2, Y_3,...Y_n\}$ in order to Z

In contrast, reason explanation has the generic form:

(2) Z was X's reason for doing Y

Which unpacks to:

(2') Z is the goal of X, and X took the world to be S, and Y would be conducive to attaining Z, were S true

The distinguishing feature of reason explanation is that the agent "takes" the world to be S. X responding to reasons is X responding to ways it takes the world to be S such that Y will be conducive to Z.

X taking the world to be S depends on an agent's higher-order repertoire (HOR) and its higher-order affordances (HOAs).

Let us suppose X is an agent that has the capacity to pursue a very abstract goal G that is pursuable across a wide range of circumstances. For example, for a person X, G might be "obtain an education".

Because G is so abstract and widely pursuable there is a number of ways of attaining G. These ways form a set of intermediate goals $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$ that constitute X's repertoire as it pertains to G. The set $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$ may be very, very large.[194] The affordances in this set are higher-order affordances (HOAs).

---

[193] Of course, agents err. Error has different senses in cases of world- and reason-responsiveness.

[194] As discussed in section 3.5 of this dissertation, there is nothing special to reason-responsiveness in a repertoire being very, very large. All agents whatsoever perpetually operate under the burden of superaffordances.

For example, where G is the goal of getting an education, we can take X's repertoire to consist of the various things a student could do: read the textbook (Ig_1), review notes (Ig_2), befriend instructor (Ig_3), and so on. The elements in this set are very, very high-order affordances.[195]

In order to act, X must be able to adopt an element from $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$. Teleological explanation of X's action is inherently contrastive, explaining why X does Ig_1 rather than $\{Ig_2, Ig_3,...Ig_n\}$.

Successful agents select from $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$ pursuant to what the world affords. Suppose that a state of the world, S_1, affords attaining G by doing Ig_1. (And S_2 affords attaining G by doing Ig_2, and so on.) Then HIg_1 is X's capacity to choose Ig_1 in response to S_1.

For example, the potential states of the world might be the textbook being very succinct (S_1), the notes being thorough (S_2), the instructor being generous (S_3), and so on.

X's capacity to select an element from the repertoire depends on X's higher-order repertoire (HOR). Like X's repertoire, HOR forms a very large set $\{HIg_1, HIg_2, HIg_3, ... HIg_n\}$. Whereas X's repertoire operates on *n*th-order affordances, X's higher-order repertoire operates on affordances of order *n+1*.

There is a crucial distinction between two capacities that X has: (1) the capacity of X to select Ig_1 because the world is S_1 and (2) the capacity to choose Ig_1 even if the world is not S_1. HOR is needed for explaining capacity (2); it is the capacity to "take" the world to be a variety of ways.

For example, HOR in the case of obtaining an education is the capacity to deploy X's repertoire regardless of worldly conditions. So, explaining why X is reading the textbook rather than reviewing notes is that X *takes* the world to be such that the textbook is very succinct. If the textbook is very succinct, then reading the textbook is conducive to getting an education.

In reason explanation, saying that X chose Ig_1 in order to bring about G is to say that X took the world to be S_1. If Ig_1 would be conducive to bringing about the agent's goal G if the world were

---

[195] Descriptions of HOAs typically take them as courses of action, schemas of action, or frameworks, rather than as simple opportunities for action. This is a pragmatic consequence of their being high-order affordances. Describing a multiply nested structure of affordances is linguistically cumbersome.

S_1, then the choosing Ig_1 is equivalent to taking the world to be S_1. The success conditions of S_1 are that the world affords attaining G by doing Ig_1. The content of HIg_1 is that the world is such that S_1 obtains. States of the higher-order repertoire have content determined by the success conditions of X's doing of any element of $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$.[196]

Reason-giving explanation is contrastive. It explains why Ig_1 happened, rather than the rest of the repertoire. If X takes the world to be S_1, and S_1 is the case then, ceteris paribus, X will succeed in bringing about G.

In this way, HOR is an approximation of belief-desire psychology in an ecological key. Roughly, X believes that P iff P is supposed by X's successful actions.[197] Elements of HOR have this structure, and for some agents at least they would carry content about states of the world required for some element of $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$ to afford G. Developing this line of thought in detail is beyond the scope of the present discussion, since connecting HOR to beliefs requires explication of a number of deep assumptions about the nature of content.[198]

Agential explanation and reason explanation are not that different. Both contrastively explain why some piece of X's repertoire was implemented rather than another. Both agential and reason explanation explain why X did Y, tacitly in contrast to other possible actions. Every successful case of reason explanation can be substituted by agential explanation. The difference between explaining Y by agential explanation and by reason explanation is that agential explanation emphasizes world-responsiveness and reason explanation emphasizes reason-responsiveness. In an agential explanation, the world bears the burden of X's action selection. In contrast, in reason explanation, X's taking the world bears the burden of action selection. X's taking the world to be a particular way is the reason X acted.

---

[196] In this way, the ecological account of agency connects to the attempt to account for the content of cognitive states through success semantics (Whyte 1990, Nanay 2013). Developing this connection is beyond the scope of this chapter.

[197] Here there is room for endless refinements. P must be part of the agent's co-constituted world, not a world that stands over and above the agent. If an agent lacks conceptual or behavioural capacities (individually or collectively) to entertain certain takings of the world, such things must be passed over. Of course, it remains open for another agent to see the impoverished agent's taking as inadequate in certain respects. But all agents are in some version of this predicament.

[198] Hutto & Myin (2013) give a good survey of the nature of content, concluding that content leads to a near-insuperable explanatory gap.

Recall from Chapter 3 that even the simplest agents carry a burden of superaffordances. For any goal, the repertoire that is potentially relevant must be winnowed down by some higher-order process. For some agents, the burden of superaffordances is carried by responsiveness to the world. For other agents, the burden is carried by a faculty for taking the world to be certain ways.

Agential explanation is pragmatically adequate for simpler systems which rely on the world to carry the burden. We can sideline the agent taking the world to be thus-and-so, instead we can typically assume the world to be a particular way. Applying reason explanation in these cases is unnecessary. The cost to the agent of determining the state of the world is close to negligible. What does the elucidating in the elucidative description is agent-world relations, namely affordances.

For example, a bacterium doing chemotaxis is well-explained by adverting to what the sucrose gradient affords the bacterium.[199]

Reason explanations are pragmatically adequate for more sophisticated cases of agency. For very, very high-order affordances, selecting from among them requires a capacity to proceed as if the world is a determinate way, even though determining how the world is might be impossible for the agent. Applying agential explanation in such sophisticated cases is unsatisfying because it makes implicit the faculty of taking the world to be a particular way which is pragmatically necessary.

For example, explaining why a person is reading a textbook by appealing to her goal of obtaining an education leans too heavily on the worldly facts that would make textbook-reading conducive to getting an education. The textbook-reading is better explained by the agent as taking the world to be particular ways, and acting from those presuppositions.

How can these intuitions about the applicability of agential and reason explanation be made more precise?

---

[199] That is, bacteria chemotax because sucrose gradients afford finding more sucrose. There is a complementary, but autonomous mechanistic explanation of this fact as well. See Barkai & Liebler (1997), Alon et al (1999), and Bi & Sourjik (2018).

Two features become increasingly important as one moves up the spectrum from life-like to mind-like agency: (i) the role of suppositions, and (ii) the attribution of failure. Simply put, suppositions about the world matter in reason explanation but not in agential explanation, and failure in agential explanation amounts to the agent doing something wrong, whereas in reason explanation there are two kinds of failure: an agent correctly responding to a reason (taking) where the taking is false, and an agent incorrectly responding to a taking, whether or not it is false. Reason explanation opens up failure modes and allows for an agent failing reasonably.

The next two section unpack these two features.

4.3 The Faculty of Taking

The thesis of this section is that higher-order repertoire—X's faculty of taking—is necessary in reason explanation. The faculty of taking is essential for agents to act under uncertainty, i.e., in conditions where determining the actual state of the world S is hard, costly, or impossible.[200]

Merely agential explanations typically do not advert to taking. Agential explanation for why X did Y presupposes that S, the state of the world that makes Y conducive to Z, is true. Because teleological explanation is contrastive, this unpacks to X did Y_1 instead of $\{Y_2, Y_3,...Y_n\}$. The explanation takes Y_1 to afford attaining Z in S. In more sketchy agential explanations, X's Y_1ing is a brute fact.

Reason explanations, in contrast, do not assume that S is true. They presuppose that X has the capacity to select elements the repertoire relevant to Z, $\{Y_1, Y_2, Y_3,...Y_n\}$ independently of how the world is. That is, reason explanations make explicit the role of higher-order repertoire, the process that selects among the elements of $\{Y_1, Y_2, Y_3,...Y_n\}$.

The faculty of taking introduces a new degree of freedom in the agent-environment system. For an agent X with a goal G, repertoire $\{Ig_1, Ig_2, Ig_3, ... Ig_n\}$ and higher-order repertoire $\{HIg_1,$

---

[200] The faculty of taking is what underwrites the capacity for representation. Not much more can be said on this point because the debate on the nature of the basic cognitive agent-world relation is extraordinarily rich. Hutto & Myin (2013, 2017) and Hutto & Satne (2015) favour the proto-intentional approach. Burge (2009, 2010) favours the proto-representational approach. Dretske (1981, 1986), Fodor (1994) Millikan (1989) take the more standard representationalist approach. Beyond these basics, many convolutions of the options are available in the philosophy of cognitive science. See Chemero (2009), Clark (2015), Steiner (2014), Aizawa (2015), and Ramsey (2015). The current discussion does not presuppose any one of these views.

HIg_2, HIg_3, ... HIg_n}, the faculty of taking is X's capacity to implement elements of {HIg_1, HIg_2, HIg_3, ... HIg_n}.

Choosing Hg_1 means X selects Ig_1. Doing Hg_1 means that X acts under the supposition that the world is S_1, and not any of the other ways the world might be. Under these circumstances X is responding to a taking. X's taking the world to be S_1 was its reason to do Ig_1.

So, in agential explanation we take X's action to respond to a state of the world S, whereas in reason explanation, we take X's action to be a response to X taking the world to be S. This capacity is a feature of higher-order repertoire. It is just the capacity to choose an element of the repertoire. The agent taking the world to be S is actually a presupposition of all teleological explanation, but in reason explanation it is explicit.

Mind-like explanations have the features they do because sophisticated agents typically act based on their faculty of taking. At some point of sophistication, explanation in such terms is indispensable. While all agents bear that burden of superaffordances, which means that all action is action under uncertainty, in mind-like explanation this is the dominant factor in making sense of agents' behaviour.

4.4 Agential and Reason Explanations have Different Failure Modes

The thesis of this section is that agential and reason explanations have different success conditions. Both explain failure cases, but they do so differently. Failure in agential explanation means the agent acted wrongly. Failure in reason explanation is more complex, allowing both for reasonable and unreasonable failure.

For agential explanations, suppose that X did Y in order to Z, but X failed to attain Z. On the agential explanation, this is because, ceteris paribus, the world was not such as to afford attaining Z. In this case, X did something wrong.[201]

[201] There is considerable subtlety hiding behind the apparent simplicity of world-responsiveness and its failures. Since on the EAA agent and world are co-constituting, the question of when an agent fails to be world-responsive is hard to articulate cleanly. This is an instance of trading the Cartesian problem of appropriate connection to the world for the problem of the disappearing environment. See Walsh (2015).

For reason explanations, suppose that Z was X's reason for doing Y, but X failed to attain Z, because the world was not S, i.e., how the agent took it to be. While the agent was wrong in taking the world to be S, it nevertheless responded to a reason appropriately given its supposition that the world was S, so in that sense X may have done nothing wrong from the reason-responsiveness perspective.

There are two distinct failure modes for reason explanation.

In the first, Z is X's reason for doing Y, but the presupposition that Y conduces to Z in S is false because the world is not S. In this case, X acted reasonably but doing Y failed to conduce to Z. X implemented the wrong repertoire given the world, but the right repertoire given the way X took the world to be.

In the second, X has a reason to do Y but it does Y* instead. This is a much more fundamental failure, seeing as it is a failure of reason-responsiveness.

Of course, the two failure modes can compound: X can fail by both being wrong about S and by failing to be reason-responsive. And for completeness, X may succeed by accident yet nevertheless fail to appropriately take the world to be, or it could succeed by accident yet do so because it failed to respond to its own reasons.

So, the failures of agential explanation point to the agent doing something wrong. Failures of reason explanation point to faulty presupposition on which reason-responsive processes ran appropriately. Agential explanation of failure cases points to failures of world-responsiveness. Reason explanation of failure cases points to failures of presupposition which is compatible with appropriate reason-responsiveness.

4.5 EAA Reason-Responsiveness: Anscombe and Davidson Revisited

The thesis of this section is that the EAA approach to reason-responsiveness integrates three key desiderata for action theory that have been typically kept apart. The EAA allows for integrating what is right about both causal and teleological approaches to action.

It has been assumed that the following three claims cannot be held together:

(1) Reasons are causes of action.

(2) Reasons teleologically explain action.

(3) Reasons are states of the agent.

Recall from Chapter 1 that Anscombe's (1957) and Davidson's (1963) accounts of reason-responsiveness treat (1)-(3) as an inconsistent set. To gloss the dispute, Davidson took reasons to be causes and states of the agent. Anscombe took reasons to teleologically explain agents' actions but because of that reasons could neither be causes not states of the agent. So Davidson would endorse (1) and (3), and Anscombe would endorse (2).

The view that (1)-(3) are inconsistent depends on assumptions that, given the arguments in this dissertation, are no longer viable.

If (1) is understood as reasons making a difference to agents, and teleological explanation is a naturalistic mode of explanation, and reason explanation is a type of teleological explanation, then reasons make a difference in the way that causes make a difference. Reason explanations, like all teleological explanations, explain by citing a set of invariances, specifically relations of conduciveness that are governed by hypothetical necessity. Both causal and teleological explanations cite change-relating invariances, although they cite change-relating invariances of quite different sorts.

Given a naturalistic account of teleology, and the view developed of reason explanation as a type of teleological explanation, then claim (2) follows immediately.

Finally, (3) has a complicated but plausible rendering on the EAA view. First, reasons are processes that can be implemented in agents on the model of complexity science and the organizational account of agency. More importantly, the agents who have these reason-responsive states are situated in their environments. Reasons are states of deeply situated systems, as such they are accessible as aspects of an agent's gross behaviour, in a way not very different from the way mere agency is a gross behavioural phenomenon. Reason-responsive behaviour is behaviour that is best explained by citing open-ended goals, which unpack to the biasing of very, very higher-order repertoire in response to very, very higher-order affordances.

Though convoluted in content, such explanations share the same form with first-order teleological explanations.

None of the above conceptual moves were available to Anscombe or Davidson. As such, the EAA can hold all of (1)-(3).

On the EAA, reasons are states of an agent that cause behaviour. The form of this causation relies on the causal architecture of complex systems with capacities for grade III downward determination. In other words, having a reason to act marshals an agent's repertoire in a way similar to what Davidson would take to be an internal cause of action. These causes are states of the agent. But these states explain teleologically, even if they are realized causally.

The EAA appears as an odd fit with holding reasons to be states of the agent because: (i) the situatedness of agents complicates the internal/external distinction, and (ii) the EAA specifies action and its explanation in terms of gross behaviour. Nevertheless, these are not deep problems.

For (i), on the EAA reasons are not *internal* states of an agent, but they are *covert* states of the agent. Reason explanation advert to non-immediate circumstances to a greater extent than agential explanation.

For (ii), while the content of a reason explanation does not advert to propositional attitudes, the agent's faculty of taking requires states whose content specifies what the world is like in a way that is indexed to an agent's affordances and repertoire. The content of such states can be specified by the success conditions of the agent's repertoire: if X does Y and Y actually conduces to Z, then the state of the HOR has the content that that Y affords Z in S. The EAA thus takes reasons to be situated, action-oriented analogues of the more familiar internal propositional attitudes.

On this view, taking is not a propositional attitude since the faculty of taking is fundamentally action-oriented rather than description-oriented. While taking is very similar to other sorts of states with intentional relations, its essential feature is as a high-level bias on the agent's high-level repertoire. Talk of the *content* of the state of taking the world to be thus-and-so is helpful shorthand.

So, the EAA view of reason-responsiveness allows for an understanding of reasons that captures aspects of how, for the very, very high-order affordances, repertoire, and goals, reason-responsiveness is different from world-responsiveness. Nevertheless, the "why?" in both cases, as Anscombe held, is the same question. Reason-responsiveness, and mind-like explanation in general, presupposes capacities by which an agent constitutes a more robust, world-independent locus of control. At the furthest reaches of this capacity, we have belief-desire psychology, which appears unrestrained in its capacity to take the world to be certain ways. Nevertheless, the difference between world-responsiveness and belief-desire psychology are only differences of degree.

4.6 Ecological Reason-Responsiveness and Predictive Processing

The thesis of this section is that the predictive processing (PP) framework in cognitive science offers a plausible architecture for how to realize the EAA account of reason-responsiveness.

To recap, the EEA view of reason-responsiveness is that Z is X's reason to Y iff Z is the goal of X, and X took the world to be S, and Y would be conducive to attaining Z, were S true. The difference between reason-responsiveness and world-responsiveness is that reason-responsiveness relies on X's faculty of taking the world to be S and to act accordingly, whether or not the world is S. This faculty of taking is essential for any agents to act at all, but for reason-responsive agents it is explanatorily indispensable. Reason-responsive agents thus attain considerable "freedom" from the exigencies of their situatedness, such that some of their actions cannot be explained any other way.

The predictive processing (PP) framework in cognitive science offers a plausible account of how the faculty of taking can be implemented in natural systems.

The PP framework is an emerging view on how to implement cognition. Its core claim is that, in acting, cognitive systems rely on an intricate and pervasive interplay of (i) internal models of the world and the (ii) signals produced by the cognitive agent's interaction with the world. The fundamental structural motif in PP is a module that compares a bottom-up and top-down

signal.[202] The top-down signal is a model of the state of the world, attempting to predict incoming sensory signal. If the model is accurate, the top-down and bottom-up signal from the world cancel out and the system carries on without needing to do anything. If there is a discrepancy in the model, the discrepancy is passed on for further processing, typically to a reiteration of the basic model at a higher level of abstraction (see figure 4).[203]

PP models use many layers of such comparisons between top-down and bottom-up signals, meaning that input is filtered through many different predictive models. The overall goal of the system constituted by many layers is to minimize prediction error, thereby keeping the system stable and functional. Prediction error can be operationalized under the concept of free energy. A cognitive agent is a complex amalgamation of predictive models at potentially many levels of abstraction as they encounter the world, whose goal is to minimize free energy.[204]

A caveat is necessary. Philosophical presuppositions of the PP framework are controversial. It has been taken to vindicate theories as disparate as neo-Cartesianism, representationalism, enactivism, and radical anti-representationalism.[205] Although enactive-ecological approaches to PP are arguable[206], here I use PP only for illustrative purposes. Ultimately, the EAA is neutral on questions of implementation.

The PP framework points to the necessity of taking the world to be a particular way. All agents must act under uncertainty, but in sophisticated agents the faculties that manage such uncertainty are explanatorily indispensable. The depth of the levels of predictive processing in a cognitive

---

[202] Slightly more technically, the fundamental organizational motif in PP is a Kalman filter, and cognitive agency necessitates a complex hierarchical arrangement of Kalman filters. See Grush (2003) for extensive discussion of this structural motif.

[203] The PP framework is discussed in depth in Friston (2009, 2010), Clark (2013, 2015, 2016), Hohwy (2016), Wiese & Metzinger (2017), Kiverstein (2018), Kiverstein et al (2019).

[204] Friston (2010) gives the classic treatment of the free energy principle (FEP) as it pertains to PP. Note well that while based in thermodynamics, the FEP is formulated in informational, not strictly physical terms, making it somewhat disanalogous with the thermodynamics discussed in Chapter 2.

[205] Grush (2003) uses a PP-like framework to argue for a version of Cartesianism. Hohwy (2016) argues against embedded cognition on the same basiss. Downey (2017) uses PP to argue for eliminativism about content. Allen & Friston (2016) use PP to argue for an enactive approach. Zahavi (2017) uses PP to point to a transcendental idealist reading of cognition. Kiverstein (2018) argues for an ecological-enactive interpretation of PP roughly along the lines of this dissertation.

[206] See Allen & Friston (2016), Kiverstein (2018), Kirchoff & Robertson (2018).

agent track, at least loosely, the explanatory indispensability of adverting to takings and presuppositions.

There is more to the faculty of taking than just the generative model. Generative models are an implementation of higher-order repertoire $\{HIg_1, HIg_2, HIg_3, \ldots HIg_n\}$. A generative model explains what it is for X to be able to take the world to be various ways, it does not yet explain what it is for X to be able to choose from $\{Ig_1, Ig_2, Ig_3, \ldots Ig_n\}$.

The connection between HOR and the affordances pertinent to a task depends on another feature of PP models, namely the ability of cognitive systems to take the generative model off-line. In these circumstances, instead of merely predicting the world to cancel out incoming signals, the generative model samples possible states of the world.[207] Attaching such off-line capacity to the system's repertoire accounts for how X might be able to select actions on the supposition that the world is a particular way.

This PP implementation can distinguish between two different failure modes of reason-responsiveness. In the reasonable case, an agent takes the world to be S and acts accordingly. The agent's failure resides in the execution after the reason was responded to, or the exigencies of the world, not in reason-responsiveness per se. In the worse failure mode, the agent takes the world to be S, but acts as if the world were S*. This amounts to an error in connecting between the generative model and the processing layers it is supposed to inform. The agent fails to execute reason-responsiveness. This failure mode amounts to a radical break with reality, and within the agent itself.

4.7 Conclusion

This chapter has presented an account of mind-like explanation that is an extension of the ecological account of agency (EAA). The EAA treats the agent as an entity that can marshal its repertoire to respond to affordances so as to attain its goals. The agent is the natural locus of teleological explanation. We explain the activities of agents by adverting to their marshaling of repertoire in world-responsive ways.

---

[207] The PP explanation of dream imagery is the operation of generative models in the absence of sensory input.

Specific natural agents exist on a spectrum ranging from relatively simple to very, very sophisticated. The features that determine position on this spectrum are multiple and controversial, but they roughly track the major transitions in evolution. Characterizing this intricate spectrum is beyond the scope of this discussion.

For simple agents, life-like explanation typically suffices to make their actions intelligible. Agential explanation takes the form: Z is a goal of X, and Y is conducive to attaining Z, given the state of the world S. Agential explanation emphasizes the world-responsiveness of the agent. The way simple agents manage the burden of their superaffordances is carried by the state of the world.

For sophisticated agents, mind-like explanation is typically required to make their actions intelligible. They require reason explanation, which is a sub-type of teleological explanation. It has the form: Z is the goal of X, and X took the world to be S, and Y would be conducive to attaining Z, were S true. In this case, Z is X's reason for doing Y. The burden of having to act in a situation where affordances are superabundant is carried in this case (in part) by the agent's faculty of taking. X's reason-responsiveness depends on its capacity to take the world to be S in a way that's independent of the state of the world.

An account of reason-responsiveness in terms of the faculty of taking allows for an ecologically embedded account of why reason-responsiveness seems desituated, in contrast to intimately situated world-responsiveness.

The extremes of the spectrum track the difference between two phenomenologically salient modes of agency: skilled coping and deliberative action.

Skilled coping is situated agency par excellence. Agential explanations are usually appropriate in cases of skillful coping, even in humans. In such cases the burden of superaffordances is carried by the world in concert with automatic or autonomous agential capacities. This mode of agency, and its dependence on life-like explanation suggests why there is a missing or attenuated subject-object dichotomy in such action. The desituated subject is simply not necessary, and subjectivity is expensive. The vast majority of mundane agency proceeds in this mode.

Reason explanations are needed for desituated agency, which takes many forms, ranging across a wide range of automaticity, salience, and language-like accessibility. Reason explanation gives us an ecological, teleological grip on situations where the agent seems to, at least some of the time, work by stepping back from its rich embedding in the world to survey the world, to try on different ways of taking the world, and to act from that capacity.

Conclusion

The main aim of this dissertation has been to argue that cognition is a mode of natural agency. Specifically, it is the desituated mode of natural agency. While many natural agents live by sophisticated responses to their situatedness, cognitive agents also have sophisticated capacities to respond to their taking the world to be particular ways. The faculty of taking is not, however, a capricious or arbitrary faculty. It is best conceived as a faculty that fills in noisy, gappy, or otherwise incomplete information about the state of the world. Cognitive agents must act under conditions of uncertainty. It may be uncertain what conditions in the world agents are responding to, and it may also be uncertain what the present conditions afford the agent. But in many instances, cognitive agents need to act.

The faculty of taking is backstopped by an evolutionary story. Cognitive agency is a successful mode of agency at least some of the time. Probably, on the evolutionary timescale, the most stable modes of agency are situated ones, since desituated agents require quite a lot of sophistication.

A natural question arises: where does the boundary between situated and desituated agency arise? On this question there is precious little consensus because there is little consensus about where cognition arose in evolutionary history. This lack of consensus is at least in part due to a fuzzy theoretical understanding of cognition itself.

The most likely scenario is that the capacity for desituated agency arose as multicellular creatures developed increasingly higher-order ways of regulating their boundary conditions.

But this is not the only option. Recent work on agential capacities in simple creatures suggests that there is much more sophistication lurking in them than typically assumed. This work is nascent, so drawing sweeping conclusions from it is premature, but it serves as an important counterweight to the intuition that cognition is something that is special to animals.

Conversely, a vast body of neuroscientific and psychological knowledge, as well as philosophical reflection on such knowledge has shaken the conviction that the life of human-level agents admits of a clean boundary between reason-responsiveness and other sorts of norm-

responsiveness. Even in our own case, the modes of agency are integrated more or less seamlessly.

In sum, the question of what exactly constitutes desituated agency is connected to other deep controversies about the nature of cognition that cannot themselves be resolved in the scope of this project. Nevertheless, viewing cognition as desituated agency can contribute to the collective project of getting clear about the nature of cognition. Situated and desituated agency are probably themselves complementary strategies, and the more we learn about them the more they might be recognized as interestingly different sides of the same coin. But for now, the difference in paradigmatic instances is clear enough.

To hold cognition as a form of agency takes a good deal of presupposition, as the discussion of the four heterodoxies has shown. We must stretch the received meanings of agency, cognition, evolution, and natural science. I hope I have shown in the first chapter that such revision is not arbitrary, but motivated by a deep consilience between the heterodox views I endorse.

Given the heterodoxies, this project sits between two more familiar views of agency: reductionism and primitivism. The discussion in chapter 2 has motivated the idea that reductionism (of a certain sophisticated sort) and primitivism (of a certain sophisticated sort) can be made complementary.

The organizational account of agency (OAA) is a sophisticated reductionist position. It is vastly superior to more mechanistic reductionist positions. It holds that ultimately the nature of agency can be given by appeal to complex, emergent but ultimately physical systems. The gross behavioural features of agency on this view reduce to elements of the causal, systemic, or thermodynamic profile of certain sorts of complex systems.

The ecological account of agency (EAA) is a sophisticated primitivist view. It is vastly superior to other primitivist views, especially those that equate agency with certain human-centric capacities such as reason-responsiveness. It takes the starting point of theorizing agency as the gross behavioural capacities of observable entities. In this way, the natural phenomenon of agency is independent of its realization.

This dialectical situation is far from settled. Both the apparatus of natural science and the indispensible, obvious features of our experience make reasonable epistemic claims on us. If possible, a theory of agency should be seamlessly integrated with natural science, hence the appeal of the OAA. But also, a theory of agency should do justice to the open-ended complexity of the norms that agents follow, hence the appeal of the EAA. Reasonable people can disagree on where the priority lies. But note that the dialectical situation between the OAA and EAA is a fruitful avenue of theoretical exploration, in ways that traditional arguments between reductionism and primitivism have not been. This is not the argument between determinism and free will. It is an argument about methodology.

The picture of the world suggested here is one that is rich in teleology. But this teleology is not occult or anti-naturalistic. It is a teleology grounded in the perfectly ordinary activity of living things. It is at once stripped of grand, cosmic significance, and imparts an intrinsic significance to many of the activities of the natural world. It is a teleology that reverses the cultural trend of disenchanting the world. It supports a liberal naturalism that is capacious, but not excessive in what it considers natural.

Finally, the arguments in this dissertation suggest a way of saving the representationalist baby from the Cartesian bathwater. The fundamental insight that cognition is desituated agency suggests that cognition is characterized by a sort of openness to the world that requires a split between the subject and object. The cognitive agent faces a complex world and must fill the gap with its own effort—its own taking the world to be determinate ways in order to act. But on the view developed here, this split is only a functional specialization of an integrated whole—the agent—that is not characterized in this way.

Agency requires a sort of openness to the world. Most modes of agency are modes of absorbed coping. Cognition is a special kind of openness. Very roughly, it is an openness that requires near constant active filling-in on the part of the agent. Of course, such filling-in is typically not arbitrary.[208] Other modes of agency within the same agent constrain the faculty of taking.

---

[208] There are probably modes of cognitive agency that approximate a sort of pure, capricious filling-in. Fantasy, hallucination, and dreaming are possible instances of this.

This project has also sketched an avenue for an ecological interpretation of the predictive processing framework. It also suggests that artificial intelligence must pay attention to the more situated modes of agency that constrain the focal, desituated, aim of artificial intelligence. If the view sketched here is right, artificial intelligence will require the production of artificial agents. Artificial intelligence may collapse the distinction between agent and artifact. But probably such "agentified" artifacts will inherit all the features of agency that make natural agency at once perfectly ordinary and deeply significant.

References

Aizawa, K. (2015). What is this cognition that is supposed to be embodied? *Philosophical Psychology*, 28(6), 755-775.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 195(6), 2459-2482.

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1997). Robustness in Bacterial Chemotaxis. *Nature* 397: 168-171.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial intelligence*, 149(1), 91-130.

Anscombe, G. E. M. (1957). *Intention*.

Ashby, R. (1956). *An introduction to cybernetics*. London: Chapman & Hall.

Ayala, F. (1970). Teleological Explanations in Evolutionary Biology. *Philosophy of Science* 37(1): 1-15.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367-386.

Barandiaran, X. E., & Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life*, 20(1), 5-28.

Barandiaran, X., & Moreno, A. (2008). Adaptivity: From metabolism to behavior. *Adaptive Behavior*, 16(5), 325-344.

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends in cognitive sciences*, 22(1), 8-20.

Barkai, N., and Leibler, S. (1999). Robustness in Simple Biochemical Networks. *Nature* 387: 913-917.

Bechtel, W. (2007). Biological Mechanisms: Organized to Maintain Autonomy. In Boogerd, Bruuggerman, Hofmeyr and Westethoff (Eds.), *Systems Biology: Philosophical Foundations*. Amsterdam: Elsevier. pp. 269-302.

Bechtel, W. and Abrahamsen, A. (2011). "Complex Biological Mechanisms: Cyclic, Oscillatory and Autonomous." In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 257-286.

Bedau, M. (1991). Can Biological Teleology Be Naturalized? *Journal of Philosophy* 88:647-55.

Bedau, M. (1992). Where's the Good in Teleology? *Philosophy and Phenomenological Research* 52: 781-805.

Bedau, M. (2002). Downward causation and the autonomy of weak emergence. *Principia: an international journal of epistemology*, 6(1), 5-50.

Bennett, M. R., Dennett, D., Dennett, D. C., Hacker, P., & Searle, J. (2007). *Neuroscience and philosophy: Brain, mind, and language*. Columbia University Press.

Bennett, M. R., & Hacker, P. M. S. (2008). *History of cognitive neuroscience*. New York: Wiley-Blackwell.

Bernard, C. (1865). *Introduction à l'étude de la médecine expérimentale*. Paris: Baillière.

Bernard, C. (1878). *Leçons sur les phénomènes de la vie communs aux animaux et aux végétaux*. Paris: Baillière.

Bi, S., & Sourjik, V. (2018). Stimulus sensing and signal processing in bacterial chemotaxis. *Current opinion in microbiology*, 45, 22-29.

Bickhard, M. (2011). "Systems and Process Metaphysics." In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 91-104.

Bishop, R. (2011). "Metaphysical and Epistemological Issues in Complex Systems" - In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 105-136.

Boden, M. A. (1999). Is metabolism necessary?. *The British Journal for the Philosophy of Science*, 50(2), 231-248.

Boden, M. A. (2000). Autopoiesis and life. *Cognitive Science Quarterly*, 1(1), 115-143.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3), 139-159.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new AI*. MIT press.

Broome, J. 1999. Normative Requirements. *Ratio* 12:398–419.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, 79(2), 251–278.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, 9(3), 399–431.

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1), 25-36.

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, 71(5), 805-819.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological psychology*, 15(2), 181-195.

Chemero, A. (2009). *Radical embodied cognitive science.* MIT press.

Christensen, W. (2012). Natural sources of normativity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 104-112.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33, 269-298.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204.

Clark, A. (2015). Predicting peace: The end of the representation wars. *Open MIND*. Frankfurt am Main: MIND Group.

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, 53, 3-27.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Davidson, D. (2001). *Essays on actions and events: Philosophical essays*. Oxford University Press on Demand.

Dawson, M. R. (2013). *Mind, body, world: Foundations of cognitive science*. Athabasca University Press.

Dennett, D. C. (1987). Cognitive Wheels: the Frame Problem in Al. *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the cognitive sciences*, 4(4), 429-452.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal.* Oxford University Press.

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*, 195(12), 5115-5139.

Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (1972). *What computers can't do*. Harper & Row.

Dreyfus, H. L. (1994). *What computers still can't do: A critique of artificial reason*. MIT press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical psychology*, 20(2), 247-268.

Dreyfus, H. L. (2013). The Myth of the Pervasiveness of the Mental. In Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

England, J. L. (2015). Dissipative adaptation in driven self-assembly. *Nature nanotechnology*, 10(11), 919-923.

Fodor, J. A. (1975). *The language of thought*. Harvard university press.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, 3, 63-109.

Fodor, J. A. (1981). *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. A Bradford Book.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3-71.

Fodor, J. A., & Pylyshyn, Z. W. (2002). How direct is visual perception?: Some reflections on Gibson's "Ecological Approach". *Vision and Mind: Selected Writings in the Philosophy of Perception*, 167-228.

Frankfurt, H. G. (1988/1978). *The importance of what we care about: Philosophical essays*. Cambridge University Press.

Friston, K. J. (2000). The labile brain. I. Neuronal transients and nonlinear coupling. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1394), 215-236.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, 13(7), 293-301.

Friston, K. J. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, 11(2), 127-138.

Froese, T. (2010). Life after Ashby: ultrastability and the autopoietic foundations of biological autonomy. *Cybernetics & Human Knowing*, 17(4), 7-49.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4), 466-500.

Fulda, F. C. (2016). *Natural Agency: An Ecological Approach* (Doctoral dissertation).

Fulda, F. C. (2017). Natural agency: the case of bacterial cognition. *Journal of the American Philosophical Association*, 3(1), 69-90.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.

Gallagher, S. (2019). Precis: Enactivist interventions. *Philosophical Studies*, 176(3), 803-806.

Ghiselin, M. T. (1994). Darwin's language may seem teleological, but his thinking is another matter. *Biology and philosophy*, 9(4), 489-492.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Gomila, A., Travieso, D., & Lobo, L. (2012). Wherein is human cognition systematic?. *Minds and Machines*, 22(2), 101-115.

Grush, R. (2003). In defense of some “Cartesian” assumptions concerning the brain and its operation. *Biology and Philosophy*, 18(1), 53-93.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4), 814.

Hall, N. (2004). Two concepts of causation. *Causation and counterfactuals*, 225-276.

Hanna, R., & Maiese, M. (2009). *Embodied minds in action*. Oxford University Press.

Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, 4, 383-427.

Haugeland, J. (2002). “Authentic Intentionality” In In M. Scheutz (ed.), *Computationalism: New Directions*. MIT Press.

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchoff, M., and Friston, K. (2019). A Multi-Scale View of the Emergent Complexity of Life: A Free-Energy Proposal. In G. Y. Georgiev et al. (eds.), *Evolution, Development and Complexity*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-030-00075-2_7, pp. 195-228.

Hofkichner, W., and Schafranek, M. (2011). "General System Theory". In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 177-194.

Hofmeyr, J.-H. (2007). The Biochemical Factory that Autonomously Fabricates itself: A Systems Biological View of the Living Cell. In Boogerd, Bruuggerman, Hofmeyr and Westethoff (Eds.), *Systems Biology: Philosophical Foundations*. Amsterdam: Elsevier. pp. 217-242.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259-285.

Hooker, C. (2011). "Introduction to Philosophy of Complex Systems". In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 3-90.

Hull, D. (1969). Essay review: What philosophy of biology is Not. *Journal of the History of Biology*, 241-268.

Huneman, P. (2016). Does Emergence Also Belong to the Scientific Image? Elements of an Alternative Theoretical Framework Towards an Objective Notion of Emergence. In *Epistemology, Knowledge and the Impact of Interaction* (pp. 485-505). Springer, Cham.

Humphreys, P. (2017). "Emergence". In *Oxford Handbook of Philosophy of Science*. Oxford University Press.

Hutto, D. D. & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT Press.

Hutto, D. D., & Myin, E. (2019). Deflating deflationism about mental representation. *What are mental representations.*

Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia*, 43(3), 521-536.

Izquierdo, E., Harvey, I., & Beer, R. D. (2008). Associative learning on a continuum in evolved dynamical neural networks. *Adaptive Behavior*, *16*(6), 361-384.

Jablonka, E., & Lamb, M. J. (2008). Soft inheritance: challenging the modern synthesis. *Genetics and Molecular Biology*, 31(2), 389-395.

Jacob, P., "Intentionality", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL=<https://plato.stanford.edu/archives/win2019/entries/intentionality/>.

Jonas, H. (1953). A critique of cybernetics. *Social Research*, 20, 172–192.

Jonas, H. (1966). *The phenomenon of life: Toward a philosophy of biology*. Northwestern University Press.

Jonas, H. (1968). Biological foundations of individuality. *International Philosophical Quarterly*, 8(2), 231-251.

Jones, D. M. (2016). *The biological foundations of action*. Routledge.

Juarrero, A. (1999). *Dynamics in action*. MIT Press.

Kant, I. (2000). *Critique of the power of judgment* (P. Guyer, & E. Matthew, trans.). Cambridge: Cambridge University Press. (First published 1793)

Kauffman, S. A. (1996). *At home in the universe: The search for the laws of self-organization and complexity*. Oxford university press.

Kauffman, S. A. (2000). *Investigations*. Oxford University Press.

Kim, J. (1989): “Mechanism, Purpose, and Explanatory Exclusion.” As reprinted in J. Kim (Ed.), *Supervenience and Mind*. New York: Cambridge University Press, 1993, pp. 237-264.

Kim, J. (1999). Making sense of emergence. *Philosophical studies*, 95(1), 3-36.

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese*, 151(3), 547-559.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical Explorations*, 21(2), 264-281.

Kitano, H. (2004). Biological Robustness. *Nature Reviews Genetics* 5: 826-837.

Kiverstein, J. (2018). Free energy and the self: an ecological–enactive interpretation. *Topoi*, 1-16.

Kiverstein, J., Miller, M., & Rietveld, E. (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese*, 196(7), 2847-2869.

Laland, K. N.; Odling-Smee, J.; Feldman, M. and Kendal, J. (2009). Conceptual Barriers to Progress within Evolutionary Biology. *Foundations of Science* 14: 195-216.

Laland, K. N.; Odling-Smee, J.; Hoppitt, W., and Uller, T. (2012). More on How and Why: Cause and Effect in Biology Revisited. *Biology and Philosophy* DOI: 10.1007/s10539-012-9335-1.

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., ... & Odling-Smee, J. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences*, 282(1813), 20151019.

Laland, K., Uller, T., Feldman, M., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, A., Oding-Smee, J., Wray, G. A., Hoekstra, H. E., Futuyma, D. J., Lenski, R. E., MacKay, T. F. C., Schluter, D., and Strassman, J. E. (2014). Does evolutionary theory need a rethink?. *Nature News*, 514(7521), 161.

Lansing, S., and Downey, S. (2011). "Complexity and Anthropology". In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 569-602.

Lewens, T. (2012). A Surfeit of Naturalism. *Metaphilosophy*, 43(1-2), 46-57.

Lewontin, R. C. (2001). *The triple helix: Gene, organism, and environment*. Harvard University Press.

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, 9, 2228.

Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing*, 7(1), 11-29.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 1-25.

Matthews, G. (2002). *The Philosophy of Merleau-Ponty*. London: Routledge.

Maturana, H. & Varela, F.H. (1980). *Autopoiesis and Cognition: The Realization of the Living.* Boston Studies in the Philosophy of Science, vol. 42. Dordrecht: D. Reidel.

Mayr, E. (1998). The multiple meanings of 'teleological'. *History and philosophy of the life sciences*, 20(1), 35-40.

Mayr, E. (1993). Proximate and ultimate causations. *Biology and Philosophy*, 8(1), 93-94.

McDowell (2013). “The Myth of the Mind as Detached.” In Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

Mitchell, M. (2009). *Complexity: A guided tour.* Oxford University Press.

Montévil, M., & Mossio, M. (2015). Closure of constraints in biological organisation. *Journal of theoretical biology*, 372, 179-191.

Mossio, M., & Bich, L. (2017). What makes biological organisation teleological?. *Synthese*, 194(4), 1089-1114.

Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer.

Moreno, A., Ruiz-Mirazo, K., and Barandiaran, X. (2011). "The Impact of the Paradigm of Complexity on Foundational Issues in Biology and Cognitive Science." In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 311-334.\

Morrison, M. (2015). “Why Is More Different?”. In Falkenburg, B., & Morrison, M. (2015). *Why More is Different. Philosophical Issues in Condensed Matter Physics and Complex Systems.* Springer, Berlin, Heidelberg.

Nagel, E. (1961). The structure of science: Problems in the logic of scientific explanation. *Journal of Philosophy* 37(142): 372-374

Nagel, E. (1977). Goal-Directed Processes in Biology. *The Journal of Philosophy* 74(5): 261-279.

Nanay, B. (2013). Success semantics: The sequel. *Philosophical Studies,* 165(1), 151-165.

Newman, S. (2011). "Complexity in Organismal Evolution." In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 335-354.

Nicholson, D. (2013). Organisms ≠ Machines. *Studies in the History and Philosophy of Biological and Biomedical Sciences* 44: 669–678.

Nicholson. D. J. (2014). The Machine Conception of the Organism in Development and Evolution: A Critical Analysis. *Studies in History and Philosophy of Biological and Biomedical Sciences* 48: 162–74.

Nicholson, D. J. (2018). Reconceptualizing the organism: From complex machine to flowing stream. In Nicholson & Dupre (Eds.) *Everything Flows: Towards a Processual Philosophy of Biology.* Oxford University Press.

Nicolis, G., & Prigogine, I. (1977). Self-organization in nonequilibrium systems (Vol. 191977). Wiley, New York.

O'Connor, T. and Wong, H. Y., "Emergent Properties", *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2015/entries/properties-emergent/>.

O'Shaughnessy, B. (2008). *The Will: A Dual Aspect Theory* (2nd ed.), Cambridge University Press.

Odenbaugh, J. (2011). "Complex Ecological Systems". In Hooker, Cliff (Ed.), *The Philosophy of Complex Systems*. Elsevier, pp. 421-440.

Odling-Smee, F. J., Laland, K. N., Feldman, M. W., & Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton University Press.

Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.

Oyama, S. (1985). *The ontogeny of information: Developmental systems and evolution*. Durham, NC: Duke University Press.

Oyama, S., Griffiths, P., & Gray, R. D. (Eds.). (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge, MA: MIT Press.

Ozawa, H., Ohmura, A., Lorenz, R. D., & Pujol, T. (2003). The second law of thermodynamics and the global climate system: A review of the maximum entropy production principle. *Reviews of Geophysics*, 41(4).

Phattanasri, P., Chiel, H. J., & Beer, R. D. (2007). The dynamics of associative learning in evolved model circuits. *Adaptive behavior*, 15(4), 377-396.

Pigliucci, M. (2009). An extended synthesis for evolutionary biology. *Annals of the New York Academy of Sciences*, 1168(1), 218-228.

Pigliucci, M., & Müller, G. B. (2010). Elements of an extended evolutionary synthesis. *Evolution: The extended synthesis*, 3-17.

Prigogine, I., & Nicolis, G. (1971). Biological order, structure and instabilities. *Quarterly Reviews of Biophysics*, 4(2-3), 107-148.

Prigogine, I. (1980). *From Being to Becoming: Time and Complexity in the Physical Sciences*. San Francisco: W.H. Freeman and Company.

Ramsey, W. (2017). Must cognition be representational? *Synthese*, 194(11), 4197-4214.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, 26(4), 325-352.

Rosen, R. (1991). *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life.* Columbia University Press.

Rosslenbroich, B. (2014). *On the origin of autonomy: a new look at the major transitions in evolution*. Springer Science & Business Media.

Roy, J. M. (2015). Anti-Cartesianism and Anti-Brentanism: The Problem of Anti-Representationalist Intentionalism. *The Southern Journal of Philosophy*, 53, 90-125.

Ruiz-Mirazo, K. J., et al. (2000). Organisms and Their Place in Biology. *Theory in Biosciences* 119: 209–233.

Ruiz-Mirazo, K. J., and A. Moreno (2004). A Universal Definition of Life. *Origins of Life and Evolution of the Biosphere* 34: 323–346.

Ruiz-Mirazo, K. J., and A. Moreno (2012). Autonomy in Evolution: From Minimal to Complex Life. *Synthese* 185: 21–52.

Ruse, M. (1972). Biological adaptation. *Philosophy of Science*, 525-528.

Salay, N. (2016). Representation: Problems and Solutions. In CogSci.

Sarkar, S., & Gilbert, S. (2000). Embracing complexity: Organicism for the 21st century. *Developmental Dynamics*, 219, 1–9.

Schetz, A. (2014). The minimalist representationalism of the ecological theory of perception. *Roczniki Psychologiczne*, 17(1), 217-235.

Schrodinger, E. (1992). *What is Life?.* 1944. reprinted Cambridge University Press, 1995, 87.

Sebastián, M. Á. (2017). Functions and mental representation: the theoretical role of representations and its real nature. *Phenomenology and the Cognitive Sciences,* 16(2), 317-336.

Setiya, K., "Intention", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/intention/>.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2, 395.

Schaffner, K. F. (2006). Reduction: the Cheshire cat problem and a return to roots. *Synthese*, 151(3), 377-402.

Schear, J. K. (Ed.). (2013). *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

Schear (2013a). “Introduction.” In Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

Schear (2013b). “Are We Essentially Rational Animals?” In Schear (Ed.) *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

Shanahan, M., "The Frame Problem", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>.

Skewes, J. C., & Hooker, C. A. (2009). Bio-agency and the problem of action. *Biology & Philosophy*, 24(3), 283-300.

Smith, M. (2012). Four objections to the standard story of action (and four replies). *Philosophical Issues*, 22(1), 387-401.

Sommerhoff, G. (1950). *Systems Biology*. Cambridge: Cambridge University Press.

Steiner, P. (2014). Enacting anti-representationalism: The scope and the limits of enactive critiques of representationalism. *AVANT. Pismo Awangardy Filozoficzno-Naukowej*, (2), 43-86.

Steiner, P. (2019). Radical views on cognition and the dynamics of scientific change. *Synthese*, 1-23.

Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

Stewart, J. R., Gapenne, O., & Di Paolo, E. A. (Eds.). (2010). *Enaction: Toward a new paradigm for cognitive science*. MIT Press.

Stoffregen, T. (2003). Affordances as Properties of the Animal-Environment System. *Ecological Psychology* 15: 115–34.

Stoljar, D. (2017). "Physicalism", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/physicalism/>.

Strevens, M. (2017). “Ontology, Complexity, and Compositionality”. in *Essays on Metaphysics and the Philosophy of Science*, M. Slater and Z. Yudell (eds.), Oxford University Press, Oxford.

Szathmáry, E., and Smith, J. M. (1995). *The major transitions in evolution*. Oxford, UK: WH Freeman Spektrum.

Taylor, C. (1964). *Explanation of Behaviour*. Routledge and Kegan Paul.

Thompson, E. (2007). *Mind in life*. Harvard University Press.

Thompson, E., & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, 28(1), 23-30.

Uller, T., & Helanterä, H. (2019). Niche construction and conceptual change in evolutionary biology. *The British Journal for the Philosophy of Science*, 70(2), 351-375.

van Duijn, M. (2017). Phylogenetic origins of biological cognition: convergent patterns in the early evolution of learning. *Interface Focus*, 7(3), 20160158.

van Riel, R. and Van Gulick, R., "Scientific Reduction", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction/>.

Varela, F. J. (1979). *Principles of Biological Autonomy*. New York: Elsevier North Holland.

Varela, F., H. Maturana and R. Uribe (1974). Autopoiesis: The Organization of Living Systems, Its Characterization and a Model. *Biosystems* 5: 187–196.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* MIT Press.

Velleman, D. (2010). There are no 'reasons for acting'. Unpublished manuscript.

Verdejo, V. M. (2015). The systematicity challenge to anti-representational dynamicism. *Synthese*, 192(3), 701-722.

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2009). Relevance realization and the emerging framework in cognitive science. *Journal of Logic and Computation*, 22(1), 79-99.

Von Bertalanffy, L. (1969). *General systems theory*. New York: George Barziller.

Wallace, R. J., "Practical Reason", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/spr2020/entries/practical-reason/>.

Walsh, D. M. (2006). Organisms as natural purposes: The contemporary evolutionary perspective. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 37(4), 771-791.

Walsh, D. M. (2008). "Teleology," in Michael Ruse (ed.), *The Oxford Handbook of Philosophy of Biology*. Oxford University Press 113-137.

Walsh, D. M. (2012a). Mechanism and purpose: A case for natural teleology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 173-181.

Walsh, D. M. (2012b). Situated adaptationism. *The environment: Philosophy, science, and ethics*, 89-116.

Walsh, D. M. (2013). Mechanism, emergence, and miscibility: the autonomy of Evo-Devo. In *Functions: Selection and mechanisms* (pp. 43-65). Springer, Dordrecht.

Walsh, D. M. (2015). *Organisms, agency, and evolution*. Cambridge University Press.

Walsh, D. M. (2018). "Objectcy and Agency: Towards a Methodological Vitalism." In *Everything Flows: Towards a Processual Philosophy of Biology*, edited by Daniel Nicholson and John Dupre. Oxford University Press.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the cognitive sciences*, 1(2), 97-125.

Weichold, M. (2018). Situated agency: towards an affordance-based, sensorimotor theory of action. *Phenomenology and the Cognitive Sciences*, 17(4), 761-785.

West-Eberhard, M. J. (2003). *Developmental plasticity and evolution*. Oxford University Press.

Wheeler, M. (1997). Cognition's coming home: The reunion of life and mind. In *Proceedings of the fourth European conference on artificial life* (pp. 10-19).

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. MIT press.

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies*, 16(3), 323-349.

Wheeler, M. (2011). Mind in life or life in mind? Making sense of deep continuity. *Journal of Consciousness Studies*, 18(5-6), 148-168.

Wheeler, M. (2016). "The Rest is Science: What Does Phenomenology Tell Us About Cognition?". In *Phenomenology and Science* (pp. 87-101). Palgrave Macmillan, New York.

Whitacre, J. (2012). Biological Robustness: Paradigms, Mechanisms and Systems Principles. *Frontiers in Genetics* 3. Article 67.

Whyte, J. T. (1990). Success semantics. *Analysis*, 50(3), 149-157.

Wiener, N. (1948). *Cybernetics, or Control and Communication in the Animal and the Machine.* Technology Press.

Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy*, 24(sup1), 207-274.

Winther, R. G. (2009). Schaffner's model of theory reduction: Critique and reconstruction. *Philosophy of Science*, 76(2), 119-142.

Wiseman, R. (2016). Routledge Philosophy Guidebook to Anscombe's Intention. Routledge.

Withagen, R., Araújo, D., & de Poel, H. J. (2017). Inviting affordances and agency. *New Ideas in Psychology*, 45, 11-18.

Woodward, J. (2002). What Is a Mechanism?: A Counterfactual Account. *Philosophy of Science* 69: S366-S377.

Woodward, J. (2011). Mechanisms revisited. *Synthese*, 183(3), 409-427.

Woodward, J. (2013). Mechanistic explanation: Its scope and limits. In *Aristotelian Society Supplementary Volume* (Vol. 87, No. 1, pp. 39-65). Oxford, UK: Blackwell Publishing Ltd.

Woodward, J. (2014). Scientific Explanation. *Stanford Encyclopedia of Philosophy*. (78 pp.)

Wu, W. (2011). Confronting Many-Many problems: Attention and agentive control. *Noûs*, 45(1), 50-76.

Zahavi, D. (2018). Brain, Mind, World: Predictive coding, neo-Kantianism, and transcendental idealism. *Husserl Studies*, 34(1), 47-61.

Figures

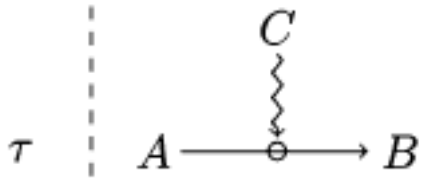

Figure 1. Constraint. A pictorial depiction of a causal process (getting to B from A) happening under a constraint C at timescale τ. C remains unchanged at the timescale τ. C enables the process of getting from A to B without itself being affected by the process. The wiggly line indicates a relationship of constraint (both limiting and enabling) (From Moreno & Mossio 2015, credit to Mael Montavil).

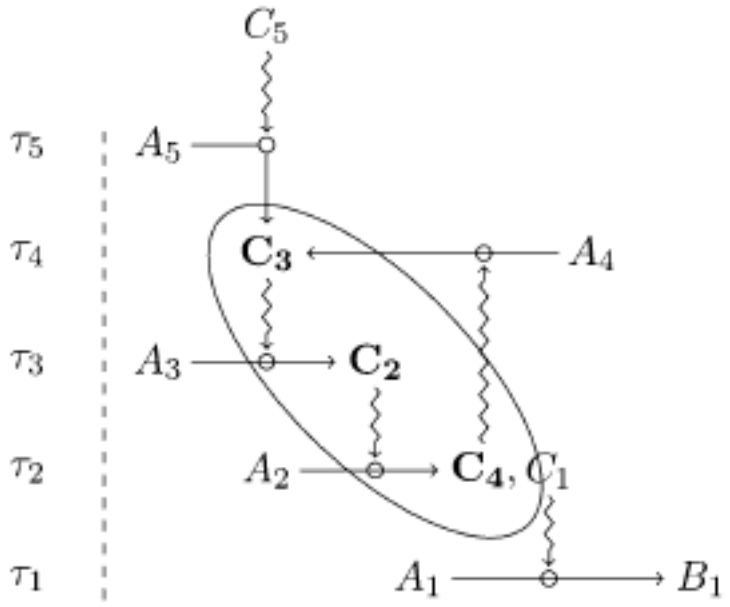

Figure 2. Closure of constraints. Straight arrows represent causal processes. C1, C2, C3, C4, and C5 are constraints, realizing conservation at the timescales on which they enable their associated processes, represented by wiggly arrows. τ1- τ5 represent five different timescales, arranged from fastest (τ1) to slowest (τ5). C2, C3 and C4 form a closed set of constraints, realizing closure of constraints whereby each of them depends on at least one other constraint in the closure and generates at least one other constraint in the closure. The closure of constraints operates at three different timescales. It also involves at least one instance of slow dependence, where the timescale of the constraint's stability is longer than the timescale of the associated process (e.g., for C2), and one instance of fast dependence, where the timescale of the constraint's stability is faster than that of the associated process (for C4). On this scheme there is no need for synchronic downward determination. (From Moreno & Mossio 2015, credit to Mael Montavil)

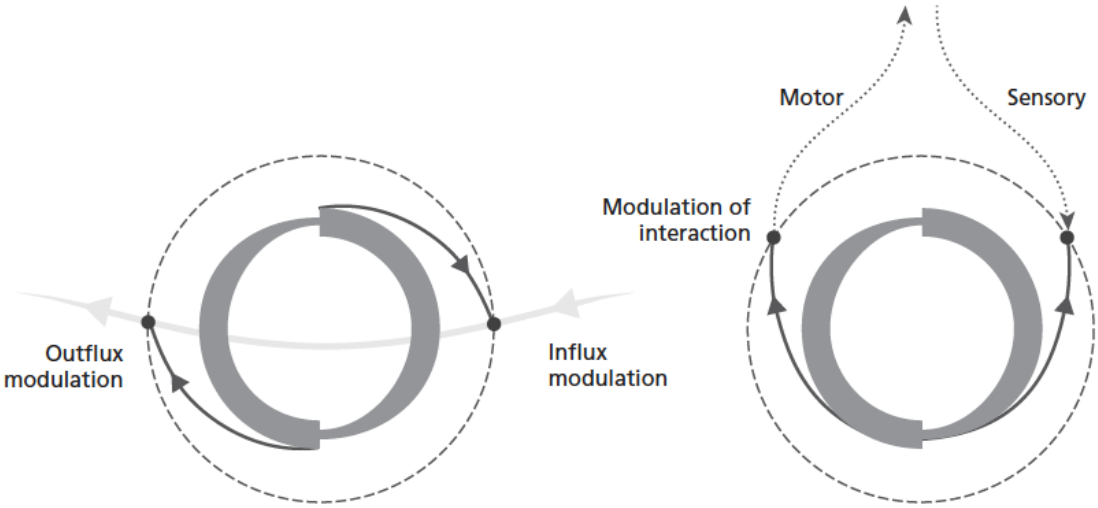

Figure 3. Constitutive constraints and interactive constraints. In this diagram, a self-constituting system (grey circle) modulates the influx and efflux of matter and energy across its boundary conditions on the left. Some such systems also have the capacity to modulate their boundary conditions by sensorimotor engagement with the environment. (Di Paolo et al 2017, 119).

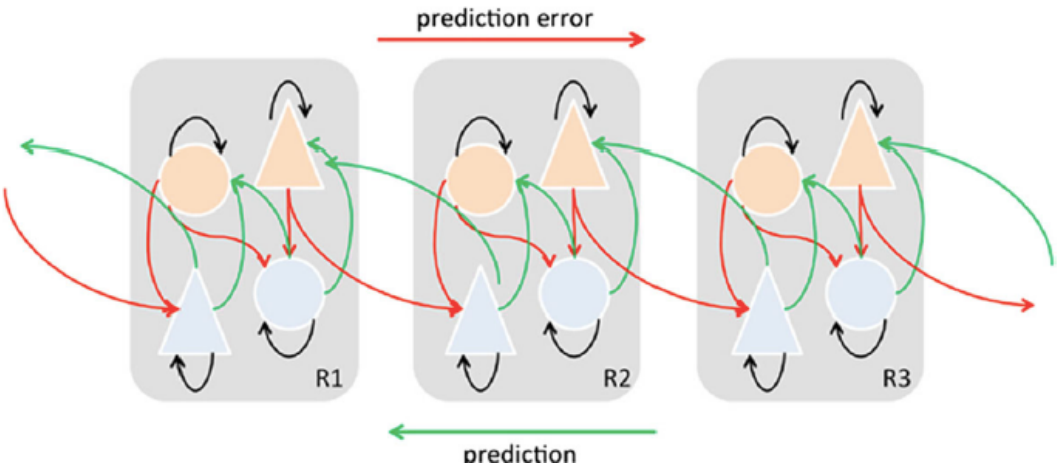

Figure 4. The fundamental organizational motif of the predictive processing (PP) framework. R1-R3 represent progressively higher levels in the stack of predictive models. All models are driven by incoming signal (left-to-right) as compared to the top-down prediction signal. The error units at each layer propagate a signal to the next layer. From Seth et al (2012).