The Mind as a Predictive Modelling Engine: Generative Models, Structural Similarity, and Mental Representation

Daniel George Williams

Trinity Hall College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

August 2018

The Mind as a Predictive Modelling Engine: Generative Models, Structural Similarity, and Mental Representation

Daniel Williams

Abstract

I outline and defend a theory of mental representation based on three ideas that I extract from the work of the mid-twentieth century philosopher, psychologist, and cybernetician Kenneth Craik: first, an account of mental representation in terms of idealised models that capitalize on structural similarity to their targets; second, an appreciation of prediction as the core function of such models; and third, a regulatory understanding of brain function. I clarify and elaborate on each of these ideas, relate them to contemporary advances in neuroscience and machine learning, and favourably contrast a predictive model-based theory of mental representation with other prominent accounts of the nature, importance, and functions of mental representations in cognitive science and philosophy.

For Marcella Montagnese

Preface

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgements

This work was supported by the Arts and Humanities Research Council.

In addition to the many intellectual debts that I note in the thesis, I would also like to thank:

- Richard Holton, whose combination of patience, encouragement, and scepticism made for an ideal supervisor;
- Jakob Hohwy for his invaluable feedback and generosity;
- Huw Price for his support, and for his appealing combination of naturalism and pragmatism that had a big influence on my philosophical development;
- the Cognition and Philosophy Lab at Monash—especially Andrew Corcoran, Stephen Gadsby, Julian Matthews, and Kelsey Palghat—for their extremely useful feedback, and for providing an ideal research community;
- Lincoln Colling, a brilliant collaborator;
- Juliet Griffin, a force of nature;
- the philosophy graduate community at Cambridge;
- Tim Crane and Marta Halina for useful advice at different stages of the PhD;
- the anonymous reviewers for the helpful and constructive feedback on the papers that I have published whilst a PhD student;
- Michael Morris and Sarah Sawyer, whose influence on me at Sussex played a large part in my decision to pursue philosophy further;
- My family—mum, dad, Beck, Josh, and Sam—for everything;
- and Marcella Montagnese, who made the second half of the PhD roughly 20 times more enjoyable than the first half.

Although almost all of the thesis is new, it draws heavily on ideas from the following published articles (cited at relevant points in the text):

Gadsby, S., & Williams, D. (2018). Action, affordances, and anorexia: body representation and basic cognition. *Synthese*. doi: 10.1007/s11229-018-1843-3

Williams, D. (2017). Predictive Processing and the Representation Wars. *Minds And Machines, 28*(1), 141-172. doi: 10.1007/s11023-017-9441-6

Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness And Cognition, 61*, 129-147. doi: 10.1016/j.concog.2018.03.003

Williams, D. (2018). Pragmatism and the predictive mind. *Phenomenology And The Cognitive Sciences*. doi: 10.1007/s11097-017-9556-5

Williams, D. (2018). Predictive coding and thought. *Synthese*. doi: 10.1007/s11229-018-1768-x

Williams, D. (2018). Predictive minds and small-scale models: Kenneth Craik's contribution to cognitive science. *Philosophical Explorations, 21*(2), 245-263. doi: 10.1080/13869795.2018.1477982

Williams, D., & Colling, L. (2017). From symbols to icons: the return of resemblance in the cognitive neuroscience revolution. *Synthese, 195*(5), 1941-1967. doi: 10.1007/s11229-017-1578-6

List of Figures

Contents

Chapter 1. Introduction and Overview

> "The fundamental nature of cognition is rooted in the tricks by which assorted representational schemas give organisms a competitive advantage in predicting" (P.S. Churchland 1986, p.14).

(1.)**1. Introduction**[1]

In this thesis I outline and defend a theory of mental representation founded on a simple idea. The idea is not original. Its basic tenets were advanced by the Cambridge psychologist, philosopher, and control theorist Kenneth Craik in the 1940s, and it has recently won substantial support from a promising body of research in contemporary cognitive science. In a slogan, the idea is this: the mind is a *predictive modelling engine*. On this view, mental representations are elements of in-the-head models that re-present the structure of worldly domains in a form that can be exploited for prediction. Explaining what this means, why one should believe it, and what some of its most important implications and limitations are—these are my chief aims in what follows. In this introductory chapter I situate these aims in their appropriate philosophical context and provide a summary of the core structure and contents of forthcoming chapters.

(1.)**2. The Philosophy of Mental Representation**

What business does a philosopher have in offering a theory of mental representation? Questions concerning the existence, importance, and nature of mental representations seem like ordinary scientific questions. It is obvious why philosophers of the *past* have consistently engaged such questions. It is much less obvious what a philosopher can contribute to such questions today—that is, in an era of rapid developments in neuroscience, cognitive psychology, and artificial intelligence research.

Nevertheless, an enormous amount of recent philosophical energy has been expended on such questions. Philosophers have advanced theories about the vehicles of mental representation with substantive implications for how the brain works—for example, that the representational system inside our heads must be structured like a language (Fodor 1975; Sterelny 1990; Schneider 2011). They have produced elaborate theories attempting to explain what makes mental representations represent—for example, that states of the brain represent whatever they have been recruited to indicate by evolution or learning (Dretske 1981; Millikan 1984;

[1] For clarity, I place chapter number identifications in parentheses and eliminate them from intra-chapter references to sections within the text.

Papineau 1984). And they have made bold pronouncements on what conditions something must satisfy to qualify as a mental representation, in some cases arguing that the bulk of practicing cognitive scientists do not know what the term "representation" means (Bennett & Hacker 2003; Ramsey 2007).

Of course, such proposals are not typically advanced in complete independence from research in the cognitive sciences. Indeed, each of the three proposals just canvassed arose from a substantial engagement with that research. Nevertheless, this engagement does invite an obvious question: what is it about the topic of mental representation that has elicited such an enormous amount of philosophical scrutiny in recent times?

Although a complete answer to this question falls beyond the scope of this thesis, a significant part of the answer can be identified relatively straightforwardly: the concept of mental representation exhibits a deeply strange dual profile in the contemporary mind sciences.

On the one hand, the concept has been *foundational* in orthodox cognitive science since the "cognitive revolution" of the late 1950s and 1960s (Bermúdez 2010). Indeed, it is so foundational that some of the most prominent cognitive psychologists of the last fifty years simply *define "*psychology as the study of mental representations" (Chomsky 1983, p.5; Gallistel 2001). From this perspective, the concept of mental representation features as part of a nexus of concepts—REPRESENTATION, COMPUTATION, INFORMATION PROCESSING—that provide the core theoretical framework for modelling the mind across the cognitive sciences, uniting research programmes that otherwise disagree sharply about the basic make-up of intelligence (Bechtel 2008; Thagard 1996). The so-called representational theory of mind is thus viewed as being "as fundamental to cognitive science as the cell doctrine is to biology and plate tectonics is to geology" (Pinker 1994, p.78; see also Newell 1980, p.136).

Against this tidy orthodoxy, however, the topic of mental representation has always been fraught with theoretical controversy and conceptual confusion. From the behaviourists' elimination of mental representations to the emergence of research traditions such as ecological psychology, embodied cognition, and enactivism, the twentieth and twenty-first century have generated a steady stream of research that seeks to either marginalise or eliminate the concept from psychological theorising. Even within orthodoxy, however, disagreements over the *format* of mental representations have generated some of the sharpest and most intractable theoretical divisions in recent cognitive science. Worse, these empirical controversies have been compounded by a set of enduring conceptual and philosophical puzzles that refuse to go

away. What *are* mental representations? How do representational *explanations* work? How can the ontologically peculiar properties of meanings and reasons that mental representations bring with them be integrated into a scientifically responsible metaphysics?

As Ramsey (2007, p.xi) puts it,

> "[T]here is nothing even remotely like a consensus on the nature of mental representation… [T]he current state of affairs is perhaps best described as one of disarray and uncertainty. There are disagreements about how we should think about mental representation, about why representations are important for psychological and neurological processes, about what they are supposed to do in a physical system, about how they get their intentional content, and even about whether or not they actually exist."

In other words, the literature on mental representation in both philosophy and cognitive science is a *mess*. It is this state of disarray and uncertainty that provides the theoretical context for a significant body of work within the philosophy of mental representation over recent decades. Although it is something of a simplification, much of this work can be understood as engaging the following three foundational—and deeply interrelated—questions (see Ramsey 2016).

(1.)3. Three Questions about Mental Representation

The first question is the most basic: what are the distinctive properties and relations in virtue of which something—a neural or computational structure, for example—functions as a mental representation in the first place? Without an answer to this question, there can be no way of adjudicating one of the most divisive questions in twentieth and twenty-first century psychology noted above: how important are mental representations to the mechanisms underlying psychological capacities? Are they central to all cognitive processes—to any process worthy of the name "cognitive" (Gallistel 2001)—or are they a "peripheral and emergent" (Anderson 2014, p.182) feature of such processes? Despite this, it is remarkably difficult to find clarity or consensus on this functional question if one turns to the scientific literature. Worse, as I will argue in Chapter 2, some of the answers that one does find in that literature are deeply unsatisfactory. One important aim of philosophical work on mental representation over recent years has therefore been to provide much-needed clarity on this issue (Bechtel 2008; Clark 1996; Godfrey-Smith 2006a; Orlandi 2014; Ramsey 2007; Van Gelder 1995).

The second question addresses not the distinctive functions of mental representations but what might opaquely be called their "nature." In fact, there are really two questions here. First, what

form do the vehicles of mental representations take in the brain? Are they structured like maps, models, images, schemas, sentences, or spreading activation patterns in high-dimensional neural networks? Although this controversy seems to be straightforwardly empirical, it is a question on which philosophers have made substantial contributions, clarifying, championing, and critiquing the implications of different research programmes for this foundational question. Second, how do mental representations acquire their contents? What makes it the case that some pattern of activity in my brain represents *dogs* rather than something else or nothing at all? Von Eckardt (2012) calls this question of content determination the "foundational problem of cognitive science." It is the lack of any consensus on how to solve it that leads Dietrich (2007, p.1) to confidently assert that "no scientist knows how mental representations represent."

Finally, the third question addresses the role of mental representations within scientific *explanation*: how does positing mental representations contribute to the explanation of psychological capacities? Or—perhaps more fundamentally—*does* the postulation of mental representations contribute to the explanation of psychological capacities? Certainly many have argued that it does not. If it does, however, as most cognitive scientists believe, how does it? Do representational explanations merely provide an instrumental "stance" or "heuristic overlay" (Dennett 1987, p.350) on fundamentally non-representational mechanisms, restricted to the "informal presentation of a theory" (Chomsky 1995, p.55)? Or do representational explanations identify features of psychological mechanisms that play a genuine causal role *qua* representations? If the latter, how can the relational properties that mental representations bear to what they represent be causally relevant to their internal role within a mechanism?

(1.)**4. A Brief Overview**

Separating these questions in the foregoing manner is slightly artificial given the important ways in which they interrelate. Further, as I will argue in Chapter 2, a good deal of contemporary work in the philosophy of mental representation either neglects the first question or conflates it with questions concerning content determination (see Ramsey 2016). Nevertheless, they have generated at least partially distinguishable threads in the enormous body of philosophical work on mental representation in recent decades, and they provide the context in which the theory of mental representation that I outline and defend in this thesis should be understood. That is, when I argue that we should understand mental representation in terms of neurally realised predictive models that share an abstract structure with target

domains in the body and world, I intend this theory to address such questions. Of course, the theory does not solve *all* the theoretical puzzles surrounding mental representation (see Section 5.3 below). Nevertheless, it does constitute genuine progress—or so I will argue. For an extremely schematic preview—to be elaborated, defended, and qualified at length in forthcoming chapters—the answers I provide to such questions are as follows.

On the functional question, I argue that mental representations should be understood as functional analogues of public representations—as structures within the mind/brain that exhibit a similar functional profile to external representations familiar from both everyday life and science and engineering. Of course, this suggestion is by no means original (Bechtel 2008; Godfrey-Smith 2006a; Ramsey 2007). Nevertheless, it does stand in stark contrast to many prominent answers to this question in contemporary philosophy. For example, it contrasts with the widespread view in neuroscience that representation is simply a matter of detection or systematic covariation and with the widespread view in philosophy that the only defining property of representations is their content. Further, it is opposed to philosophical work that seeks to get a grip on the phenomenon of mental representation by appeal to nebulous terms such as "aboutness" or the mind's "cognitive contact" (Kriegel 2008, p.308) with the world. By focusing on how public representations *work*—their role as tools for representation users in the service of goals—my approach to this question places a strong emphasis on functional properties of the sort that are relevant to the explanatory concerns of science.

On the second question, I draw on research from cognitive neuroscience and machine learning to argue that mental representation should be understood in terms of *generative models*. Specifically, I draw on research according to which a fundamental function of the mammalian neocortex is model-based *prediction*, which underlies core psychological capacities such as unsupervised learning, perception, action, and imaginative simulation. As such, mental representation should be understood in terms of neurally instantiated models of the world. As is familiar from ordinary scale models and from influential work on modelling in the philosophy of science, models capitalize on structural similarity to their targets (Giere 2004; Godfrey-Smith 2006b; Weisberg 2013). In the case of mental models, I will argue that the structure being re-presented is the complex networks of causal and statistical relationships among features of the body and environment. As such, a vision of the mind as a predictive modelling engine vindicates an "iconic" (similarity-based) understanding of mental representation. This view of mental representation was heretical throughout much of the twentieth century, written off as conceptually problematic and inconsistent with materialism

(Fodor 1984; Goodman 1969; Sterelny 1990). Extolling the virtues of this heresy is one of my chief aims in what follows.

Finally, an answer to the explanatory question is relatively straightforward. If our psychological capacities are dependent on prediction, and successful prediction is dependent on the accuracy of our mental models, then our psychological capacities are dependent on the accuracy of our mental representations—on the degree to which they recapitulate the causal-statistical structure of target domains. A vision of the mind as a predictive modelling engine thus provides a simple yet powerful framework for understanding how the representational properties of our mental states can play a genuine causal role in the mechanisms responsible for our psychological capacities.

(1.)**5. Clarifications**

Such—in extremely brief and schematic outline—is the account of mental representation that I will outline and defend in this thesis. Before I provide a summary of the structure and contents of forthcoming chapters, I want to note three clarifications that will be important to bear in mind in what follows.

(1.)**5.1. A Framework**

First, the theory of mental representation that I articulate and defend in this thesis is intended as a substantive proposal about *the world*—specifically, about that part of the world that includes certain animals and other systems capable of constructing mental representations of the world. It is not intended as conceptual analysis or a priori philosophy. There are possible worlds in which the account is mistaken, one of which may be *this* world if the scientific theories and empirical research that I draw on turn out to be incorrect.

Nevertheless, the theory is not itself a scientific theory. Rather, it is intended to subsume a set of promising scientific theories and research programmes, spell out their implications for our understanding of the nature of mental representation, and distinguish such implications from other theoretical approaches to mental representation. Although I do draw on concrete instantiations of the account of mental representation that I defend in contemporary neuroscience (see below), the account itself is intended to be consistent with multiple views concerning the specific algorithms implicated in neural information processing, their implementation, and so on. Of course, the account is not so schematic as to be consistent with anything. Of the many theoretical approaches to the topic of mental representation that it is

inconsistent with, there will be three that will feature prominently in this thesis: first, the idea that mental representation is marginal or non-existent; second, the idea that mental representation is primarily about "detection" or "indication"; and third, the idea that the representational system inside our heads is a mental language for expressing the contents of our folk psychological states.

(1.)5.2. Synthesising and Systematising

The second clarification is this: as stressed above, a vision of the mind as a predictive modelling engine is obviously not original to me. Fodor (2008, p.3) has recently recalled the experience of writing "The Language of Thought" in the early 1970s: "I remember thinking (and not without a certain satisfaction) that it didn't contain a single new idea." A similar sentiment applies here.

Most obviously, the scientific research that I draw on is not my own. As I argue in Chapter 3, the first scientist to propose a vision of the mind as a predictive modelling engine—at least as I will present this idea—was the mid-twentieth-century psychologist and control theorist Kenneth Craik (1943).[2] For example, Craik (1943) explicitly argues that "one of the most fundamental properties of thought is its power of predicting events," which gives it its "immense adaptive and constructive significance" (p.50), and which is made possible because an organism "carries a 'small-scale model' of external reality and of its own possible actions within its head" (p.61)—a small-scale model "which has a similar relation-structure to that of the process it imitates" (p.51).

It is only in the previous two decades, however, that an emphasis on model-based prediction as one of—if not the—fundamental components of intelligence has become central to the cognitive sciences (Bubic et al. 2010; Friston 2005; Kveraga et al. 2007). For example, it is now not uncommon to read claims such as the following in both popular and academic scientific works: that "the capacity to predict the outcome of future events—critical to successful movement—is, most likely, the ultimate and most common of all global brain functions" (Llinás 2001, p.21); that "prediction is not just one of the things your brain does. It is the *primary function* of the neocortex, and the foundation of intelligence" (Hawkins &

[2] Hermann von Helmholtz is typically credited as the grandfather of contemporary interest in the "predictive mind" (sometimes alongside the Arabic polymath Ibn al Haytham writing a millennium ago) (see Clark 2016, p.307; Hohwy 2013, pp.5-6). These thinkers are really the originators of a "constructivist" understanding of perception in terms of knowledge-driven *inference*, however, which need not include active prediction. Further, neither focused on neurally realised structural models. (See also Chapter 3, footnote 9).

Blakeslee 2004, p.89); and that "envisaging the future is one of the primary reasons we have a neocortex" (Kurzweil 2012, p.52). As Bubic et al. (2010, p.2) note in an influential review paper, "the term 'predictive brain' depicts one of the most relevant concepts in cognitive neuroscience." Barr (2011, p.v) concurs, reporting a broad consensus that prediction is "a universal principle in the operation of the human brain." "The essence of intelligence," write Yann LeCun and his colleagues at Facebook AI Research, "is the ability to predict" (Henaff et al. 2016, p.1).

The influential theory that takes this emphasis on prediction to its extreme is *predictive processing*, a theory in cognitive and computational neuroscience that models the neocortex as a hierarchically structured organ of probabilistic prediction (Clark 2013; Friston 2005; Friston et al. 2017; Hohwy 2013; Rao and Ballard 1998; Seth 2015). In this theory, the brain is an organ for minimizing *prediction error*, the mismatch between internally generated model-based predictions of its evolving sensory inputs and the sensory inputs themselves. Because it represents the maximal expression of a vision of the mind as a predictive modelling engine, I draw extensively on this theory in this thesis. Nevertheless, it sits at the extreme end of a continuum of theories in cognitive neuroscience that stress the fundamental role of generative model-based prediction in the functioning of the brain, and it is this broader set of theories that provides the scientific foundation for the account of mental representation that I articulate and defend.

In addition to this scientific research, I also draw heavily on a body of broadly philosophical work in what follows. Some of this work explicitly traces the implications of an understanding of mental representation in terms of predictive modelling. The chief influences here are Andy Clark (2016), Rick Grush (2004), Jakob Hohwy (2013), and Dan Ryder (2004). In addition, however, I have also sought to relate these ideas concerning predictive modelling to a heterodox tradition in both theoretical psychology (Gallistel 2001; Johnson-Laird 1983; Palmer 1978) and philosophy (P.M. Churchland 2012; Cummins 1996; O'Brien and Opie 2004; Gładziejewski and Miłkowski 2017; Isaac 2013; Ramsey 2007) that advocates an understanding of mental representation in terms of isomorphism, homomorphism, or structural resemblance.

I have stressed the provenance of these ideas to emphasise the following point: the originality in what follows resides not primarily in the ideas themselves, but in the concerted attempt to

bring a set of ideas together to yield a systematic theory of mental representation, which I then apply in novel areas.

(1.)**5.3. The Mind is Not Only a Predictive Modelling Engine**

The final clarification is this: the theory of mental representation that I defend in this thesis is founded on a vision of the mind as a predictive modelling engine. Crucially, however, it is not based on the claim that the mind is *only* a predictive modelling engine. History has not looked kindly on claims to the effect that the mind is only one thing—an elaborate stimulus-and-response mechanism, a large associative network, a symbol-manipulating device, and so on. I bet that history will never look kindly on such claims: psychological capacities in humans and other animals have a complex evolutionary history rooted in multiple different ancestral environments and "design problems" (Dennett 1995), and they are subserved by a diverse range of neural and—in some cases, at least—bodily and environmental structures (Clark 2008). As such, the claim that I defend in what follows is simply that *one* of the things that the mind "is"—and certainly not the only thing—is a predictive modelling engine, and that this fact is especially important when it comes to understanding the nature of mental representation.

Most importantly, the theory of mental representation that I defend does not capture everything that falls under the category of mental representation. The most notable phenomena that I will not address in what follows are the suite of human capacities subserved by public representational technologies such as natural language, mathematical symbols, and so on. It is plausible that a significant component of what makes human cognition genuinely distinctive relative to the capacities exhibited by other animals is precisely such representational tools, the accumulated bodies of cultural information they contain, and the social practices in which their tokenings are subject to critique and evaluation (P.M. Churchland 2012; Clark 1996). Instead, my target is pre- or non-linguistic mental representation—that is, the kind of mental representation that exists prior to and in many cases independently of public symbol systems, and thus the kind that we share with many other animals.

As Paul and Patricia Churchland (P.S. Churchland 1986; P.M. Churchland 1995; 2012) have argued over many years, what philosophy needs is a neuroscientifically informed understanding of mental representation that honours the profound continuities between humans and animals that lack public symbol systems. Without such a framework, we will not understand the nature of our own minds, and we will not grasp the powerful ways in which

public symbols might augment and transform them. It is such a framework that I intend to develop and defend in what follows.

(1.)**6. Thesis Overview**

With these clarifications in mind, I will now offer a brief sketch of the contents and structure of forthcoming chapters.

In "Chapter 2: The Representation Wars," I provide a brief overview of the central controversies in cognitive science and philosophy concerning the existence, importance, and nature of mental representations. I argue that a frustrating characteristic of these "representation wars" is the lack of any consensus on how to *adjudicate* them—specifically, a lack of consensus on when representational explanations are appropriate in the first place and on what a theory of mental representation should achieve. To rectify this, I motivate the claims that mental representations should be understood as mental analogues of public representations and that a theory of mental representation should be exclusively motivated by the explanatory constraints of our best cognitive science, and not—as many philosophers assume—folk psychological or semantic intuitions.

This chapter provides the methodological foundation for future chapters. After clarifying what a theory of mental representation should achieve, I set out to provide such a theory.

In "Chapter 3: Kenneth Craik's Hypothesis on the Nature of Thought," I outline some important but largely neglected ideas advanced by Kenneth Craik in the early 1940s. Specifically, I extract from Craik's work three core insights that provide the schematic framework for understanding mental representation that I then elaborate and defend throughout the rest of the thesis: first, an account of mental representation in terms of neurally realised models that share an abstract structure with target domains in the world; second, an appreciation of prediction as the core function of such models; and third, a regulatory understanding of brain function.

The next four chapters elaborate on Craik's proto-theory and situate them in the context of contemporary cognitive science.

In "Chapter 4: Models as Idealised Structural Surrogates," I draw on work from the philosophy of science and elsewhere to outline the chief characteristics of model-based forms of representation. I argue that models should be understood as idealised, holistic, and interest-

relative structural surrogates for target domains and I illustrate these characteristics by reference to cartographic maps, scale models, and dynamic models of the sort used in science.

In "Chapter 5: Could the Mind be a Modelling Engine?" I show how this functional profile of models can be understood in the context of neural mechanisms that underlie psychological capacities. I explain how the concept of structural similarity should be understood when it comes to mental models and features of the body and environment, what determines the targets of such models, and how the structural similarity between mental models and target domains can make a causal contribution to the successful exercise of psychological capacities.

In "Chapter 6: Generative Models and Predictive Minds," I introduce the contemporary formal ideas and empirical research that go some way towards vindicating an understanding of the mind as a predictive modelling engine. First, I introduce the concept of generative models and I explain their varied functions. Second, I introduce Bayesian networks as a way of understanding the structure of generative models. Finally, I turn to research in cognitive neuroscience that models the neocortex as a hierarchically structured prediction machine, drawing especially on predictive processing.

In "Chapter 7: Predictive Processing and Craik's Three Insights," I show how the formal ideas and empirical research outlined in Chapter 6 map onto the three core insights that I extracted from Craik's work in Chapter 3. First, I show how generative models can be understood in terms of the structural similarity-based account of mental representation outlined in previous chapters. Second, I show how predictive processing and related work in the cognitive sciences both vindicate and extend Craik's conviction that model-based prediction underlies our core psychological capacities. Finally, I show how predictive processing can be understood within the context of the regulatory understanding of brain function championed by Craik and other cyberneticists.

With this overarching theory of mental representation in place, in the next three chapters I then argue that it can answer three of the most serious challenges to representationalist accounts of the mind as advanced over the past century or so.

In "Chapter 8: The World Is Not Its Own Best Generative Model," I consider the so-called "replacement hypothesis" (Shapiro 2010): the hypothesis that an organism's body in interaction with the environment replaces the need for mental representations in cognitive processes. I argue that proponents of the replacement hypothesis neglect the predictive character of sensorimotor processing and that a predictive model-based account of mental

representation can accommodate the phenomena that motivate "radical" embodied views of cognition.

In "Chapter 9: Modelling the Umwelt," I consider the argument that the pragmatic nature of intelligence implies a profoundly "organism-relative" understanding of the mind-world relation that is inconsistent with the structural similarity-based account of mental representation that I defend. I argue that claims of radical organism-relativity are implausible, and I show how a generative model-based account of mental representation can clarify this topic and accommodate our representation of phenomena such as affordances and response-dependent properties.

In "Chapter 10: Generative Intentionality," I consider the classic argument that representational content is so metaphysically problematic that it should be eliminated from scientific theorising. I contrast the generative model-based account of mental representation that I defend with the folk psychology-inspired account that has dominated philosophical work on content determination in recent decades and I sketch in broad strokes how content determination and the vexed issue of misrepresentation should be understood within probabilistic generative models.

Finally, I conclude the thesis in "Chapter 11: Conclusion" by briefly summarising the account of mental representation that emerges from previous chapters before identifying three of the most important areas for future research that it gives rise to.

Chapter 2. The Representation Wars

> "We, as cognitive psychologists, do not really understand our concepts of representation. We propose them, talk about them, argue about them, and try to obtain evidence in support of them, but we do not understand them in any fundamental sense" (Palmer 1978, p.259).

(2.)1. Introduction

Are there such things as mental representations? If so, how important are they to cognitive processes? What form do they take?

These questions define what might reasonably be called the "representation wars" (Clark 2015; Williams 2017) in cognitive science and philosophy, a divisive and seemingly intractable set of debates concerning the existence, importance, and nature of mental representations within the mind.[3] In this chapter I briefly review these controversies and argue that a frustrating characteristic of the representation wars is the lack of consensus on how to adjudicate them. Specifically, what is needed are answers to two fundamental questions. First, what are the distinctive properties and relations in virtue of which something qualifies as a mental representation in the first place? Second, what should a successful theory of mental representation achieve? I criticise widespread answers to these questions in the philosophical literature and motivate two claims: first, that mental representations should be understood as functional analogues of public representations within the mind/brain; second, that a theory of mental representation should be exclusively answerable to the explanatory constraints of our best cognitive science, and not—as many philosophers assume—folk psychological and semantic intuitions

I structure the chapter as follows.

In Section 2 I argue that controversies over the existence and importance of mental representations within the mind are hampered by the lack of consensus on what mental representations *are*. In Section 3 I criticise three prominent views in the philosophical literature that bear on this question and I motivate the view that mental representations should be understood as mental analogues of public representations. In Section 4 I introduce controversies over the format of mental representations and the determination of their contents. In Section 5 I argue against a widespread identification of a theory of mental representation

[3] Here I extend the term "representation wars" to cover not just debates concerning the importance of mental representations in cognitive processes but also controversies over their format and content determination, which have often been just as heated and divisive.

with the attempt to "naturalize" folk psychology and I motivate alternative naturalistic constraints on such a theory.

I conclude in Section 6 by summarising the foregoing lessons and identifying their relevance for future chapters. As will be clear, the overall aim of this chapter is to defend an unforgiving naturalism as the methodological framework for approaching questions concerning mental representation. Too much philosophical work on this topic has been divorced from the explanatory concerns of cognitive science and overly focused on a project of metaphysical domestication.

(2.)**2. The Existence and Importance of Mental Representations**

How important are mental representations to the mechanisms underlying our psychological capacities? This question lays the foundation of the representation wars that have raged for over a century in both psychology and philosophy (see Chemero 2009). It is often said that what most sharply distinguished the interdisciplinary research programme of cognitive science that emerged in the late 1950s from the behaviourist movement it superseded was its commitment to the centrality of mental representations in cognitive processes (Bechtel 2008; Bermúdez 2010). Since then, this commitment has helped to define orthodoxy across the mind sciences, uniting research traditions that otherwise disagree sharply about the basic make-up intelligence. Indeed, these disagreements often hinge on the *format* of these mental representations (Thagard 1996; see Section 4 below).

Against this orthodoxy, however, there has always been a steady tradition of psychological research that seeks to marginalise or in some cases eliminate the concept of mental representation from psychology. The most salient example is the tradition of ecological psychology pioneered by Gibson (1979), itself inspired by work from the traditions of American pragmatism and phenomenology (see Chemero 2009), and its descendants and close cousins in contemporary research in embodied cognition and enactivism (Anderson 2014; L. Barrett 2011; Gallagher 2017; Hutto & Myin 2012; Varela et al. 1993). As we will see in Chapter 8, much of the work within these research programmes emphasises how embodied and dynamic interactions with the environment replace the need for internal structures to *stand in* for—to re-present—that environment.

Importantly, these scientific challenges to representationalist accounts of the mind have been coupled to a steady stream of more philosophically-oriented scepticism. For example, there are arguments that representational explanations in cognitive science are irremediably confused,

mis-applying concepts that belong at the personal-level to sub-personal mechanisms (Bennett & Hacker 2003; Ryle 1949; Wittgenstein 1953), that the structures implicated in cognitive processes do not genuinely function as representations (Ramsey 2007), that semantic properties are causally irrelevant to the functioning of cognitive mechanisms (Stich 1983), and that semantic properties are so metaphysically problematic that they should be eliminated from science (Quine 1960; Rosenberg 2015).

(2.)**2.1. The Constitutive Question**

In one way or another, controversies over the existence and importance of mental representations in psychological mechanisms all hinge on whether the structures implicated in such mechanisms *qualify* as mental representations. If we are to adjudicate such controversies, then, we need an answer to what I will call the *Constitutive Question*:

> (Constitutive Question) What are the distinctive properties and relations in virtue of which something qualifies as a mental representation?

If we had an answer to the Constitutive Question, debates concerning the existence and importance of mental representations in the mind would be straightforward to adjudicate: we would just check whether the component parts of psychological mechanisms exhibit the relevant properties and relations. Nevertheless, such debates have not been straightforward to adjudicate. Controversies concerning the existence and scope of mental representations in the mind continue to rage in both cognitive science and philosophy. In a recent introduction to perceptual psychology comparing representational and non-representational theories of perception, for example, Rogers (2017, p.6) writes that "you may find it remarkable that there is still no real consensus as to which of these approaches to the subject is more appropriate."

Why have such debates proven so resistant to a consensus empirical resolution?

Although there are likely many answers to this question, the one that I want to focus on here is this: in addition to first-order controversies concerning the mechanisms underlying intelligence, the representation wars are riddled with disagreements over answers to the Constitutive Question—that is, disagreements concerning how psychological mechanisms would have to operate to qualify as representational (Bechtel 2008; Rowlands 2015). For example, answers in both the philosophical and psychological literature range from specifications of conditions for representational status so weak that *any* conceivable cognitive

system would be representational to specifications of conditions so strong that it becomes difficult to see how anything *could* qualify as a mental representation (Ramsey 2007).

Without settling the Constitutive Question, there can be no way of adjudicating how important mental representations are in psychological mechanisms. Given that one of my chief aims in this thesis is to argue that mental representations are profoundly important in psychological mechanisms—and thus that nonrepresentational paradigms in cognitive science are mistaken (see Chapters 8-10)—it will therefore be useful to address this issue.

(2.)**3. Three Answers to the Constitutive Question**

Any exhaustive summary of answers to the Constitutive Question is beyond the scope of this thesis. Nevertheless, by way of motivating what I think is a useful answer to this question—a way of understanding mental representations that renders debates over their existence both substantive and interesting—I will first outline two alternative and superficially attractive views in the philosophical and neuroscientific literature that bear on the question.

(2.)**3.1. Deferring to Science**

One obvious response to the Constitutive Question states that mental representations are just whatever cognitive scientists say they are—that we can or at least should just defer to the use of the term "mental representation" in cognitive science in answering the Constitutive Question.

There are at least two powerful reasons for endorsing this view. First, it gives concrete expression to the idea—surely correct—that science should not be limited by the conceptual constraints of commonsense (Ladyman et al. 2014). This deference to science is a central feature of the philosophical naturalism that I will assume throughout this thesis (see below). Second, it is often noted—correctly—that MENTAL REPRESENTATION is a theoretical concept, and not itself a term of ordinary discourse (Pitt 2017). It is therefore plausible that it should mean whatever cognitive scientists intend by it—that its meaning should be fixed by the theories in which the concept features, rather than by pre-theoretic stipulation.

As will be clear below (Section 5), I agree with the spirit of this view: questions concerning the existence, importance, and nature of mental representations *should* be exclusively settled by science, broadly construed. Nevertheless, it is one thing to say that science should determine whether there are such things as mental representations and another thing to say that

contemporary scientific usage of the term "mental representation" should determine how we answer the Constitutive Question.

To see why these two things should be kept apart, it will be helpful to see how the term "representation" is often used in contemporary cognitive science and the problems that this gives rise to.

(2.)**3.1.2. Detection as Representation?**

Throughout much of cognitive neuroscience and work in artificial intelligence (especially in neural network modelling), the term "representation" is often used to denote nothing more than a state of a cognitive system that responds selectively to some (usually distal) environmental condition (O'Reilly and Munakata 2000, p.25). As Friston and Price (2001, p.279) put it,

> "Here a representation is taken to be a neuronal event that represents some "cause" in the sensorium. It can be defined operationally as the neuronal responses evoked by the cause being represented."

On this view, a state of a cognitive system qualifies as an internal representation if it is differentially responsive to specific environmental conditions. Paradigmatic mental representations are thus things like edge detectors in the primary visual cortex that respond selectively to edges at certain orientations in the visual field and neural states in the so-called "fusiform face area" that respond selectively to specific faces. Importantly, this understanding of representation has been embraced by many philosophers. Chalmers (2003, p.111), for example, defines "a notion of *functional representation*, on which P is represented roughly when a system responds to P." In many philosophical elaborations of this idea, it is often added that these detectors must also have the evolved or acquired function of responding to or systematically covarying with the relevant condition (Dretske 1988; Millikan 1984).

As several philosophers and cognitive scientists have pointed out, the problem with this widespread view is that differential responsiveness does not seem to constitute representation (Beer 1995; Cummins 2010; Orlandi 2014; Ramsey 2007; Van Gelder 1995). Specifically, there seems to be a commonsense and important distinction between a system that merely detects, indicates, or tracks environmental conditions and a system that relies on internal representations of such conditions (Orlandi 2014). For this reason, there does not seem to be any explanatory loss in replacing all talk of "representation" in such contexts with talk of things like *causal relays* and *tracking states* (Ramsey 2007).

Worse, this account of representation massively over-generalises, including as representations all kinds of things that do not seem to be representational at all. For example, states of our digestive and immune systems are selectively responsive to specific inputs; indeed, it is their function to be so responsive. It seems odd to say that they therefore *represent* such inputs, however (see Ramsey 2007). As Orlandi (2014, p.52) puts it,

> "Virtually all biological and organic systems, including plants and other very simple mechanisms have features that evolved in response to environmental pressure—that is, features that lawfully co-vary with some stable worldly fact and that have the function of doing so."

As such, the problem with understanding representation in this way is that it effectively settles debates over the importance of mental representations in psychological mechanisms by definitional fiat (Orlandi 2014; Ramsey 2007). No view—actual or possible—denies that intelligent agents possess inner states that respond selectively to environmental conditions. If "representation" is defined in this way, then, *any* conceivable cognitive system will make use of internal representations. As such, the representational theory of mind would not be a substantive and interesting scientific hypothesis. It would be a banality.

One might respond that our conceptual deference should be reserved for legitimate or useful applications of the term "representation" throughout cognitive science. For example, many cognitive scientists explicitly deny that detection or systematic covariation is sufficient for representational status (Beer 1995; Brooks 1991). To say this, however, is to effectively abandon the view: to distinguish between legitimate and non-legitimate applications of the term "representation" in this way, one must be drawing on a source independent of cognitive scientific practice.

As the foregoing remarks suggest, one obvious place to look for this independent source is our ordinary understanding of representation. Specifically, although MENTAL REPRESENTATION is a theoretical concept, it is composed from concepts—MENTAL and REPRESENTATION—that have their home in ordinary thought and discourse. The main problem with any identification of detection or covariation with representation seems to be that it fails to conform to our ordinary understanding of representational function. That is, there is a big gulf between things like maps, models, graphs and sentences on the one hand and states that are merely differentially responsive to certain conditions on the other. This suggests that a chief constraint on any answer to the Constitutive Question is that it should conform to our ordinary (pre-theoretic) understanding of representation. I return to this suggestion below (S3.3).

Importantly, although I stated above that many philosophers embrace the view that something qualifies for representational status if its systemic function is to detect or systematically covary with specific conditions, it may be that such philosophers are not trying to explain *what makes something a representation* but rather how states that *are* representations get their contents (Dretske 1998; Millikan 1984; see Section 4 below). As such, one might point out that they are not vulnerable to the foregoing critique (Orlandi 2014, pp.113-4). If this is correct, these philosophers must provide some independent account of what makes something a mental representation. As we will see next, however, one of the biggest problems with the philosophical literature on mental representation is its neglect of this issue in favour of an almost exclusive concern with questions concerning content determination.

(2.)**3.2. Mental Representation as Mental Content**

The second view that I will consider is probably the most widespread in the philosophical literature. It states that mental representations are just mental states with semantic properties or *content*, where this is generally understood in terms of satisfaction conditions of some kind (see Burge 2010; Pitt 2017; Rescorla 2013; Ryder 2009a). For example, Cummins (1989, p.11) writes that "the central question about mental representation is this: What is it for a mental state to have a semantic property? Equivalently, what makes a state (or an object) in a cognitive system a *representation*?"

This answer to the Constitutive Question implies that the fundamental debate over the existence of mental representations boils down to whether mental states have semantic properties. In conjunction with the widespread thesis that the only *real* properties are those that can be reduced to properties of the sort described in the physical sciences, it thus encourages the widespread view that issues concerning the existence of mental representations hinge on whether semantic properties can be reduced to non-semantic "natural" properties (Hutto & Myin 2012; Rosenberg 2015; see Section 5 below). As Fodor (1984, p.232) puts it,

> "The worry about representation is above all that the semantic…will prove permanently recalcitrant to integration in the natural order; for example, that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is, at a minimum, the framing of *naturalistic* conditions for representation."

Despite the prominence of this view in the philosophical literature, I think that it is unattractive. Specifically, the focus on semantic properties tell us at best half the story. It neglects the issue—in many ways more important when it comes to adjudicating controversies over the importance

of mental representations in cognitive processes—of what it is to *function* as a mental representation (Ramsey 2007). A long tradition in philosophy holds that 'representing is a functional status or role of a certain sort, and to be a representation is to have that status or role'' (Haugeland 1991, p. 69; see Peirce 1931-59). An answer to the Constitutive Question that focuses exclusively on content offers no insight into what this functional role *is*. It doesn't specify the distinctive *jobs* that mental representations are supposed to perform within the cognitive systems of which they are a part.

To see this, note that an exclusive focus on questions concerning content determination neglects the question of why some parts of systems are proper objects of semantic evaluation in the first place. This is important because semantic characterisations of systems are cheap: it is *possible* to describe just about anything as instantiating contentful states such as beliefs and desires (Dennett 1987; Ramsey 2007). One of the things we want from an answer to the Constitutive Question is an explanation of when such attributions are explanatorily useful and when they are vacuous.

Wheeler (2005, p.6) voices much the same worry:

> "Although the issue of how to specify the meanings of the representations that we (allegedly) have has received library loads of philosophical attention, the question of under what circumstances it is appropriate to engage in representational explanation at all remains curiously underexplored."

For this reason, an exclusive focus on content is unsatisfactory. It neglects the question of what it is to *function* as a mental representation—the distinctive jobs that mental representations perform—and thus the question of why some structures are proper objects of semantic evaluation in the first place.

(2.)**3.3. Functional Analogues of Public Representations**

The two views that I have considered in this section both falter on the same problem: they do not automatically accommodate the plausible view that mental representations should be understood as structures that perform distinctive jobs in the cognitive systems of which they are a part. The fourth view that I will describe—and endorse—tackles this problem directly. It states that mental representations are *structures implicated in the production of our psychological capacities that function analogously to public representations* (Bechtel 2008; Godfrey-Smith 2006a; Ramsey 2007).

To unpack this, by "psychological capacities" I mean things like our capacity to perceive, imagine, attend, predict, plan, reason, and so on (Bechtel 2008). Beyond pointing to paradigmatic examples of such capacities, I do not have any sophisticated account of what falls within this domain. These capacities are produced by underlying mechanisms. In general, there is a useful distinction to be made between two levels here: the level of the psychological capacities produced by the mechanism—for example, our capacity to perceive, imagine, reason, plan, and so on—and the component parts and operations of the mechanisms responsible for producing such capacities (Bechtel 2008). On the view defended here, mental representations are component parts of such psychological mechanisms that exhibit a similar functional profile to public representations such as maps, symbols, pictures, models, and so on.

The concept of similarity here is crucial. No doubt there are many differences between prototypical public representations and what goes on within the head. The view outlined here does not require that *all* characteristics of public representations are transposed over to mental representations. Rather, it just requires similarity to what Godfrey-Smith (2006a) calls the "empirical skeleton" of public representation-use and what in future chapters I will call the "core functional profile" of such use. For this reason, the concept of similarity should be understood as a graded notion. It may be that some things are much *better* examples of mental representations than other things, and that in some cases it is simply indeterminate whether a structure qualifies as a mental representation.

(2.)**3.3.1. A Defence**

There are at least two reasons for endorsing this answer to the Constitutive Question.

First, it accurately captures the way in which the concept of mental representation has been understood throughout much of the history of both philosophy and psychology. As we will see below (Section 4.1), for example, throughout the history of philosophy it is "external representations [that] have inspired competing theories of mental representation," generating important hypotheses concerning the nature of human perception, thought, and reasoning (Prinz 2012; p.115; see Waskan 2006). Further, this trend has continued throughout the history of twentieth and twenty-first century philosophy where theories of mental processes based on analogies with public representational media such as language (Fodor 1975) and maps (Braddon-Mitchell & Jackson 1996) abound.

The history of scientific psychology is no different. Prior to the cognitive revolution of the late 1950s, for example, it characterises both Kenneth Craik's (1943) proposal that organisms

contain "small-scale models" of the external world inside their heads and Tolman's (1948) famous hypothesis that rats (and other mammals) navigate their environments through consultation with "cognitive maps." Since the cognitive revolution, it characterises the widespread view that intelligence arises through internal operations on structures such as mental models, sentences, graphs, and so on (Bermúdez 2010). As Bechtel (2008, p.160) puts it, "cognitive scientists and neuroscientists… are proposing that there are such things as internal representations and that these work much like the external representations which we… use in our daily lives."

Of course, as we saw above, this view does not characterise all uses of the expression "mental representation" throughout psychology and philosophy. For example, it is difficult to reconcile with the view that differential responsiveness to specific stimuli is constitutive of representational status. Nevertheless, the view defended here helps to explain why this usage of the term strikes so many as problematic: there is a large gulf between *selective response* and the kinds of structured external representations that are familiar from everyday life.

Second, this answer to the Constitutive Question focuses on functional properties of the sort that are relevant to the explanatory concerns of science and thus provides a way of rendering the representation wars both substantive and interesting. That is, whist it accommodates the emphasis on the importance of semantic properties central to the previous view, public representations are also *tools* that perform specific kinds of jobs for representation-users in virtue of the characteristics that they have. It is not trivial that there are functional analogues of such things inside the head. It is a substantive functional claim about the way in which psychological mechanisms operate.

(2.)**3.3.2. Objections**

Despite this, I think that there are at least two serious objections that this answer to the Constitutive Question confronts.

First, there is the classic objection that there *cannot* be functional analogues of public representations within the head (Bennett & Hacker 2003; Ryle 1949; Wittgenstein 1953). According to this challenge, public representations essentially require interpretation. To say that there are functional analogues of public representations inside the head, then, would imply that there is a homunculus inside the head that interprets such representations—a suggestion that invites regress (who interprets the homunculus's mental representations?).

As I hope to show in later chapters, I think that this objection is mistaken. We can identify the core functional profile of public representation-use in a way that abstracts away from features such as interpretation, social conventions, and intelligence. Of course, this claim requires a substantial defence. As such, if one understands the first objection not as denying that there *could* be functional analogues of public representations inside the head but as denying that the relevant conditions are *in fact* met, it deserves to be taken seriously (Godfrey-Smith 2006a). If there are functional analogues of public representations inside the head, this should be surprising—a substantive discovery of a non-trivial truth.

Second, one might argue that this answer to the Constitutive Question is too restrictive. Even if one concedes that theories of mental representation in cognitive science and philosophy have often been modelled on our understanding of public representations, why rule out the possibility of other legitimate technical accounts of representation that do not conform in any obvious way to our ordinary understanding of public representations? Indeed, it might seem strange to suggest that the brain's representational capacities must be limited to the kinds of external representations that we happen to have come across in the public domain.

Strictly speaking, I could concede much of this objection given my aims in what follows. That is, what my argument in this thesis requires is that the view under discussion here provides a *sufficient* condition for the attribution of representational status and that certain other views—for example, that representation is detection or that representation is just a matter of instantiating semantic properties—do not. Godfrey-Smith (2009, p.31), for example, argues that the understanding of representation defended here—what he calls the "basic representationalist model"—constitutes just "one path" by which "a physical system can acquire a well-motivated semantic description."

Nevertheless, I also think that this objection is less persuasive than it appears.

First, conceiving of mental representation in the manner outlined here obviously does not imply that the brain cannot be doing incredibly complex things that seem alien to anything we have discovered in the external world. The point is that *if* we are to characterise such activity as representational, we must be relying on some independent understanding of representation, and it is difficult to see where else this independent understanding could come from. Indeed, one way of making sense of certain anti-representationalist arguments is that the brain's activity is so alien relative to our ordinary understanding of public representation-use that it should not be considered representational (e.g. Van Gelder 1995; see Chapter 8). This is a powerful

argument that deserves to be taken seriously. Conceiving of mental representation in the manner outlined here enables us to do so.

Second, it is important to distinguish the functional and representational idiosyncrasies of specific kinds of public representations from what might be thought of as the generic functional characteristics that they all share. For example, it is plausible that all public representations function as tools that representation-users exploit as proxies or surrogates for targets that they want to coordinate their behaviour with. Godfrey-Smith (2009, p,33), for example, suggests that the basic character of public representation-use occurs when "the state of one thing (X) is consulted in dealing with another (Y)." As such, when certain philosophers focus on functional properties such as *standing in for, informing,* or *decoupleability* as marks of mental representation, these need not be understood as competing answers to the Constitutive Question. Rather, they can be understood as attempts to extract the generic functional profile of public representation-use. Indeed, it is difficult to make sense of these claims if we do not understand them in this way. Why else should we think of these properties as characteristics of *representation* rather than something else?

For these reasons, I think that the view defended here should be understood as at least providing a sufficient condition for the attribution of representational status and plausibly as providing a necessary condition as well. Of course, so far I have said very little about how the functional profile of public representation-use should be understood and how this might map on to the behaviour of the neural mechanisms that underlie our psychological capacities. For one kind of public representation, at least, I take up this topic at much greater length in forthcoming chapters.

(2.)**3.4. Summary**

Controversies over the existence and importance of mental representations falter on the lack of consensus on what mental representations are. In this section I have motivated the view that mental representations should be understood as functional analogues of public representations within the mind. In the next chapter I will turn to the work of Kenneth Craik who drew upon exactly this concept of mental representation to argue that humans and other animals contain *models* of the world inside their heads.

First, however, I will turn to the other core component of the representation wars.

(2.)**4. Format and Content**

Controversies over the existence and importance of mental representations within the mind comprise just one facet of the representation wars. The other facet concerns not the existence of mental representations but what might opaquely be called their "nature." Although it is something of a simplification, I think that controversies in this area have concerned two core dimensions of mental representation: representational format and content determination. In this section I briefly outline the nature of these controversies before stepping back in the next section to ask a more philosophical question: what should a theory of mental representation achieve? What do we want from such a theory?

(2.)**4.1. Format**

Controversies over the format of mental representation concern the properties of representational *vehicles* relevant to their computational or inferential roles within the cognitive systems of which they are a part.[4] Different representational formats in this sense have important consequences for the kinds of manipulations and transformations the relevant representations permit. Consequently, disagreements concerning representational format have generated some of the sharpest fault lines in debates concerning how the mind works throughout the history of both philosophy and cognitive science (Bechtel 2008; Bermúdez 2010; Thagard 1996).

The history of philosophy gives rise to two canonical theories of representational format in this sense (Prinz 2012; Waskan 2006). On the first, mental representations are analogous to images or pictures, and thinking is the manipulation and interrogation of those images. Historical proponents of this view include Aristotle, Locke, and Berkeley. On the second, mental representations are analogous to linguaformal or mathematical symbols, and thinking is a matter of rule-governed symbolic concatenation and inference. Historical proponents of this view include Hobbes, Leibniz, and Kant.

In the history of twentieth and twenty-first century cognitive science, it is often said that the chief dividing line runs instead between *symbolists* and *connectionists*: namely, those who model mental representations as formally individuated discrete symbols of the sort that feature in digital computation, and those who model representations as distributed patterns of activity across interconnected nodes in a neural network (see Bechtel 2008; Bermúdez 2010). This

[4] Although questions about representational format are therefore related to questions about representational function, there is nevertheless an important conceptual distinction between the two: the former concerns the *structure* of representational vehicles; the latter concerns the distinctive *jobs* the representation performs—that is, "how some neural state actually comes to play a representational role in the first place" (Ramsey 2016, p.5).

taxonomy is probably both too coarse-grained, however, neglecting the enormous variety of views about representational format within these traditions, and also insufficient to capture a large body of research within the cognitive sciences that straddles this divide. Nevertheless, I will not attempt a thorough overview of the enormous and enormously complex literature on representational format here.

(2.)**4.2. Content**

The second focus of debates concerning the nature of mental representation concerns representational *content*. Specifically, it concerns those properties and relations responsible for *determining* the contents of mental representations.

As noted above, representational content in this sense is typically understood in terms of satisfaction conditions of some kind. As Hutto and Myin (2013, p.x) put it,

> "Just what is content? At its simplest, there is content wherever there are specified conditions of satisfaction. And there is true or accurate content wherever the conditions specified are, in fact, instantiated."

For example, beliefs have truth conditions, percepts are often thought to have accuracy conditions, desires have satisfaction conditions, and so on.

Although there is some dissent here (e.g. Burge 2010), a widespread view in the philosophical literature is that we need a "naturalistic" theory of content (Fodor 1987; Millikan 1984; Ryder 2009b). That is, it is widely assumed that content cannot be an explanatory primitive within cognitive theory. Instead, it must be explained in terms of properties and relations of the sort countenanced by the natural sciences. As Sterelny (1990, p.111) puts it,

> "Mental properties are ultimately to be explained by the non-mental; by more fundamental, because more general, features of our world. The representational theory of mind isn't committed to any vulgar reductivism. But representational properties cannot be unexplained primitives in our total theory of the world."

As we will see below, the "explanation" here is typically understood to consist of an extensionally equivalent translation of a set of statements describing the contents of some proprietary set of representational vehicles into a set of naturalistic statements that do not include any semantic concepts (Fodor 1987). The philosophical literature contains many "theories of content" in this sense: causal theories (Fodor 1984), co-variance theories (Dretske 1988; Fodor 1987), functional role theories (Block 1986), teleological theories (Millikan 1984),

and so on (see Ryder 2009b). Each of these theories seeks to identify those features of the natural world responsible for imbuing mental representations with their contents.

For reasons that will become clear, I will not provide a summary of this enormous literature here. Instead, I want to step back and ask a prior question: what do we want from a theory of representational content to begin with? What constraints are there on such a theory? What counts as success?

(2.)**5. Constraints on a Theory of Mental Representation**

There is no great mystery concerning what a theory of representational format should achieve: it should correctly describe those properties of representational vehicles relevant to their systemic roles within the brain. Of course, this has proven remarkably difficult to do, with constraints on such theories spanning low-level details of neural circuitry to high-level behavioural phenomena. Further, it is a topic on which empirically-minded philosophers have made substantial contributions. Nevertheless, the project is an ordinary scientific one and answerable in principle to experimental evidence from a range of different sources.

When it comes to theories of representational content, however, things are much murkier. In this domain the bulk of theorising has been undertaken primarily by philosophers. Further, although the philosophical literature contains an enormous number of "theories" of representational content alongside sharp disagreements concerning which of these theories is the most promising, the theories in question are not generally tested against experimental evidence or indeed against anything recognisable as empirical evidence at all. Given this, what are they tested against? What are these theories trying to achieve? As Stich (1992, p.243) puts it,

> "While there is no shortage of debate about the merits and demerits of various accounts of mental content, there has been remarkably little discussion about what a theory of mental representation is supposed to do: What question (or questions) is a theory of mental representation supposed to answer? And what would count as getting the answer right?"

I think it is important to distinguish between two different answers to these questions in the philosophical literature. The first answer appeals to the project of what I will call (following Fodor 1987) "naturalistic psychosemantics."

(2.)**5.1. Naturalistic Psychosemantics**

I noted above that a theory of representational content seeks to explain the contents of mental representations in naturalistic terms. This raises the question: the contents of *which* mental representations? According to the project of naturalistic psychosemantics, the mental representations in question are our *propositional attitudes*: mental states such as beliefs, desires, and intentions that form the basic categories of our folk psychological understanding of ourselves (Fodor 1987). As Warfield and Stich (1994, p.4) write,

> "Mental representations are theoretical postulates invoked by philosophers and cognitive scientists in an attempt to analyse or explain propositional attitudes, such as belief and desire, *that play a central role in commonsense psychology*" (my emphasis).

A theory of representational content in this sense should identify those properties and relations in virtue of which our propositional attitudes mean what they do. Given the widespread view that the contents of our propositional attitudes are compositional functions of the contents of our concepts, the challenge becomes one of explaining in naturalistic terms what determines the mapping from our concepts (e.g. DOG, CAT, TABLE, HUMAN) to those individuals, properties, and kinds in the world that they refer to (Fodor 1987; Sterelny 1990).

I will return in more detail to this project in Chapter 8. For now, I want to consider the two questions raised by Stich above: what question does naturalistic psychosemantics try to answer? What would count as getting the answer right?

As we have seen, the question it seeks to answer is this: is it possible to translate a set of statements describing the semantic properties of our folk psychological states into a set of naturalistically kosher statements that are free of semantic concepts? As such, the conditions of success are transparent: a theory of representational content must provide an adequate reductive translation. Crucially, because the relevant semantic properties being reduced are those ascribed to mental states by folk psychology, philosophers do not need to leave the armchair to evaluate such theories. For example, they can rule out theories that do not imply highly determinate contents because we *know* that folk psychological states have highly determinate contents (Millikan 1984). They can rule out theories that imply that concepts are not shared between people because we *know* that people share the same concepts (Fodor 1987). And they can rule out theories that suggest that a fully functioning human being spontaneously constructed from the fortuitous arrangement of matter produced by a lightning strike on a swamp would not mentally represent anything because we *know* that such a "swamp man" would have mental representations (see Ryder 2009b for a review).

As Stich (1992) points out, although this research programme is called "naturalistic" psychosemantics, its basic methodology reflects an analytical project in philosophy that goes back at least to Plato and Socrates. In this project, the aim is to produce extensionally equivalent non-circular definitions of our folk concepts. The chief difference is that within naturalistic psychosemantics the definitions can only include terms countenanced by "metaphysical naturalism."

(2.)**5.2. Abandoning Commonsense Constraints**

I think that it is a fundamental mistake to identify this project with the project of providing a theory of representational content (see Churchland 2012; Cummins 1989; Stich 1992). There are two reasons for this.

First, there is an enormous amount of research within the contemporary cognitive sciences on mental representation that has nothing to do with identifying the truth-makers for propositional attitude ascriptions (Bechtel 2008; Cummins 1989). These include mental representations such as the mental models first posited by Craik (1943) that we will explore in the next chapter, the cognitive maps posited by Tolman (1948), the symbolic data structures involved in parsing the syntactic structure of linguistic expressions central to work in generative linguistics (Chomsky 1980), the spreading activation patterns in high-dimensional neural networks central to work in machine learning (Rumelhart & McClelland 1986), the causal Bayesian networks central to a large body of work within artificial intelligence (Russell & Norvig 2010), and so on. Fodor (1985, p.78) writes that "the full-blown Representational Theory of Mind… purports to explain how there *could be* states that have the semantical and causal properties that propositional attitudes are commonsensically supposed to have." None of the representational structures just outlined were posited to account for the truth of propositional attitude ascriptions, however. Focusing on the propositional attitudes posited within folk psychology is thus parochial at best (Bechtel 2008, p.169).

Second, it is not obvious why *any* theory of representational content should be constrained by folk psychological and semantic intuitions (Cummins 1989; Stich 1992). In general, we do not hold scientific research answerable to its capacity to integrate commonsense intuitions. In this sense the project of naturalistic psychosemantics is essentially a project of *metaphysical domestication* (Ladyman et al 2014): it starts with the prejudice that the basic structure of propositional attitude psychology and the network of semantic intuitions that go along with it are correct and then seeks a home for them in our metaphysics. Even setting aside the fact that

this project is widely regarded as a failure (Hutto & Satne 2015), it is a strange project to begin with. Why *should* our science of the mind/brain be hospitable to folk psychological or semantic intuitions?

In what follows I will assume that there is no reason to think that it should be. As Stich (1992, pp.251-2) puts it, a theory of mental representation should not "much care about the commonsense conception of mental representation. The intuitions and tacit knowledge of the man or woman in the street are quite irrelevant."

Given this, what should a naturalistic theory of mental representation try to achieve? Before answering this question, I want to briefly address two likely objections.

First, proponents of naturalistic psychosemantics might argue that a theory of mental representation must accommodate our folk psychological and semantic intuitions not *because* they are our intuitions but rather because we have independent empirical reason to believe these intuitions are broadly correct (Fodor 1987; Sterelny 1990). Interpreted in this way, this motivation for naturalistic psychosemantics is consonant with the perspective I am defending here. Nevertheless, the only way to evaluate this claim is by turning to our most promising cognitive science.

Second, I argued above that our answer to the Constitutive Question should be constrained by our ordinary understanding of representation. I am now arguing that *theories* of mental representation should not be so constrained. Isn't there a tension here? If we should defer to science when it comes to characterising the nature of mental representations, why should we not also defer to science in determining what conditions something must satisfy to qualify for representational status in the first place?

The tension here is illusory. In saying that our answer to the Conceptual Question should be constrained by our ordinary understanding of external representations, I am proposing that a useful way of understanding the representational theory of mind—one which honours how the theory has historically been understood and renders it both substantive and interesting—is as the claim that psychological mechanisms implicate functional analogues of external representations. Just as the *truth* of this theory should be exclusively settled by science, however, the *nature* of mental representations—if there are such things—should also be an exclusively scientific issue. For an analogy, consider the hypothesis that a biological structure functions as a pump (see Ramsey 2007). This claim draws on our ordinary understanding of pumps; if the structure does not conform in some important way to this functional profile, it is

simply not a pump. Nevertheless, the *truth* of the theory and the characteristics of the structure that occupies this functional profile if the theory is true are scientific matters.

(2.)**5.3. A Naturalistic Theory of Mental Representation**

A theory of representational content should not be answerable to folk psychological or semantic intuitions. What constraints are there on such a theory, then? As Cummins (1989, p.16) puts it, "once you abandon… the idea that the theory of mental representation must yield contents for intentional states [i.e. the propositional attitudes], you *need* a few constraints." I will follow Cummins in his recommendation for a useful methodology for the philosophy of mental representation:

> "first determine what explanatory role representation plays in some particular representation-invoking scientific theory or theoretical framework; then ask what representation has to be—how it is to be explicated—if it is to play that role" (Cummins 1989, p.145).

To expand on this, I think that a theory of mental representation should satisfy the following two basic constraints.

First, it should be exclusively motivated by our best science. Of course, reasonable people can disagree at this point concerning what our most promising scientific research programmes are, but any theory of mental representation should be motivated by cognitive science, not commonsense intuitions. Specifically, it should take the following form: this is how mental representation should be understood *if* such and such a theory or research programme is correct. As such, there is no guarantee that the theory will be hospitable to folk psychology. Indeed, it would be surprising if cognitive science were not surprising (Ladyman et al. 2014). In this way a theory of mental representation

> "seeks to say what mental representation really is, not what folk psychology takes it to be. And to do this it must describe, and perhaps patch up, the notion of mental representation as it is used by the best cognitive science we have available" (Stich 1992, p.252).

The second constraint is really a consequence of the first, but it is worth making it explicit. It is that a theory of mental representation should reveal how the postulation of mental representations within a given theory or research programme contributes to the genuine *explanation* of the relevant psychological capacities. For now, I want to leave open the question of how exactly explanation should be understood except to stress a basic causal constraint: the representational properties of mental representations must be *causally* relevant to their systemic

roles (see Chapter 5). Cognitive scientists are interested in mental representations because they are supposed to perform distinctive jobs within cognitive systems. To make sense of this, they must play a role within the causal order. A central job for a theory of mental representation is thus to explain what distinctive roles they play and how they play them—what *difference* they make to cognitive systems that harbour them (Cummins 1989).

(2.)**6. Conclusion**

Questions concerning the existence, importance, and nature of mental representations in psychological mechanisms have generated some of the deepest and most enduring controversies concerning how the mind works. In this chapter I have argued that a frustrating characteristic of these controversies is the lack of any consensus on two basic questions: first, what are the distinctive properties and relations in virtue of which something qualifies as a mental representation in the first place? Second, what should a successful theory of mental representation achieve? I argued that standard answers to these questions in the philosophical literature are implausible and I motivated two claims: that mental representations should be understood as mental analogues of public representations, and that a theory of mental representation should be exclusively answerable to the explanatory constraints of our best cognitive science.

With these clarifications in mind, I now turn to providing an account of mental representation that honours these answers.

Chapter 3: Craik's Hypothesis on the Nature of Thought

> "So we may try to formulate a hypothesis concerning the nature of thought and reality, see whether any evidence at present available seems to support it, and wait to find whether the evidence gained in future gives any support to it or to slightly modified forms of it" (Craik 1943, p.47).

(3.)**1. Introduction**

In this chapter I draw on the work of the mid-twentieth century psychologist, philosopher, and control theorist Kenneth Craik. I argue that one can extract from Craik's work an attractive—albeit highly schematic—proto-theory of mental representation founded on three core insights that I will elaborate and defend in later chapters: first, an account of mental representation in terms of neural models that capitalize on structural similarity to their targets; second, an appreciation of prediction as the core function of such models; and third, a regulatory understanding of brain function.

Kenneth Craik studied philosophy and psychology as an undergraduate at Edinburgh University before moving to the University of Cambridge in the late 1930s.[5] Before he arrived there, his supervisor at Edinburgh told Sir Frederic Bartlett, the first professor of experimental psychology at Cambridge, "Next term I am going to send you a genius!" (Zangwill 1980, p.2). At Cambridge, Craik did pioneering theoretical and experimental work on visual psychology and physiology, winning a Fellowship at St John's College in 1941. During the war, he worked primarily as an applied research scientist contributing to the design of military technology such as gun predictors and tracking devices—a contribution that earned him the status of the first Director of the Medical Research Council's Applied Psychology Unit in 1944. His life was cut short at the age of 31 less than a year later in a cycling accident two days before VE day. Of Craik's intellectual brilliance, Zangwill (1980, p.11) writes:

> "Just as his Professors considered him to be the best student who had ever come their way, so did his colleagues regard him as a man of altogether outstanding intellectual distinction. One of his Cambridge contemporaries spontaneously remarked, some 30 years after his death, that Kenneth was the greatest man he had ever met."

Most of Craik's research was published posthumously, except for a short monograph, "The Nature of Explanation," which I will draw from below. In the immediate decades after his

[5] Here I draw on Gregory (1983) and Zangwill (1980).

death, his ideas were largely ignored (Gregory 1983). Although certain historically-minded cognitive scientists have since highlighted the prescience of Craik's work (e.g. Boden 2006; Russell & Norvig 2010), it is rarely considered in any detail, and it has remained effectively invisible in philosophy.[6]

Given this relative obscurity, why am I devoting a chapter of this thesis to Craik's work? In the last chapter I argued that a theory of mental representation should be exclusively answerable to the explanatory constraints of our best cognitive science. Turning back to the views of an obscure psychologist from the early 1940s might therefore seem strange.

Nevertheless, there are at least three reasons for including Craik's work in this thesis.

First, Craik is an important figure to consider in the context of the contemporary representation wars. Writing in a period of maximal scepticism about the concept of mental representation both in psychology (Skinner 1938; Watson 1913) and philosophy (Ryle 1949; Wittgenstein 1953), Craik's hypothesis was deeply heretical in the intellectual climate in which he advanced it (Gregory 1983). It is interesting, then, and—I will argue—*instructive* to consider the motivation for this heresy in a context in which scepticism about the importance of internal representations in cognitive processes is on the rise once again (Anderson 2014; Chemero 2009; Gallagher 2017; Hutto & Myin 2012; Ramsey 2007; Wilson & Golonka 2013).

Second, recent years have seen an explosion of work in both philosophy and the cognitive sciences that rediscovers many of Craik's insights: the centrality of prediction to intelligence (Barr 2011; Bubic et al. 2010; Llinás 2001); an understanding of mental representation in terms of structural similarity (Cummins 1989; Grush 2004; Ryder 2004); and an increased focus on regulation and predictive control as principles of brain functioning (L.F.Barrett 2017; Sterling & Laughlin 2015). Nowhere is this more evident than in predictive processing (Clark 2016; Friston 2005; Friston et al. 2017; Hohwy 2013), a theory in cognitive neuroscience that I will draw heavily from in forthcoming chapters. Craik provides an especially insightful—but largely unappreciated—historical precursor to this efflorescence of recent work on the "predictive mind" (Williams 2018d).

Finally, I believe that Craik was simply correct in his overall account of mental representation. That is, I will argue in forthcoming chapters that an understanding of mental representation founded on neurally realised models that recapitulate the structure of worldly domains in the

[6] Notable exceptions to this are Horst (2016) and McGinn (1991). Grush (2004) and Ryder (2004) also both mention Craik as influences although they do not describe his work.

service of prediction is theoretically attractive and empirically well-motivated in light of recent advances in contemporary cognitive science.

I proceed as follows. In Section 2 I briefly outline the historical context and aims of Craik's work. In Section 3 I provide a concise summary of Craik's "hypothesis on the nature of thought." Sections 4, 5, and 6 then extract three ideas from Craik's work: first, an account of mental representation in terms of neural models that capitalize on structural similarity to their targets (S4); second, an appreciation of prediction as the core function of such models (S5); and third, a regulatory understanding of brain function (S6). I conclude in Section 7 by summarising the foregoing lessons and outlining the aims and contents of future chapters.

(3.)**2. Historical Background**

As the title suggests, Craik's overarching aim in "The Nature of Explanation" is to provide an account of the nature and function of explanation. Although the work covers a variety of different topics related to this aim, the book can be decomposed into two main parts: first, Craik (1943, p.120) seeks to defend what he takes to be the commonsense view—the view held by "the man in the street"—that explanation "means giving the causes of things and saying why they happen" within an objective world of "external lasting things"; second, he seeks to advance a "hypothesis on the nature of thought" motivated by this commonsense account of the nature and function of explanation.

Craik's approach is fiercely naturalistic. The preface of the book bemoans the lack of progress in philosophy, which Craik (1943) attributes both to a fruitless "a priorism" which has historically characterised the discipline (p.9) and to the remnants of this a priorism in "the verbal and formal precision of the symbolic logicians" (p.120) of his day. In place of such methodologies, Craik (1943, p.i) suggests that "some definite contribution to the solution of philosophical problems may come from the application of the various experimental methods which have advanced the sciences."

Craik's hypothesis on the nature of thought develops his broader views about the nature of explanation. The monograph opens with the question, "What do we mean by "explaining" anything?" (Craik 1943, p.6). Craik's answer sets the foundation for all that follows: a "man who can explain a phenomenon," he writes, "can *predict it*, and *utilise it* more than other men," such that "the power to explain involves the power of insight and *anticipation*" (Craik 1943, p.7, my emphasis). In this way an explanation can be understood as "a kind of *distance-receptor* in time, which enables organisms to adapt themselves to situations which are about to arise"

(Craik 1943, p.7, my emphasis). This account of explanation leads Craik to propose a theory of intelligence, according to which the predictive abilities of certain animals are grounded in neural models.

Before elaborating on this proposal, it is worth reviewing just how heretical it was in the scientific and philosophical context in which Craik advanced it.

In psychology, the dominant (although by no means exclusive) research programme was *scientific behaviourism* as practiced by figures like John Watson and B.F. Skinner.[7] For convenience, this research programme can be characterised in terms of a commitment to the view that all animal behaviour can be accounted for in terms of habituation, classical and operant condition, and the associative mechanisms inside animals' nervous systems that mediate such conditioning (Skinner 1938; Watson 1913; see Rey 1997 for a review).[8] As noted in the previous chapter, it is often said that what most sharply differentiates this approach to intelligent behaviour from that which superseded it was its wholesale rejection of mental representations in psychological explanations (Bechtel 2008; Ramsey 2007). As Gallistel (2001, p.9691) puts it, "mental representations were banned by the behaviorists." Why?

Although it is something of a simplification, I think that one can identify the following three general sources of scepticism as especially important.

The first was a simple consequence of its account of the mechanisms underlying behaviour: if all behaviour is a function of habituation and conditioning, mental representations are unnecessary (see Section 5 below). This account was itself partly explained by broader theoretical commitments, however. Probably the most important of these—the second source of scepticism—was the conviction that content-bearing mental states are somehow scientifically disreputable. The arguments for this claim were heterogenous and not always coherent (Dennett 1978, pp.53-70). The core idea, however, was that the broadly positivist methodology underlying behaviourism required the elimination of publicly unobservable entities and sources of justification such as folk psychology and introspection as guides to the mechanisms underlying intelligence (Skinner 1938). Given the conviction among the behaviourists that the postulation of mental representations falls foul of these methodological

[7] Although scientific behaviourism's influence was chiefly in North America (see Bechtel 2008), Craik himself had extensive knowledge of—and was deeply influenced by—this work (see especially Craik 1966; Zangwill 1980).

[8] Note that describing behaviourism in this way is highly idealised and slightly distortive, given the heterogeneity of work that fell within that tradition (see Ramsey 2007).

strictures, they were thus one of the most conspicuous casualties of behaviourism's psychological approach (Gallistel 2001).

These two considerations interacted with a third source of scepticism, according to which populating the mind or brain with mental representations invites either regress or vacuity. Specifically, because representations require *interpretation*, mental representations would require in-the-head agents to use and interpret them. These homunculi, however, would have to be as intelligent as the very behaviours that psychologists were trying to explain. As such, positing mental representations merely creates the *illusion* of explanation. As Skinner (1971, p.200) put it, "Science does not dehumanize man; it *de-homunculizes* him… Only by dispossessing him can we turn to the *real* causes of behavior" (my emphasis). More sophisticated forms of this general worry were advanced by both Ryle (1949) and Wittgenstein (1953).

Craik must have been aware of these critiques of mental representation. He was a trained psychologist and philosopher at the University of Cambridge where Wittgenstein's influence dominated, and his early research was greatly inspired by Thorndike's work on trial-and-error learning (Zangwill 1980; see Craik 1966). Nevertheless, he also evidently thought that non-representational psychology is bereft of the theoretical tools to explain intelligence—that postulating mental representations is not just conceptually coherent but positively necessary for the explanation of our psychological capacities. To understand why Craik thought these things, one must turn to his "hypothesis on the nature of thought" (Craik 1943, p.61).

(3.)**3. Craik's Hypothesis on the Nature of Thought**

In a slogan, Craik's hypothesis is this: the mind is a *predictive modelling engine*. Specifically, Craik (1943) argues that "one of the most fundamental properties of thought is its power of predicting events," which gives it its "immense adaptive and constructive significance" (p.50), and which is made possible because an organism "carries a 'small-scale model' of external reality and of its own possible actions within its head" (p.61)—a small-scale model "which has a similar relation-structure to that of the process it imitates" (p.51).

In an oft-quoted passage, he speculates:

> "If the organism carries a "small-scale model" of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with

> the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it" (Craik 1943, p.61).

Craik's (1943, p.90) hypothesis is not motivated by a priori reflection or conceptual analysis but from a naturalistic perspective in which "adaptation to environment [is] one of the most fundamental principles of behaviour," a conviction he likely inherited from William James and John Dewey (and of course Darwin). That is, he posits a process of model-based prediction as the best explanation of the kinds of adaptations to our environment exhibited by both human beings and "the thinking of [other] animals," which, Craik (1943, p.59) argues, reflects "on a more restricted scale the ability to represent."

There are three features of Craik's work that will be important in what follows: first, an account of mental representation in terms of neural models that capitalize on structural similarity to their targets; second, an appreciation of prediction as the core function of such mental models; and third, a regulatory understanding of brain function. I will consider them in turn.

(3.)**4. Mental Models**

The first of three important ideas that I will extract from Craik's work is this:

> (Insight #1) Mental representation should be understood in terms of neural models that capitalize on structural similarity to their targets.

To unpack this, it is important to first note that Craik (1943, p.49) assumes a strong form of materialism as the appropriate methodology for psychology, making a "plea for physical explanation" of psychological capacities. If there are mental representations, then, they must be components of neural mechanisms. What kind of components? Craik's answer embodies the understanding of mental representation defended in the previous chapter: they will be neural structures that perform similar functional roles to public representations—specifically, the public representations that help us to *predict* and *control* our environments. As Craik (1943, p.61) puts it,

> "Most of the great advances of modern technology have been instruments which extended the scope of our sense-organs, our brains or our limbs. Such are telescopes and microscopes, wireless, calculating machines, typewriters, motor cars, ships and aeroplanes. Is it not possible, therefore, that our brains themselves *utilise comparable mechanisms to achieve the same ends and that these mechanisms can parallel phenomena in the external world as a calculating machine can parallel the development of strains in a bridge*?" (my emphasis).

Craik thus reasons by analogy. First, he argues that we successfully predict and control our environments effectively by the construction and manipulation of external representations such as "calculating machines" (see below)—that this is the "*one* way that 'works'… with which we are familiar in the physical sciences" (Craik 1943, p.53). Second, he then concludes from the importance of prediction and predictive control to adaptive behaviour more generally (see below) that neural mechanisms must make use of structures that function "comparably" to such public representations—that "it does not seem overbold to consider whether the *brain* does not work in this way—that it imitates or models external processes" (Craik 1943, p.53, my emphasis). As Gregory (1983, p.236) puts it, "Craik supposes that hundreds of millions of years ago brains incorporated technologies we have invented over the last few centuries."

In one sense, there is of course nothing original about this part of Craik's approach. The idea of understanding mental goings-on by analogy to public artefacts or technologies of the time is plausibly as old as philosophy itself. Indeed, Aristotle's (*De Anima* [1984]) suggestion that we understand the mind-world relation by analogy with the imprint that a signet ring leaves on wax to mark the provenance of a letter is a perfect—if extremely primitive—example of this approach. Further, at first pass this suggestion looks hopeless as a response to the conceptual worries about mental representation raised above: don't public representations essentially require *interpretation*? As such, how can positing in-the-head analogues of public representations help?

To understand Craik's (implicit) answer to this question, one must turn to the *kind* of public representation that he draws on as the appropriate analogy for understanding mental representation: *physical models*.

(3.)**4.1. Models and Structural Similarity**

For Craik (1943, p.51), a model is just "any physical or chemical system which has a similar *relation-structure* to that of the process it imitates." Further, this appeal to a similar relation-structure is intended to be understood in a straightforwardly naturalistic way:

> "By "relation-structure" I do not mean some obscure non-physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the process it parallels…Thus, the model need not resemble the real object pictorially; Kelvins' tide-predictor, which consists of a number of pulleys on levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects…" (Craik 1943, p.51).

Craik's appeal to a shared "relation-structure" between a model and its target is supposed to capture the fact that they resemble one another not with respect to their first-order properties (e.g. colour or size) but rather with respect to their *relational organisation*. The idea that models in general are representations that capitalize on structural similarity of this kind is a popular view that I will be returning to at greater length in future chapters. Although Craik does refer to ordinary scale models to illustrate his argument—for example, a scale model of the *Queen Mary*—the principal kind of models he appeals to are those exploited by the analogue computing devices of his time. Lord Kelvin's tide-predicting machine, for example, was a widely used special-purpose analogue computer designed in the late 19th century that harnessed a complex system of pulleys and wires to predict the ebb and flow of sea tides, combining "oscillations of various frequencies so as to produce an oscillation which closely resembles in amplitude at each moment the variation in tide level at any place" (Craik 1943, pp.51-2). Likewise, the "calculating machine" referred to above was Bush's "differential analyser," the world's first general-purpose analogue computer, designed to solve differential equations using wheel-and-disc mechanisms (O'Brien & Opie 2010).

The specific (and often highly complex) details of these mechanical devices need not concern us. What such analogue mechanisms have in common, however, and what obviously drew Craik to them, is "that the components of each computer… are assembled to permit the computer to perform as a model, or in a manner analogous to some other physical system" (Truit & Rogers 1960, p.3; quoted in O'Brien & Opie 2010, p.3). Specifically, they are designed to generate predictions of the behaviour of complex dynamical systems by in effect *simulating* their dynamics through the clever arrangement of specific material components. As Craik (1943, p.51) puts it, such devices involve

> "the 'translation' of the external processes into their representatives (positions of gears, etc.) in the model; the arrival at other positions of gears, etc., by mechanical processes in the instrument; and finally, the retranslation of these into physical processes of the original type."

The differential analyser, for example, "consists of a collection of basic building blocks which can be interconnected so that they are governed by the same set of equations as those describing the system to be analyzed" (Smith & Wood 1959, p.1; quoted in O'Brien & Opie 2008, p.60). Such systems are thus "analogue" devices in that they employ functional *analogues* of target systems: cleverly arranged material parts and operations that can be exploited for mimicking and thus *predicting* the behaviour of such systems (Craik 1943, 53; O'Brien & Opie 2010).

For Craik, the significance of such analogue systems is that they reveal how a physical machine can be configured to mirror the relational structure of a target domain *without instantiating the same material properties that sustain that structure*. For this reason, Craik concludes that there is no in-principle reason why a complex physical machine like the *brain* could not function in a similar way, installing models that generate predictive simulations of processes in the world that the organism must interact with. He asks, "what structures and processes are required in a mechanical system to enable it to imitate correctly and to predict external processes…?" (Craik 1943, p.57) and argues—convincingly—that there is no *a priori* reason why the brain could not possess such structures and processes. "It is *possible*," Craik (1943, p.115) writes, "that a brain consisting of randomly connected impression synapses would assume the required degree of orderliness [for model-based prediction] as a result of experience" (my emphasis).

In this sense Craik was an important pre-cursor to the structuralist account of mental representation in both philosophy and psychology that I will return to in future chapters. However, it is important to stress that for Craik the appeal to structural similarity between neurally realised models and target domains is not intended to answer traditional philosophical concerns about meaning or content determination (see Chapter 2, Section 5). Instead, it features as part of a proto-scientific explanation of specific *capacities* that intelligent organisms exhibit (see below). As Horst (2016, p.165) puts it,

> "Craik's interest…seems to have been not with semantics… but with adaptive cognition: if structural similarities between a model and the system modelled are important, it is because this allows the model to function as a mental surrogate for the target system."

(3.)**4.2. Rehabilitating Representationalism**

Above I noted three sources of anti-representationalism in the behaviourist tradition: first, that the mechanisms of classical and operant conditioning that underlie behaviour do not require mental representations; second, that support for the existence of mental representations derives entirely from dubious practices of introspection and folk psychology; and third, that there is something conceptually problematic about appeals to mental representations—that such appeals invite regress or vacuity.

With respect to the first of these claims, Craik would have agreed. Where he disagreed, however, was in the assumption that simple processes of conditioning are adequate to account for intelligence. I address this point in the next section.

The second claim—that mental representations arise only from dubious folk practices of introspection and interpretation—is easily rejected by Craik. Specifically, he acknowledges that his hypothesis on the nature of thought might be mistaken, but he is adamant that it *is* a scientific hypothesis. According to Craik, certain organisms exhibit certain capacities—most fundamentally, *foresight* (see below)—and the best explanation of such capacities is that they have structural models of the world in their nervous systems. This claim evidently does not rely on introspection, a priori insight into the nature of reality, or conceptual analysis, as Craik is eager to point out. It might be *wrong*, but that is an empirical issue.

Perhaps most importantly, Craik's account of mental representation also seems to be immune to the objection that mental representations would require intelligent agents to use and interpret them. Implicit in Craik's argument is the claim that one can isolate the core functional profile of model-based prediction in a way that abstracts away from issues concerning interpretation and intelligence. This core functional profile, according to Craik, is the exploitation of structural similarity between a model and some target domain to guide prediction and control. Although in the public case this activity is typically associated with intelligence and interpretation, Craik's argument is that there is no *a priori* reason this must be the case.

For these reasons, Craik's "hypothesis on the nature of thought" provides a genuinely novel (in his day) and attractive approach to understanding mental representation that was not vulnerable to the criticisms of representationalist accounts of the mind advanced by the behaviourists. Nevertheless, suppose one grants that there *could* be structures inside the head that re-present the structure of worldly domains in this way. What is the value of such models? Why should we think that they play the extensive role in intelligence that Craik alleges?

Craik (1943, pp.81-2) is acutely sensitive to this worry:

> "Some may object that this reduces thought to a mere "copy" of reality and that we ought not to want such an internal "copy"; are not electrons, causally interacting, good enough?"

His answer is simple:

> "The answer, I think, is…that only this internal model of reality—this working model—enables us to predict events which have not yet occurred in the physical world, a process which saves time, expense, and even life" (Craik 1943, p.82).

This brings us to the second idea that I will extract from Craik's work.

(3.)**5. The Predictive Mind**

Craik's claim that the brain commands small-scale models of external reality is motivated by a commitment to the importance of prediction to intelligence. Specifically, Craik (1943, p.50) argues that "one of the most fundamental properties of thought is its power of predicting events," which gives it its "immense adaptive and constructive significance," an insight he credits to "Dewey and other pragmatists" (my emphasis). This predictive capacity, he notes, is "not unique to minds, though no doubt it is hard to imitate the flexibility and versality of mental prediction" (Craik 1943, p.51). Given that such flexible predictive capacities in science and engineering are subserved by models of target domains—or, as Craik (1943, p.53) puts it, that "this is one way that "works," in fact the only way with which we are familiar in the physical sciences"—he concludes that the predictive capacities exhibited by intelligent animals are themselves underpinned by models inside their brains. The second insight that I will extract from Craik's work, then, is this:

(Insight #2) The core function of mental models is prediction.

This emphasis on prediction was at odds with the behaviourism of his day (see below). Nevertheless, as Craik (1943, p.50) himself points out (referencing Dewey), it was not unique to him. What plausibly was original to him, however, was the use of the analogy with prediction in science and engineering to propose the hypothesis that such predictive capacities are underpinned by mental models (Boden 2006, p.210; Horst 2016).

One aspect of this inference is a conditional that seems reasonable: *if* there are mental models of target domains inside our heads, we have at least a very schematic explanation of our capacity to predict the behaviour of such domains. This link between predictive success and the structural correspondence between models and targets is a theme that I will return to at greater length in future chapters. There are two other related features of Craik's views here that are much less transparent, however. First, why did he think that prediction *is* so central to adaptive success? Second, even if mental models would provide an explanation of predictive capacities, why think that they provide the *only* or even *best* explanation of such capacities? Craik seems to want to draw a fairly direct inference from prediction to mental representation. How plausible is this inference?

It will be instructive to consider these two issues in reverse order.

(3.)**5.1. Prediction and Representation**

How plausible is the inference from the predictive success (allegedly) exhibited by intelligent animals to the hypothesis that they house mental models inside their heads? Superficially, at least, there is something compelling about it. Non-representational frameworks in psychology have often been associated with highly *reactive* accounts of intelligence. In behaviourism, for example, an organism's behaviour is a function of previous associations between stimuli, responses, and outcomes (rewards and punishments), such that the locus of behavioural control is effectively the environment itself (Skinner 1938). Further, as we will see in Chapter 8, non-representational movements in the late 20th century of the sort associated with Rodney Brooks (1999) typically advanced what Anderson (2003, p.97) calls "highly reactive models of perception, bypassing the representational level altogether."

In addition, one might bolster Craik's argument by appeal to more recent philosophical considerations. An influential idea in the literature, for example, is that "representation-hungry" (Clark & Toribio 1994) tasks are tasks in which an intelligent system must coordinate its choice or behaviour with states of affairs to which its access is restricted, either because they are spatiotemporally distant or because the relevant phenomena are abstract or non-existent. *Future* states of affairs seem to fall into this category: an intelligent system cannot interact "directly" with such events. If a system must coordinate its behaviour with predicted outcomes, then, it is natural to think that it must exploit some internal proxy for the domain generating such outcomes.

Despite such attractions, I think that any simple inference from the role of prediction in mental life to internal modelling is untenable.

First, it is crucial to distinguish predictions as the sorts of things that people consciously engage in—that is, consciously held truth-evaluable judgements about future events—from predictive or anticipatory capacities more generally (Anderson & Chemero 2013). In the first case, the idea that prediction relies on internal representation is trivial; after all, the prediction itself plausibly just is a representation. In the current context this point is important. Craik's aim is to generalise from the personal-level case to the sub-personal mechanisms inside our heads which underlie a wide range of adaptive behaviours. If this is right, it must be the latter kind of prediction—predictive capacities in general—that he has in mind as the relevant motivation for positing internal models. But—at least if such predictive capacities simply involve tracking correlations—it is not obvious that they *do* require representation (Anderson & Chemero 2013).

To see this, consider a second problem for any simple inference from prediction to representation: there is a perfectly straightforward sense in which the anti-representationalist behaviourist approach of Craik's day could explain predictive capacities. In classical conditioning, for example, a dog might learn to associate the sound of a bell with the presence of food. After a while, we might say that she has acquired predictive knowledge: upon hearing the bell, she "predicts" food. In this sense the very "usefulness of classical conditioning is predictive" (Godfrey-Smith 2018, p.227). Likewise with operant conditioning: your dog gives you a cuddle when you get home and you give her a treat; she barks at the neighbour in the morning and you gently tap her on the nose. After a while she learns what outcomes—rewards and punishments—her behaviour "predicts." Given this, does the dog also acquire a "small-scale model" of external reality and of her possible actions? One can imagine trying to capture this "knowledge" in terms of a model in which the values of one variable (<food> or <cuddle>) systematically covary with the values of another (<bell> or <reward>). We might say that this model "recapitulates" the relations among worldly features in a manner that can guide action: when the dog hears the bell, it prepares itself for the intake of food.

Clearly, however, this way of talking is grossly misleading. It conveys no additional explanatory advantage at all. The dog's behaviour has simply become sensitive to correlations between previous stimuli, responses, and outcomes.

(3.)**5.2. The Link Between Prediction and Modelling**

This challenge helps to bring out something important about Craik's understanding of prediction, which is this: he infers the presence of internal models from a flexible predictive capacity that goes far beyond the "predictions" facilitated by classical and operant conditioning just outlined. In previous work (Williams 2018d) I have called this more substantial predictive capacity "counterfactual competence." A system that exhibits counterfactual competence is not stuck exclusively in a past history of associations linking stimuli, responses, and outcomes (rewards and punishments). Rather, it models the causal structure of the world such that it can generate and act on predictions of the outcomes of a range of possible interventions upon it, even if these interventions have not occurred before.

This reading is supported both by the analogies that Craik draws upon and his discussion of the kinds of predictive capacities that intelligent agents exhibit.

With respect to the former, for example, Craik's understanding of model-based prediction bears no resemblance to the kind of "predictive knowledge" engendered through conditioning just

outlined. Specifically, Craik argues that that it is distinctive of models used in engineering that they do not involve learning primarily through trial and error. As he puts it, building models enables "bridge design… to proceed on a cheaper and smaller scale than if each bridge in turn were built and tried by sending a train over it, to see whether it was sufficiently strong" (Craik 1943, p.52).

Further, it is this counterfactual competence that Craik thinks is central to thought. "In the particular case of our own nervous systems," he writes, "the reason why I regard them as modelling… is that they permit *trial of alternatives*" (Craik 1943, p.52, my emphasis). Specifically, our nervous systems confer upon us an ability not just to track relationships among stimuli and responses, but to run predictive simulations of the likely outcomes of worldly interventions, irrespective of whether these have been carried out before. This is clear, for example, in Craik's (1943, p.61) overview of the benefits that a small-scale model of external reality would confer: namely, capacities to "try out various alternatives" and "conclude which is the best of them."

On this reading, Craik's primary focus is on the capacity of organisms to engage in a kind of rational action-planning: to generate actions on the basis of predictions of their likely effects. For example, he draws an explicit distinction between two different kinds of purposive activity: "that which proceeds along any line which has been successful up to the present, though there may be no clear 'idea' of the goal ahead," and "that which proceeds towards a goal clearly envisaged" (Craik 1966, p.90). He writes that the latter "would seem to involve ideas or concepts or images of the ends to be achieved," and notes that "if these are to be considered as having a neural basis it must presumably take the form of 'models' of real activities and situations" (Craik 1966, p.90).

This interpretation thus connects Craik's work to the contemporary distinction often drawn between "model-free" and "model-based" reinforcement learning (Dayan & Daw 2008). It suggests that the principal difference he wanted to highlight was something like that between what Dennett (1995) calls "Skinnerian creatures" whose adaptive strategies are restricted to learning through trial and error (i.e. operant conditioning) and "Popperian creatures" who can test behavioural strategies "offline" by running internal experiments in their heads. In this way "the possessor of a nervous system is thus able to anticipate events instead of making invariable empirical trial" (Craik 1943, pp. 120-121, my emphasis added). As Dawkins (1976, p.78) puts the point,

> "Survival machines [i.e. biological systems] that can simulate the future are one jump ahead of survival machines who can only learn on the basis of overt trial and error. The trouble with overt trial is that it takes time and energy. The trouble with overt error is that it is often fatal."

Nevertheless, I now think that this reading captures only part of what Craik had in mind when he referred to the constructive and adaptive significance of prediction.

First, Craik thought that model-based prediction extends far more broadly into the animal kingdom and the roots of adaptive success than one would expect if its only function is the explicit modelling of the potential outcomes of one's possible interventions upon the world.

Second, some of the phenomena he appealed to are much broader than rational action-planning. For example, he notes a general capacity to "react to future situations before they arise" and "utilise the knowledge of past events in dealing with the present and future," and the technological analogies he draws on include predictive tracking mechanisms such as anti-aircraft gun predictors (Craik 1943, p.61). Likewise, he also notes that "neural or other mechanisms can imitate or parallel the behaviour and interaction of physical objects, and so supply us with information on physical processes which are not directly observable to us" (Craik 1943, p.99).

Finally, and relatedly, one important function of prediction that Craik (1966) emphasised in an unpublished book manuscript that he was yet to finish at the time of his death was its utility in overcoming signalling delays that would otherwise plague sensorimotor processing. This predictive capacity, he argues, involves

> "the "modelling", by the brain, of the sequence of events whose consequence is sought, so that this model may predict the answer earlier than it occurs in the course of external nature. In other words, by the ability to model… the brain is able to outrun the physical processes which are too rapid for it and so can, on the average, forestall and anticipate the course of nature" (Craik 1966, p.27).

For these reasons, I think it is clear that Craik thought the utility of mental prediction goes beyond its role in action-planning. Nevertheless, I will set this complex exegetical matter aside here. For my purposes, the aspect of Craik's work that is relevant is this: insofar as there are mental models inside our heads, their core function is not, say, pattern recognition or abstract logical reasoning but a kind of flexible predictive capacity that was not accounted for by the reactive mechanisms posited by the behaviourist movement of his day. As such, the structural correspondence between mental models and target domains highlighted above provides the fuel

for a specific kind of success: predictive success. This is an idea that will play a central role in future chapters.

First, however, I turn to a third important idea found in Craik's work.

(3.)**6. The Cybernetic Brain**

The analogies that Craik draws between science and what happens inside the head might suggest a commitment to an influential metaphor both in psychology and philosophy for understanding brain function, or at least that aspect of the brain involved in representing the world: *the brain as scientist* (see Gregory 1980; Hohwy 2013).

The origins of this metaphor can probably be traced back at least to Helmholtz (e.g. 1867)—an interesting figure to consider in the current context, as he is most often considered the grandfather of current interest in the idea of the "predictive mind" (see Friston 2005; Hohwy 2013).[9] For Helmholtz, the brain is in roughly the same position as a scientist seeking to infer the objective hidden state of the world from observations. The observations are the patterns of proximal stimulation the brain receives, and it must draw on such evidence and its prior knowledge to reliably infer their most probable causes—that is, the hidden structure of the world responsible for generating such data. In contemporary incarnations of this idea, it is suggested that the brain must effectively function like a hypothesis-testing machine, formulating hypotheses about states of the world and testing them against incoming evidence (Gregory 1980)—an idea we will return to at greater length in Chapter 6. As this brief overview presents, it is really a metaphor of the brain as an idealised impartial scientist—that is, a scientist concerned only with attaining the truth and evaluating evidence in a maximally impartial way, and thus a scientist with no counterpart in the actual world.

The brain as scientist metaphor has been highly influential and fruitful in guiding research. Despite this, many have argued that it presents an implausible vision of neural function (Anderson 2014; Bruineberg et al. 2016). Insofar as biological agents evolve from a process of natural selection, we should expect their behaviour to be guided—fundamentally, at least—by *practical*, not epistemic or inferential, functions (Tooby & Cosmides 2015; see Chapters 7 and 8). The brain is a metabolically expensive organ. This gives us good reason to expect what

[9] See Hohwy (2013, pp.5-8) for the history. Although Helmholtz is often credited as the grandfather of the *predictive* mind, his chief contribution seems to have been to stress the importance of knowledge-driven *inference*, which can be accommodated in purely feedforward mechanisms with no active prediction (see Marr 1982; Teufel & Nanay 2017).

elsewhere I have called (developing work from Akins 1996) "narcissistic nervous systems": nervous systems whose responsiveness to the environment is mediated entirely through the relevant organism's idiosyncratic practical interests (Williams 2018b; see Chapter 9).

Some have taken such considerations to weigh heavily against any vision of the model-building brain (e.g. Anderson 2014; Bruineberg et al. 2016). Craik evidently did not. He shares with the brain as scientist metaphor the notion that the brains of at least some organisms must genuinely model the structure of the ambient environment to function effectively within it. Nevertheless, it is also clear that he is sensitive to the fundamentally pragmatic function of these models. Craik (1943) recommends that we consider "the growth of symbolising power from… [considerations] of survival-value" (p.60) and argues that the brain's models should be answerable to norms of "cheapness, speed, and convenience" (p.52). Given this, he emphasises that the brain's model-use is "bound somewhere to break down…"—to fail to capture every aspect of reality with complete accuracy (Craik 1943, p.52). In this case, however, the brain's models are no different from many other models used in science, engineering, and elsewhere. They are *good enough* for the job but nevertheless selective and idealised in certain respects. As Craik (1966, p.68) puts it, "often the reproduction [i.e. model] is not identical with the original—it is intentionally different in order to be more convenient for the purpose in hand."

In a previous article (Williams 2018) I argued that a better metaphor than a scientist is an *engineer* when it comes to understanding Craik's approach to explaining brain function. This was supposed to capture the sense in which the brain's modelling activity is relative to specific practical goals, and not "representation for representation's sake" (Wilson 2002). As noted above, Craik himself was an engineer who made pioneering contributions to the design of military technology during the Second World War, and it is clear throughout his work that this background influenced his treatment of the strategies available to the brain (see Boden 2006; Gregory 1983). The metaphor of the brain as an engineer was intended to capture an important similarity between science and the task of the brain—namely, the exploitation of modelling, prediction, and so on—whilst nevertheless honouring an important difference between the two—namely, the fundamentally *pragmatic* function of neural activity.

I now think that this metaphor obscures rather than illuminates, however. No doubt the brain is in some important sense the *product* of something akin to engineering (i.e. through natural selection) (Dennett 1995), but—beyond the combined use of modelling and pragmatism—the metaphor of the brain *as* an engineer breaks down. Most engineers are primarily concerned

with building things—bridges, reservoirs, buildings, and so on—in a way that has no obvious parallel in the brain, and they are subject to very different practical constraints.

More importantly, however, Craik himself advanced a much more specific hypothesis about brain function that was not metaphorical. Although it is not at the centre of "The Nature of Explanation," it comes across explicitly in work that was published posthumously. It is a conception of the brain as a *regulator*.

(3.)**6.1. The Brain as an Element in a Control System**

In work written after "The Nature of Explanation," Craik advanced a conception of the brain as an "automatic regulating system" (Craik 1966, p.16) or "an element in a control system" (Craik 1948). This perspective evidently reflected his background in control theory and the role it played in the design of military technology such as radar-guided missiles and anti-aircraft gun predictors during the war. In such self-corrective "servomechanisms," information concerning deviations from goal states is fed back to correct behaviour. (The simple domestic thermostat is thus an example of a servomechanism in this sense). For Craik and a number of other theorists in the 1940s, many of whom were steeped in these military technologies, the strategy of exploiting negative feedback to generate sophisticated goal-directed behaviour in this way played a substantial role in the emergence of a general theoretical perspective on adaptive behaviour that Norbert Wiener would christen "cybernetics" in 1948 (Wiener 1948; see Boden 2006; Pickering 2010).

As a research programme, cybernetics sought to understand self-organization and purposeful behaviour in both artificial and biological systems in terms of circular causal chains in which information about the outcomes of the system's actions is detected ("fed back") and compared to goal states to cease, adjust, or prolong those actions (Boden 2006). In this way "all purposive behaviour may be considered to require negative feedback," such that "if a goal is to be attained, some signals from the goal are necessary at some time to direct the behaviour" (Rosenblueth et al. 1943, p.2). Error-correcting "teleological mechanisms" in this sense thus

> "show behaviour which is determined not just by the external disturbances acting on them and their internal store of energy, but by the relation between their disturbed state and some assigned state of equilibrium" (Craik 1966, p.12).

Like others before him, Craik recognised that this process of error-correction in the service of maintaining equilibrium is central to adaptive success in biological systems (see Bechtel 2008,

ch.6 for a historical review). That is, all biological systems must exchange matter and energy with the environment to maintain the structural integrity and constancy of their internal states in the face of a general tendency towards disorder (Anderson 2014; Bechtel 2008). This requires mechanisms that are responsive to deviations from this internal stability. "Living organisms," Craik (1966, pp.12-3) writes, are thus

> "in a state of dynamic equilibrium with their environments. If they do not maintain this equilibrium they die; if they do maintain it they show a degree of spontaneity, variability, and purposiveness of response unknown in the non-living world. This is what is meant by "adaptation to environment" … [Its] essential feature… is stability—that is, the ability to withstand disturbances."

In systems like us, this internal stability requires the regulation of essential variables such as "body temperature, blood Ph, blood oxygen level and circulation rate" (Craik 1966, p.33). Like other cyberneticists (e.g. Ashby 1952, 1956), Craik recognised that this core task of *self*-regulation demands that biological systems coordinate their behaviour with those features of the *external* environment that threaten to undermine it. Something as basic as temperature regulation, for example, "involves mechanisms at all levels from the sympathetic nervous system, through learned responses (such as going to open a window if the room feels too hot) up to the highest technical achievement—the invention of thermostatically controlled devices" (Craik 1966, p.35).

For Craik, this basic imperative to maintain what Walter Cannon (1929) had in 1929 named "homeostasis" provides a general theoretical perspective on the brain as an "automatic regulating or stability-restoring system" (Craik 1966, p.16). The third insight that I will extract from Craik's work, then, is this:

> (Insight #3) The fundamental task of the brain is regulation.

Crucially, this regulatory perspective on brain function provides a fundamentally pragmatic context in which to understand the emergence of mental models that will play an important role in later chapters. Nevertheless, so far I have said nothing to explicitly connect this regulatory account of brain function to the strategy of predictive modelling outlined in the previous two sections. Further, any such connection might seem strange given that cybernetics is typically situated in a tradition of research in psychology and artificial intelligence that is *hostile* to representationalist accounts of adaptive success (see Pickering 2010).

This is a crucial issue that I will return to at greater length in Chapter 7. For now, however, I will set it aside. My aim in this section has simply been to extract this dimension of Craik's work. As we will see in future chapters, it is another aspect of his approach to understanding the mind that seems impressively prescient, and the combination of this perspective with an emphasis on the adaptive value of predictive modelling is a proposal that will play an important role when we turn to contemporary research in the cognitive sciences that supports a conception of the mind as a predictive modelling engine.

Before concluding this section, I should note that this third insight has a much less plausible claim to substantial originality than the previous two (see Boden 2006). That is, a regulatory understanding of brain function has historical roots that go back much further than Craik and it was developed concurrently by a number of different theorists working in the 1940s (see Bechtel 2008, ch.6). Further, although I have characterised Craik as a cybernetician, it is important to stress that his work was "not directly influential in the American development of Cybernetics" (Gregory 1983, p.233). It was only in 1966 that Warren McCulloch arranged for Craik's previously unpublished work to be released in a monograph entitled *The Nature of Psychology*. As Boden (2006, p.217) puts it, shortly after Craik's death

> "cybernetic discussions—and cybernetic machines—were…springing up on both sides of the Atlantic… It was evident from Craik's (posthumous) papers, and from his book, that he'd anticipated the core themes. In brief: although he wasn't part of the cybernetics *movement,* he was an early, and highly significant, *cybernetician*."

The pioneering British cyberneticist Grey Walter made a similar remark in a letter to a friend shortly after meeting Norbert Wiener in 1947:

> "We had a visit yesterday from a Professor Wiener, from Boston… He represents quite a large group in the States… These people are thinking on very much the same lines as Kenneth Craik did, but with much less sparkle and humour" (quoted in Boden 2006, p.224).

(3.)**7. Conclusion**

As I hope to show in future chapters, Craik's "hypothesis on the nature of thought" was in many respects extremely prescient. Nevertheless, it was also speculative and highly schematic. In the next four chapters I therefore elaborate on the framework for understanding mental representation outlined in this chapter and situate it in the context of contemporary cognitive science. In Chapter 4 I expand on the chief characteristics of model-based representation. In Chapter 5 I show how these characteristics can be understood in the context of the neural

mechanisms underlying psychological capacities. In Chapter 6 I draw on research from cognitive neuroscience and machine learning that supports a conception of the mind as a predictive modelling engine. In Chapter 7 I then show how this research maps on to the three insights outlined in this chapter.

Chapter 4: Models as Idealised Structural Surrogates

> "The only PERFECT model of the world, perfect in every little detail, is, of course, the world itself" (Teller 2001, p.410).

(4.)1. Introduction

In the previous chapter I introduced three insights from the mid-twentieth century psychologist, philosopher, and control theorist Kenneth Craik: first, an account of mental representation in terms of neural models that capitalize on structural similarity to their targets; second, an appreciation of prediction as the core function of such models; and third, a regulatory understanding of brain function.

The aim of this chapter is to outline in more detail the chief characteristics of model-based representation. Specifically, I draw on work from the philosophy of science to argue that models should minimally be understood as *idealised structural surrogates* for *target domains*.

Turning to work in the philosophy of science might seem strange in the context of an investigation into the nature of mental representation. The kinds of representations trafficked in science are paradigmatic *public* representations. Perhaps we could learn something about such public representations by turning to the cognitive science and philosophy of mental representation—indeed, several philosophers of science have conducted extremely fruitful work in just this vein (Nersessian 1999; Thagard 2012)—but it is not immediately obvious what insights from the philosophy of science could contribute to our understanding of mental representation.

In Chapter 2, however, I suggested that mental representations can be understood as functional analogues of public representations, and in the previous chapter we saw Craik's suggestion that neural structures function as models "comparable" in important ways to scale models and those models that feature in science and engineering. Given the enormous body of work in the philosophy of science focusing on scientific representation, then, this suggests that one might be able to extract from such work a better understanding of the characteristics by which mental representations perform their systemic roles within our heads. That is the first reason for turning to the philosophy of science.

The second reason is related. Throughout most of the twentieth century, much of the philosophy of science was pre-occupied with a fundamentally *linguaformal* understanding of scientific representation. As Godfrey-Smith (2003, pp.186-7) puts it,

> "In [the] twentieth-century [most] philosophy treated theories as *linguistic* entities, as collections of sentences. So when people tried to work out what sorts of relationships theories have with reality, they drew on concepts from the philosophy of language. In particular, the concepts of *truth* and *reference* were emphasized. A good scientific theory is a true theory."

This understanding of scientific representation was heavily influenced by the logical empiricist tradition that dominated the philosophy of science in the middle decades of the twentieth century (see Giere 2006; Godfrey-Smith 2003). As the influence of this movement began to wane and naturalistic philosophers of science began to pay much more attention to the actual *practice* of science, many philosophers argued that this exclusively linguaformal understanding of scientific representation is inadequate to this practice. Among other flaws, it neglects the crucial role of representational structures such as scale models, mathematical models, and—especially in recent decades—computational simulations (see Frigg & Hartmann 2018; Weisberg 2013). Such forms of representation seem to play central roles in scientific theory construction, evaluation, and experimentation, as well as in the creative components of scientific discovery that had been taboo throughout most of the logical empiricist era. As Giere (2004, pp. 734-4) puts it, "a focus on the *activity* of representing fits more comfortably with a model-based understanding of scientific theories" (my emphasis). It is not obvious, however, that the structural units, semantic properties, and forms of reasoning associated with language-like systems of representation are adequate tools with which to understand the character of model-based representation (Giere 2004; 2006). To quote Godfrey-Smith (2003, p.187) again, mathematical models

> "are abstract mathematical structures that are supposed to represent key features of real systems in the world. But in thinking about how a mathematical model might succeed in representing the world, the linguistic concepts of truth, falsity, reference, and so forth do not seem to be useful. Models have a different *kind* of representational relationship with the world from that found in language."

The second reason for turning to the philosophy of science, then, is to get some insight into this different kind of representational relationship.

Importantly, the models I focus on here are *representational models* (sometimes called *models of phenomena*)—namely, models that represent the structure of target domains (Frigg & Hartmann 2018). These contrast with *models of data*, which function as corrected or regimented versions of "raw" data collected from observation, and with a sense of "model" prominent in formal logic and often associated with the "semantic view" of scientific theories,

according to which models are set-theoretical structures consisting of a set of objects, properties, relations, and functions, which provide an interpretation of a set of uninterpreted axioms (Suppes 1960; see Frigg & Hartmann 2018).

The models I focus on are "tools for representing the world" (Giere 2004, p.3). They include everything from a toy model of a car, the double helix model of DNA, the dynamic models used by forecasters to simulate likely weather outcomes, and rational choice models used in the social sciences. To begin with, however, it will be useful to start with the maximally simple example of *maps*, which, as Giere (2006, p.72) points out, "have many of the features of representational models."

(4.)**2. Maps**

Figure 1 provides a two-dimensional map of the area of Rochester in Kent where I grew up:

Figure 1. A map of part of Rochester, taken from https://www.google.co.uk/maps

This simple map situates key features of the area—roads, significant buildings and landmarks (the castle, parks, nightclubs, etc.), and so on—on a two-dimensional array that preserves the relative spatial locations of such features. If someone is new to the area and doesn't know their way around, this map could help them to navigate. For example, they could use the map to figure out how to get from Watts Avenue to Rochester Cathedral.

The example is trivial, but it serves to illustrate several points that will be important in what follows, each of which is carried over into the case of model-based forms of representation more generally.

(4.)**2.1. Target Domains**

First, to understand how maps like this work, one must distinguish between the map and its *target*, the domain that the map is intended to be a map *of* (Cummins 1996). One can only use the map in Figure 1, for example, if one knows that it is a map of this specific region of Rochester. In general, a map is a tool that we consult in order to coordinate various kinds of behaviour with a target domain—in order to navigate, plan routes, learn about the layout of a territory, and so on.

(4.)**2.2. Structural Similarity**

Second, how do maps like this work? That is, how do such maps enable us to navigate an area and plan routes within target domains? The answer that immediately jumps out is this: maps are in some sense *like* the terrains they represent. Specifically, the two-dimensional spatial structure of the map *resembles* the actual geographical structure of the relevant terrain: the former sustains a similar pattern of spatial relationships among its parts as the latter (Giere 2004). For this reason, one can use the map to *infer* numerous features of its target. For example, one can infer from the fact that Rochester Cathedral is closer to the train station than The Esplanade on this map of Rochester that this pattern of distance relationships is replicated among the corresponding places in Rochester itself.

Of course, the structural similarity between the map and the terrain is not sufficient to explain its function. Most fundamentally, the use of such a map depends on an enormous amount of background knowledge and capacities. Nevertheless, this should not detract from the importance of structural similarity for maps once these background conditions are in place: to anticipate a core theme that I will address in future chapters, someone who uses such a map to navigate her way from Watts Avenue to Rochester Cathedral is *dependent* on the structural resemblance between the map and this part of Rochester. If the similarity did not obtain—if this was a map of Bagdad, for example—her navigational efforts would fail.

Philosophers are often sceptical of appeals to similarity, especially in the context of discussions of representation (Goodman 1969; Sterelny 1990). A classic worry, for example, is that appeals to similarity are explanatorily vacuous because everything resembles everything else in some respect or other (Goodman 1969).

Return to the case of the map, however. Although it is true that this map could be defined as similar to an infinite number of systems along an infinite number of possible dimensions, only

one of these resemblance relations is *relevant* to the map's function: namely, the resemblance between the spatial layout of the map and the layout of its target domain (Giere 2004; Godfrey-Smith 2006b). When focusing on the similarity between a map and a target domain, then, the focus is on *relevant similarity*, not similarity as such. And—crucially—what is relevant is highly context-sensitive. As Teller (2001, p.401) puts it, the "problem [of similarity] dissolves as soon we notice that the demand for a general account of similarity can't be met because what is going to count as a relevant similarity depends on the details of the case at hand."

I think that a source of confusion here is a widespread tendency in philosophy to characterise representation as a *dyadic* relation between the representation and what it represents (see O'Brien & Opie 2004). This characterisation is central to Goodman's (1969) classic dismissal of the relevance of resemblance to representation, for example. If we think of representation in this way, the challenge becomes one of specifying a dyadic relation in virtue of which a representation represents some given phenomenon (see Crane 2003). In this context, appeals to similarity are obviously hopeless for familiar reasons: mere similarity cannot explain the fact that a representation represents something, because similarity is ubiquitous. Worse, it also has several logical properties that render it a non-starter in this context: similarity is reflexive and symmetrical, for example, but representation is not (Goodman 1969).

One response is to give up on resemblance (Crane 2003). Another—the one I will pursue here—gives up on the idea of representation as a *dyadic* relation. Instead, one should think of representation as a *triadic* relation that implicates not just the representation and what it represents but also the *representation-user* that uses the former to represent the latter (Millikan 1984; O'Brien & Opie 2004; Peirce 1931-58).

Importantly, once we think of representation in this way, the challenge is not to specify a dyadic relation that makes it the case that a representation represents some specific thing, but rather to identify the relation between the representation and a target that explains *how* an agent can successfully use the former as a representation of the latter (Giere 2006; O'Brien & Opie 2004). It is in this context that the appeal to relevant similarity plays a role: what explains the fact we can use a map as a representation of a target terrain is that it recapitulates—resembles—the geographical structure of that terrain.

To illustrate the context-sensitivity of appeals to similarity in this context, consider a map of tube connections in London:

Figure 2. A map of tube connections in London (taken from https://tfl.gov.uk/maps/track/tube)

In this map, the relevant similarity between the map and the target domain does not relate their *geometric* structure but rather their *topological* structure. Specifically, it is the similarity between the pattern of topological relationships between elements of the map and tube connections in London that explains—once a number of background conditions are in place—one's capacity to use the former to navigate the latter.

(4.)**2.3. Idealization**

A third important feature of maps is that they are often heavily *idealized* (Giere 2004; Weisberg 2013). There are at least two components to this that should be kept apart and that will be important in what follows.

First, maps are highly *selective*. They abstract away from most of the features of a target domain to focus only on those features that are relevant for the map's function.

Second, maps are rarely—and perhaps never—perfectly accurate. As Giere (2004, p.3) notes, "in the limit, the only perfect map of a territory would be the territory itself, which would no longer be a map at all." It is often said that maps exploit an isomorphism or homomorphism between the map and some target domain (Camp 2007). The problem with this view, however, is that "no reasonably detailed map can be accurate enough to exhibit a literal isomorphism with identifiable features of a real geographical structure" (Giere 2004, p.4). Even a highly

accurate map of the sort exhibited in Figure 1 is forced to extract two-dimensional spatial structure from a complex three-dimensional terrain distributed on a spherical planet. Many other maps will exhibit even greater distortions. As such, "talk of similarity [rather than isomorphism] makes explicit the fact that one does not expect there to be a perfect fit between models and the world" (Giere 2006, p.66).

Importantly, this means that accuracy in maps is always a matter of *degree* (Godfrey-Smith 2003; O'Brien & Opie 2004). Maps can be *better* or *worse*. Consider Figure 3, for example. It displays a hastily drawn route from Watts Avenue to Rochester Cathedral of the sort you might draw on a napkin for a visiting friend. This napkin map is highly inaccurate, in the sense that at best it preserves only a set of extremely coarse-grained relationships among parts of the represented terrain. Nevertheless, it is probably not *so* inaccurate that it couldn't be of use to a friend who doesn't know the area, and who wants to make her way to the Cathedral. That is, it is accurate *enough* for that goal. One can imagine a continuum of maps from this maximally simple one to the highly accurate one used by Google shown in Figure 1. At no point in that continuum, I would suggest, do maps suddenly go from being "false" or "wrong" to "true" or "veridical." With maps, there are degrees of accuracy, and how much accuracy a map-user needs is dependent on her *interests*—a point I return to below.

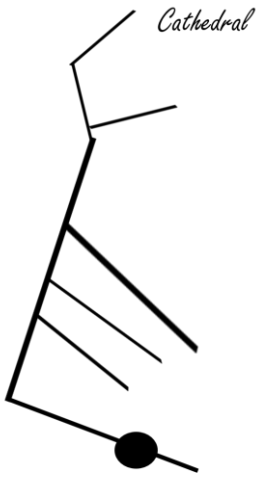

Figure 3. A hastily drawn map tracing an effective route to get from somewhere in Watts Avenue to the Cathedral. The pattern of lines only partially replicates the actual spatial information of this region of Rochester as seen in Figure 1.

Both these dimensions of idealization reflect two species of idealization familiar from the philosophy of science.

The focus on selectivity captures what is sometimes called *Aristotelian idealization*, which involves "stripping away… all properties from a concrete object that we believe are not relevant to the problem at hand," allowing us "to focus on a limited set of properties in isolation" (Frigg & Hartmann 2018, p.5). The focus on the lack of perfect correspondence captures *Galilean idealization*, which involves various species of deliberate distortions—for example, models in physics of point masses moving on frictionless planes, or the assumption often used in economic models that agents are omniscient (Frigg & Hartmann 2018, p.5). As with maps, both forms of idealization often feature together in scientific models.

(4.)**2.4. Two Species of Accuracy**

Accuracy in maps is always a matter of degree. It is important, however, to distinguish at least two different dimensions of representational virtue when using a map, and accuracy—as I am using the term, at least—is only one of them.

"Accuracy" refers to the dimension of representational virtue that we have already seen: to what extent does the map faithfully re-capitulate the relevant structure of some domain? As noted, this dimension of representational evaluation compares the degree of similarity between the entire map and the relevant features of a target domain. Relative to this standard of evaluation, misrepresentation amounts to a map that fails to adequately capture the relevant structure of a target.

A second dimension of representational evaluation occurs in the context of *using* a map that is already sufficiently accurate. For example, one needs to *locate* oneself on the map as one navigates a territory. In this case, we can evaluate to what extent one's guess corresponds to one's actual position in the territory. Here we might be more tempted to think the issue is not one of degree—that one is either located at the place one guesses, or one is not. It is important to bear in mind, however, that such judgements are made relative to a given map, whose correspondence to a target domain *is* one of degree. Strictly, this means that the second form of representational evaluation is only possible once one has a sufficiently accurate map in the first place. Further, often one's location of oneself on a map does not need to be very precise, and one might think that the degree of precision required is relative to one's practical goals in using the map. This brings me to the fifth characteristic of maps.

(4.)**2.5. Interest Relativity**

Giere (2004, p.4) writes that "maps are *interest relative*, and necessarily so." This interest relativity falls out directly from what has been written so far, but it is worth making it explicit. As noted above, maps vary enormously in both the features of a territory that they represent and the degree of accuracy with which they represent them. It is the role of *interests*—the contingent interests of the relevant map-makers and map-users—that determines these two characteristics. Crucially, this applies even when one's interest is to make a map that is maximally accurate. In such cases, the reason one wants a map and will not settle for the territory itself is that the map is more easily accessed and exploited, which will entail a complex trade-off: achieving as much accuracy as possible subject to the constraint that it can still be used as a map. Nevertheless, often one's interest in using a map is purely pragmatic. Under such conditions, a map can be *too* accurate, in which case it is a bad map, and a highly inaccurate map can be excellent if it is sufficiently accurate to serve one's ends in a convenient way.

(4.)**2.6. Summary**

To summarise, then, maps function as idealised structural surrogates for target domains. Importantly, I believe that each of the characteristics I have identified as features of map-based representation extend to model-based forms of representation in general (see Giere 2004). Because maps are a maximally simple example of a model, then, they thus provide a useful illustration of these broader characteristics. Nevertheless, maps of the sort just described are quite limited if we want to gain a better understanding of mental representation. Although humans and other animals do need to represent the spatial structure of their environments, typical maps do not help much with the phenomenon of *prediction* that we saw was central to Craik's hypothesis on the nature of thought outlined in the previous chapter.

It will be helpful, then, to turn to a kind of model that incorporates not just spatial structure but *dynamic* structure.

(4.)**3. Orreries**

Most maps are *static* representations—namely, representations of the enduring spatial structure of physical objects or a broader terrain. Since human beings have been designing models, however, they have been concerned with models of *dynamic* phenomena: models of parts of reality that change over time in regular or law-governed ways that can therefore be used in the service of *prediction*. One of the most influential examples of such models in the philosophical literature is an *orrery* (a physical model of the solar system) (Ryder 2009b).

An orrery contains physical structures that stand in for the significant elements of the solar system (e.g. the Sun, planets, and sometimes their moons) in a way that preserves their spatial relationships to one another and often their relative sizes and proportionality. In addition, however, it also contains a system of gears that ensures that the pattern of regularities as the relevant structures orbit the Sun *mirrors* (when accurate) the actual motions of the real planets and moons in space. (In this way orreries provide a simple example of the analogue machines that we saw Craik appeal to in the previous chapter.) This dynamic characteristic confers a fundamentally *predictive* capacity. Specifically, the structure of the orrery ensures that the relative positions of some planets determines the positions of others in such a manner that one can use the former to predict or infer the latter. Further, one can use the orrery to predict the evolution of planetary positions for a large set of possible initial conditions.

Importantly, an orrery illustrates each of the characteristics associated with maps above.

First, one must distinguish between the model and its target—in this case, the solar system. Second, the orrery capitalizes on structural similarity: the spatial structure and motions of the model planets recapitulates the relative spatial structure and regularities of the solar system. Third, this similarity is heavily *idealized*: an orrery neglects most features of the solar system, and no orrery can be wholly accurate. (The dynamics sustained by the system of gears can only approximate the actual motions of the planets). Fourth, orreries admit of at least two species of representational evaluation: first, with respect to the similarity between the orrery and the solar system; second, with respect to one's attempts to use the orrery to model the *current* state of the solar system. Finally, orreries are interest-relative: which features of the solar system and how much accuracy are required in the model depends on the contingent function (ornamental, educational, predictive) of the orrery.

(4.)**4. Models in Science**

Maps and orreries are both *physical models*. Although such physical or "concrete" models are used in science and engineering, it is now more common to use *mathematical models* or *computational models* (which on some accounts are just a species of mathematical models) (Weisberg 2013). As Godfrey-Smith (2003, pp.187-8) puts it,

> "Mathematical models… are abstract mathematical structures that are supposed to represent key features of real systems in the world… Abstract mathematical models might be thought of as attempts to use a general-purpose and precise framework to represent *dependence relationships* that might exist between the parts of real systems. A mathematical model will

> treat one variable as a function of others, which in turn are functions of others, and so on. In this way, a complicated network of dependence structures can be represented."

In the philosophy of science, much of the focus has been on the use of dynamic models specified by differential equations. Dynamical models of this kind use calculus and differential (or sometimes difference) equations to describe the relationships between a set of quantities that change over time. The dynamics of a swinging pendulum, for example, can be captured in terms of the changing values along three different dimensions: its angle, its angular velocity, and time. As such, every possible state of the pendulum over some time period can be represented in terms of these three numbers (and thus plotted in three-dimensional space). A dynamic model will have equations describing the evolution of the pendulum as a function of the relationships between these variables. If one combines such a model with a specification of the pendulum's initial conditions, the model can be used to *predict* its temporal evolution. As such, one can use the model to predict how its behaviour *would* evolve under a large set of possible initial conditions.

As with the previous models I have considered, there is an obvious sense in which a mathematical model of this kind is also dependent on structural similarity. As Godfrey-Smith puts it above, such mathematical models are supposed to accurately represent a complex structure of dependence relationships. *Unlike* the previous examples I have considered, however, the mathematical symbols used to express this model evidently do not themselves instantiate this structure. Rather, it is natural to say that they *specify* an *abstract* structure whose resemblance to some target domain can then be evaluated (Giere 2004; Godfrey-Smith 2006b; Weisberg 2013).

Nevertheless, for most mathematical models of this kind one can then construct a system that instantiates the relevant structure it specifies. If—as will almost certainly be the case—the model is highly idealised, this instantiation of the model will differ in many ways from the real-world system it is intended to represent. If the differences are minimal or irrelevant, such an instantiation can still be highly useful. For example, there are computer programs with which one can convert a dynamic model into a simulation of the target system. If one feeds in the initial conditions of the system, the simulated world will then automatically unfold as the relevant equations dictate.

A familiar example of this is the dynamic models used by weather forecasters (see Ryder forthcoming). Such models contain elements (variables) that systematically map onto positions

in the atmosphere, where the elements can take on different values depending on different weather conditions—on whether there is rain, snow, a hurricane, or clear sky at the relevant position, for example. Such models are useful—they can be used for prediction and thus forecasting—to the extent that the variables in the model covary in a way that mirrors the actual structure of the environment.

Dynamic models such as this serve numerous different functions for model-users. I will mention just a few.

First, they facilitate *prediction* in a conventional sense of that term: if we have a dynamic model, we can use it to predict the evolution of a system as a function of its initial conditions. For example, we can exploit a dynamic model of a pendulum to specify its trajectory given a specification of its angle and angular velocity and starting conditions, and we can exploit a dynamic model of the sort used by weather forecasters to generate predictive simulations of likely weather outcomes under a range of possible conditions (see Ryder forthcoming).

Second, models facilitate what might also be thought of as *synchronic prediction*. Specifically, one can often use a dynamic model of the covariational relations among the features of a domain to read off the values of some variables by specifying the values of other variables. That is, because the model captures the structure of a domain, one can use the values of some variables to "*fill in"* the values of other variables. To illustrate this, Ryder (forthcoming) uses the helpful example of a model of a sink with variables that map onto the sink's functionally important parameters: the radial positions of knobs that release water, the temperature at the tap, which knob releases hot water and which releases cold water, the flow rate at the tap, and so on. With a model of the covariational structure of sinks (i.e. a set of equations that describes the dependencies among these variables), Ryder points out, one can use information about the values of some variables (for example, the radial position of the knob responsible for releasing hot water) to infer the values of others (for example, the temperature of the water).

Third, dynamic models also generally facilitate action planning and intervention. That is, just as one can use a model to predict how the domain would change or evolve under certain interventions, one can use such models to learn how the domain *would have to change* to yield certain desirable outcomes. For example, one can change the value of the temperature variable for the water in one's sink to a desirable temperature and the model will tell you how you would have to change the values of other variables—the radial position of the knobs—to generate that temperature (Ryder forthcoming).

More generally, these functions illustrate that models can be exploited for what Swoyer (1991, p.449) calls "surrogative reasoning": because the model—when accurate—mirrors the structure of its target domain, one can use the model as a *surrogate* for that target domain, reasoning "directly about a representation in order to draw conclusions about some phenomenon that it represents." In this way the model-user can "cognitively process the structure of the target by manipulating the structure of its representation" (Cummins 2010, pp.101-2).

Importantly, this characteristic of models is central to the purely exploratory functions that play such a major role in the use of models in science. Indeed, several theorists have argued that it is *distinctive* of model-based science to exploit a fundamentally *indirect* form of representation of the world (Godfrey-Smith 2006; Weisberg 2013). As Godfrey-Smith (2006, p.726) puts it, "the modeler's strategy is to gain understanding of a complex real-world system *via* an understanding of a simpler, hypothetical system that resembles it in relevant respects." In this sense "the obvious strategy that contrasts with model-based science is to avoid any deliberate detour through merely hypothetical systems, and seek to directly represent the workings of the real-world system we are trying to understand" (Godfrey-Smith 2006, p.730).

Nevertheless, it is important to stress that this "strategy" of model-based science is just that: a strategy. That is, because models function as idealised structural surrogates for target domains, they *facilitate* this indirect mode of representation and understanding. This indirect strategy is not *constitutive* of models, however. As we will see in future chapters, this structural surrogacy can also be exploited for more interactive—less indirect—forms of coordination with target domains.

(4.)**5. Conclusion**

To summarise, then, models can be understood as *idealised structural surrogates* for *target domains*. Of course, so far this analysis has applied exclusively to *public* models—namely, material models "out there in the world" that multiple agents can scrutinise and exploit, or abstract structures specified by intelligent agents using language or mathematics. In the previous chapter I outlined Craik's view that we get a better understanding of our own minds by understanding such public models. With this more thorough analysis of public models in hand, we are now in a better position to evaluate Craik's proposal that their core functional profile can be transposed into the head.

Chapter 5. Could the Mind be a Modelling Engine?

> "What makes sophisticated cognition possible is the fact that the mind can operate on something that has the same structure as the domain it is said to cognize" (Cummins 1994, pp.297-298).

(5.)1. Introduction

In Chapter 3 I extracted a proto-theory of mental representation from Kenneth Craik in which mental representation is understood in terms of idealised predictive models that share an abstract structure with target domains in the world. In the previous chapter I outlined in more detail the chief characteristics of model-based representation. I argued that models can be understood as holistic, interest-relative, idealised structural surrogates for target domains. Following Craik, the hope was that by focusing on the characteristics of public models we might gain a better understanding of the nature of mental representation.

It is here that we confront a problem, however: the construction and use of public models in science, engineering, and elsewhere exists against a backdrop of interpretative social practices brimming with intelligence and often ingenuity. It is obviously no good importing such characteristics directly into the head. Cognitive science posits mental representations to *explain* intelligence.

In Chapter 3 I outlined the beginnings of a response to this objection that I attributed to Craik. According to this response, even though public models do exist against the backdrop of intelligence and interpretation, one can nevertheless isolate their core functional profile—what Godfrey-Smith (2006) calls the "empirical skeleton" of public representation-use—in abstraction from such features, and there is no a priori reason why this core functional profile could not be carried over into neural mechanisms.

The previous chapter was in part an attempt to elaborate on this core functional profile of models. This chapter is devoted to showing that—and how—it can be transferred into the activity of neural mechanisms that underlie an organism's psychological capacities.

At the core of a model-based account of mental representation is the concept of structural similarity. Section 2 of this chapter begins with an overview of two classic challenges to similarity-based accounts of mental representation: that such accounts are inconsistent with materialism, and that similarity has properties—ubiquity, reflexivity, and symmetry, for example—that representation does not. I argue that the first challenge can be answered by focusing on the abstract character of structural similarity and that the second challenge can be

answered by adopting a *triadic* account of mental representation outlined in the previous chapter.

Sections 3, 4, and 5 then explain how three features of model-based representation can be carried over into the case of mental representation. Section 3 identifies the worldly structure that is relevant to mental models. Section 4 explains how one can make sense of the concept of targets for mental models. Section 5 explains in what way the psychological capacities of an organism that relies on mental models are causally dependent on the structural similarity between such models and their representational targets. I conclude in Section 6 by summarising how this chapter provides an account of mental representation that satisfies one of the constraints on such an account outlined in Chapter 2: namely, that a theory of mental representation must explain—at least in broad outline—how the postulation of mental representations can contribute to the causal explanation of psychological capacities.

(5.)**2. The Challenge to Similarity**

If the previous two chapters are along the right lines, at the centre of model-based representation is the concept of structural similarity. If we are to make sense of mental models, then, we must be able to make sense of similarity between structures in the brain and worldly domains.

For many, this already undermines the prospects of a model-based account of mental representation. In the twentieth century a consensus arose in philosophy according to which similarity-based accounts of mental representation are hopeless (Crane 2003). Fodor, for example, writes:

> "There are, so far as I know, only two sorts of naturalistic theories of the representation relation that have ever been proposed. And at least one of these is certainly wrong. The two theories are as follows: that C [i.e. the description of the representation relation] specifies some sort of resemblance relation between R [representation] and S [what it represents]: and that C specifies some sort of causal relation between R and S… The one of this pair that is certainly wrong is the resemblance theory" (Fodor 1984, p.233).

Likewise, Sterelny (1990, p. 65) writes:

> "Before this century, particularly in the empiricist tradition, resemblance seemed an obvious key to the explanation of mind-world relationships… This idea is all but universally abandoned."

There are many grievances that philosophers have raised for similarity-based views of representation, but there are two that will be of special importance here.

The first is that similarity-based accounts of mental representation are inconsistent with even a very vapid materialism. This argument is straightforward: it becomes impossible to believe that the mind literally inherits the properties of the things we represent once one situates the mind in the physical world. Neural states that represent bananas are not yellow and the neural states that represent tables do not have four legs.

In principle, at least, this objection is easily undermined. Structural similarity does not require that a system of representations inherits the first-order properties of representational targets: their shape, size, colour, and so on. Rather, it just requires that the system of representations sustains a similar *pattern of relationships* among its parts as the target domain (Gallistel 2001; O'Brien & Opie 2004; Palmer 1978). Crucially, this does not require that the *kinds* of relations in the two structures are the same. Although in maps we typically re-present the spatial layout of a domain with patterns of spatial relationships among elements on the map, this is incidental. In a thermostat, for example, the pattern of spatial relationships among bimetallic strip curvatures recapitulates the pattern of relationships among ambient air temperatures (O'Brien 2015).

The key idea here is that of a structure-preserving mapping between the elements of two domains—what Palmer (1978) calls the *representing world* and the *represented world*—such that the important relations between elements of the latter are mirrored by relations between elements of the former. O'Brien and Opie (2004, p.8) offer a formalisation of this notion in terms of what they call "second-order resemblance":

> "Suppose SV = (V, ℜV) is a system comprising a set V of objects, and a set ℜV of relations defined on the members of V. The objects in V may be conceptual or concrete; the relations in ℜV may be spatial, causal, structural, inferential, and so on… We will say that there is a second-order resemblance between two systems SV = (V, ℜV) and SO = (O, ℜO) if, for at least some objects in V and some relations in ℜV , there is a one-to-one mapping from V to O and a one-to-one mapping from ℜV to ℜO such that when a relation in ℜV holds of objects in V, the corresponding relation in ℜO holds of the corresponding objects in O. In other words, the two systems resemble each other with regard to their abstract relational organisation."

In the case of mental representation, then, what is required is that the brain's representational vehicles can sustain similar patterns of relations as the relevant features of target domains,

subject to the constraint that such representations will be idealised. Given the structural complexity of the brain, there is no a priori reason for thinking that this is impossible (Churchland 2012; O'Brien & Opie 2004). Nevertheless, this mere in-principle possibility does not take us very far. We still need to know what the relevant worldly structure is that the brain's mental models are supposed to capture. I address this question in the next section.

The second class of challenges to similarity-based accounts of mental representation contend that similarity has the wrong properties to explain representation. Three important such properties are as follows (see Goodman 1969). First, similarity is ubiquitous: everything resembles everything else in some respect or other. (Note that this problem becomes even worse in virtue of focusing on a graded notion of similarity rather than strict notions like isomorphism or homomorphism (Shea 2014)). Second, similarity is reflexive: all things resemble themselves, but representations (typically) do not represent themselves. Finally, representation is symmetric: if A resembles B in respect R, B resembles A in respect R as well. Representation does not (typically) have this characteristic, however. Mona Lisa did not represent Da Vinci's painting.

Again, in principle I think that these challenges are easily answered. Recall the triadic understanding of representation outlined in the previous chapter. In this triadic context, the challenge is not to analyse a dyadic representational relation in virtue of which one thing represents another thing but rather to explain *how* an agent successfully uses one thing—the representation—as a surrogate for another thing—its target. It is here that appeals to structural similarity become relevant. The proposal is that specifying the structural similarity between the representation and its target offers at least part of an *explanation* of an agent's capacity to successfully use the former as a surrogate for the latter.

Again, however, this does not take us very far. What it shows is that any model-based account of mental representation must be able to make sense of two phenomena: first, the concept of a *target domain* for mental models; second, the idea that psychological mechanisms *exploit* the structural similarity between mental models and such target domains in the service of effective functioning. I turn to the first of these issues in Section 4 and the second in Section 5.

(5.)**3. Making Sense of Structural Similarity**

I will assume that scepticism about structural similarity-based accounts of mental representation is not directed at the idea that the world (including the brain) *has* structure. Specifically, I will assume a vapid form of commonsense and scientific realism throughout:

there really are electrons, neurons, synapses, neurotransmitters, organisms, tables, words, causal relations, spatial relations, and so on (although see Chapter 9 for some important qualifications). As this list illustrates, however, the world has lots of different *kinds* of structure. Which structure is relevant to the case of mental representation?

In the case of the brain, this question relates directly to the issue of representational *formats* outlined briefly in Chapter 2. Roughly, it is just the question: what are the relevant properties of the brain's representational vehicles and the important relations between them? The qualifier "relevant" is again crucial here. Whatever the brain's representational vehicles are, they will have many properties (e.g. colour) that are both irrelevant to their systemic role as representations and irrelevant to the relation they stand in to their representational targets.

I will not take a stand in this chapter on the nature of the brain's representational vehicles (see Chapter 6 and 9 for some important possibilities). Nevertheless, it is important to note the structural *complexity* of the brain's potential representational capacities. The brain contains tens of billions of neurons, most of which form thousands of synaptic connections. The structural complexity of potential synaptic connections and patterns of activity distributed across the brain's neural works is thus immense (Churchland 2012). Once we have a liberal notion of structural similarity that is substrate independent (that does not require similar *kinds* of relations among representations and elements of the target), then, the idea that the brain is able—at least in principle—to inherit "the same structure as the domain it is said to cognize" (Cummins 1994, pp.297-298) should be uncontroversial.

(5.)**3.1. Prediction and Regularity**

For present purposes, a more important issue than the structure of the brain's representational vehicles is the *worldly* structure that is relevant to the case of mental representation. As noted above, the world has numerous kinds of structure. Which structure is relevant for capacities like perception, sensorimotor control, imagination, reasoning, and so on?

To answer this question, recall the second component of Craik's proto-theory of mental representation outlined in Chapter 3: namely, that the core function of the brain's mental models is *prediction*. Given that prediction is only possible in the presence of regularity, Craik's perspective leads us to the view that the target of mental models is the world's *regularity structure*. Most abstractly, we can understand this concept of regularity structure in terms of the complex networks of causal and correlational relationships among environmental

variables—or, to anticipate a theme that will emerge in later chapters, those networks *relevant* to the organism's practical ends.

I will assume a simple understanding of causal relationships here according to which—roughly—one variable X is causally relevant to another variable Y if and only if systematic interventions on the values of X would result in associated changes in Y. More precisely, this so-called "interventionist" account can be summarised as follows:

> (Causal Relevance) X causes Y if and only if there are background circumstances B such that if some (single) intervention that changes the value of X (and no other variable) were to occur in B, then Y would change (Gładziejewski & Miłkowski 2017, p.342; see Woodward 2003, 2008).

This only captures the skeleton of Woodward's detailed and subtle account of causal relevance, but it will be sufficient for my purposes here. I will assume that causal relationships comprise a genuine feature of the world that organisms like us must engage with, although I will not assume any metaphysical story about what makes causal statements true. Importantly, the concept of intervention here does not require human intervention of the sort brought about in controlled experiments; any change in X would count. Further, causal relevance evidently does not require *deterministic* causal relationships: smoking causes cancer, but not everyone that smokes gets cancer.

A useful way to think about correlational relationships here is in terms of the concept of *mutual information*, a measure of statistical dependency that is generalised to cover networks of variables (Friston 2002, pp.230-1). Roughly, two variables X and Y are mutually informative if observation of the values of one reliably reduces uncertainty about—that is, predicts—the values of the other. Again, this notion of correlation is probabilistic: observing the newspaper someone reads reliably reduces uncertainty about their political views, but the relationship is of course not deterministic.

The distinction between correlation and causation is important when it comes to prediction. Although one can exploit mutual information for prediction, there is one kind of prediction—prediction of the likely effects of interventions on the world—where mutual information alone is insufficient. To take a simple example, the state of the light bulb is correlated with the state of the light switch, but one cannot exploit this information alone to predict the likely consequences of one's intervention upon either entity (e.g. breaking the light bulb will not

change the state of the light switch). For this, one needs to know the asymmetric causal relationship that connects them.

Crucially, the simplicity of these examples reflects expository convenience and should not be taken to imply that the worldly structure relevant to the case of mental representation consists primarily of simple dyadic causal or correlational relationships between variables. The kind of pairwise correlations between things like lightning and thunder that preoccupied the British empiricists are the exception in nature. Most causal and correlational relationships exist as parts of highly complex multivariate networks.

(5.)**3.2. Sources of Mutual Information**

Is there anything more specific that can be said about the structure of such networks? One of the most interesting suggestions I know comes from Dan Ryder (2004; forthcoming), drawing on work from Millikan (1999) and others (Boyd 1991; Kornblith 1993). Ryder argues that patterns of regularity organised around *sources of mutual information* (SOMI) provide the fundamental target of the neocortex's representational capacities.

The simplest way to understand this idea is by focusing first on classic homeostatic property cluster accounts of natural kinds (Boyd 1991). On such accounts, natural kinds are characterised by sets of systematically correlated (i.e. mutually informative) properties where such correlations are unified and explained by some underlying reason. For example, water is characterised by mutually informative properties such as liquidity under certain conditions, boiling at 100 degrees, translucence, and so on, where the correlations between such properties are explained by H20, the chemical structure of water. As such, one can exploit observations of subsets of such properties to infer the presence of the kind *water*, which can in turn be used to predict the presence of other correlated properties. Likewise for non-chemical kinds: biological kinds license similarly reliable inductive generalisations, where correlations among properties are instead explained by common evolutionary histories rather than underlying chemical structure.

As Millikan (1999) points out, this basic structure extends beyond natural kinds. For example, non-natural but real kinds such as tables or mobile phones have multiple correlated properties explained by factors such as their function or aetiology, and the systematic correlations exhibited by individuals are explained by their persistence through the time. Further, as Ryder (2004, p.214) puts it,

> "Events and processes, whether particulars or kinds, are also sources of correlation, for instance Halloween, World War 2, biological growth, and atomic fusion. Dynamical systems, like homeostatic systems, economic systems, eco systems, organisms, and many artifacts, have an underlying causal structure that is a source of correlated variation among the system's global effects."

Perhaps most importantly, this general way of understanding regularity patterns also seems to map onto contemporary ways of understanding both cortical information processing and work on machine learning in artificial neural networks (Ryder 2004; forthcoming). For example, deep learning architectures trained to learn the relevant structure of some input data—for example, natural images—can be understood as recursively extracting the sources of mutual information (i.e. correlations) contained in the data: systematic correlations in the raw input data (e.g. pixel values) are explained by interactions among variables registered at the level above (e.g. oriented edges), correlations among which are in turn explained by interactions among variables registered at the level above them (e.g. spatial motifs), and so on (Lecun et al. 2015). Likewise, the hierarchical structure of information encoding in cortical networks can plausibly be understood in the same way: each level of the representational hierarchy is concerned with identifying the sources of spatiotemporal correlations among features registered at the level below (Ryder forthcoming).

I will return in more depth to this idea of recursively extracting latent features of the environment responsible for generating observed regularities in future chapters. For now, this important idea of patterns of regularity organised around SOMIs should illustrate how general this kind of regularity pattern is in the world.

It is an interesting question just *how* general it is—that is, how much of the world's regularity structure relevant to cognitive processes can be captured by appeal to this concept. Ryder (2004; forthcoming) argues that *all* information processing in the neocortex is organised around sensitivity to SOMIs. For my purposes here, I do not need to endorse this strong claim. What is important is just the idea that the environment (including the organism) exhibits complex, nested regularities, which—in principle, at least—neural networks can model and exploit for prediction. In the next chapter we will explore a concrete suggestion for how this actually occurs in the brain.

(5.)**4. Target Domain**

In the previous chapter I stressed that one must be able to distinguish between models and their targets. A specification of structural similarity between (accurate) models and target domains is intended to explain an agent's capacity to successfully use the former as a surrogate for the latter (see below Section 5). Importantly, however, a model's target domain cannot itself be *determined* by structural similarity: any given model is structurally similar to an infinite number of possible things along an infinite number of possible dimensions (Cummins 1996). In the case of public models, the specification of their target is easy: *we* decide. This story obviously cannot be carried over to the case of *mental* models, however. What determines their targets, then?

In fact, there are really two questions here.

First, I argued above that the generic target of mental models is the environment's regularity structure. What makes this the case? The second question concerns the targets of *specific* mental models or regions of a broader model. For example, suppose—to simplify greatly—that the models in our visual system target that aspect of the environment's causal and statistical structure implicated in generating the spatiotemporal patterns of stimulation at our retinal cells. What makes this the case?

It is tempting to become exasperated with questions like this. Neuroscientists and psychologists do not spend their time worrying about such broadly "metaphysical" questions, so why should we? I am sympathetic to this temptation. Nevertheless, I think that the exasperation here—and the neglect of such questions by working scientists—reflects the fact that the problem can be straightforwardly solved, not that it is unreal.

First, then, consider that if an animal commands a mental modelling engine of the sort hypothesised by Craik, this is not an accident. That is, if there is a general-purpose modelling engine inside our heads, it presumably has the *function* of capturing the regularity structure of the ambient environment—and not, say, fortuitously arranged grains of sand on a beach in Normandy. We can therefore appeal to this function in determining its target. Of course, there is a question of how exactly to understand the concept of function here—for example, whether it should be understood in terms of the evolved function of the relevant mechanism or some ahistorical account of function. I will return to this question in Chapter 7.

Now turn to the target of specific mental models or specific regions of broader mental models. Insofar as the generic target of the mind's modelling engine is the world's regularity structure, we can specify the target of specific models by identifying that aspect of the world's causal-

statistical structure causally implicated in their production. For example, to simplify greatly once more, the target of mental models realised in visual systems is the "visual world" (the shapes, sizes, locations, textures, etc., of visible objects) because that is the aspect of the world causally implicated in their production.[10] As we will see in the next chapter, this maps on nicely to theories in cognitive neuroscience in which neurally realised *generative models* target the process responsible for generating the data to which they are exposed.

Ryder (2004; forthcoming) advances a similar view. He argues that just as scale models are often constructed from a template—for example, a toy car is constructed by consulting the actual car that it is a model of—we can also talk of the templates from which *mental* models are constructed: namely, that aspect of the world that they are designed to mirror. In this way "evolution designs a model making machine, and the operation of this machine constitutes learning," such that the brain (specifically, the neocortex) has evolved to construct models of an open-ended variety of templates: namely, those features of the world that generate the information to which the brain is exposed (Ryder forthcoming, p.67).

This is a story that fits nicely with the work on cortical information processing that I will outline in the next two chapters.

(5.)**5. Exploiting Mental Models**

I have argued that we should understand the structural similarity between models and their targets in the context of a triadic understanding of representation that includes:

(1) The model;

(2) The target of the model;

(3) The representation-user that exploits the structural similarity between (1) and (2) in the service of goals.

As always, it is helpful to focus on a maximally simple example. Consider a person, Sally, exploiting the structural similarity between the spatial layout of a map of London and the layout of London itself in navigating her way from Big Ben to Nelson's Column. In cases like this, the map functions as a tool for achieving a goal. Specifically, the representation-user's success

[10] Other authors appeal not to the causal history of mental models—or at least not only to their causal history—but to their *use*, such that a mental model targets domain D because the representation-user exploits the mental model in coordinating its behaviour with D (e.g. see Ramsey 2007). One obvious problem with this suggestion, however, is that cognitive systems can engage with mental models in a way that is purely "offline" (see Chapter 7).

in achieving this goal is *dependent* on the accuracy of the map. From an outsider's perspective, we can therefore appeal to the accuracy of the map as part of a *causal explanation* of the representation-user's success: if Sally's map had been inaccurate, she would not have made it to Nelson's Column. In this sense structural similarity is what Godfrey-Smith calls an "exploitable relation" (2006a) and "fuel for success" (1996; see also Shea 2014).

As Gładziejewski and Miłkowski (2017) point out, the concept of causal explanation here can be understood in terms of the interventionist account of causal relevance outlined above. Let the variable X stand for the degree of structural similarity between the map and the relevant part of London and let Y stand for the degree of success that Sally has in achieving her goal. Sally's success is *causally dependent* on the structural similarity between her map and London in the following sense: systematic interventions on X would result in associated changes Y, such that *if* the similarity did not obtain, Sally would not be successful. Of course, X and Y must here be understood as continuous variables, and it may be—as pointed out in Chapter 4—that *too much* similarity (high values for X) would in fact hinder Sally's goals. This does not undermine the basic story, however.

In the case of public models such as maps, this schema is relatively straightforward to understand. There is no deep mystery in attaching success conditions to Sally's navigational endeavours and checking whether they are satisfied. How can this story be carried over to *mental* representations, however?

Gładziejewski and Miłkowski (2017) address this problem by appeal to the neomechanistic account of explanation, according to which cognitive systems are understood as a collection of mechanisms (see Bechtel 2008). A mechanism in this sense can be understood as "a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel & Abrahamsen 2005, p.423). The "functioning of the mechanism is responsible for one or more phenomena," which are typically understood in cognitive systems as *capacities* (Bechtel & Abrahamsen 2005, p.423). Mechanisms are thus individuated at least in part functionally: they are mechanisms *of* perception, attention, spatial navigation, motor control, and so on. As such, these mechanisms can *malfunction*: "any mechanism responsible for some capacity C… can *fail* to realize or enable C" (Gładziejewski & Miłkowski 2017, p.341).

Gładziejewski and Miłkowski argue that proper functioning in psychological mechanisms that implicate internal models is *causally dependent* on the structural similarity between such

models and target domains in just the manner outlined above: systematic interventions on this similarity relation would be associated with changes in the degree to which the mechanism successfully implements its function. For example, they argue that the successful operation of the neural mechanism that underlies spatial navigation in rats (and other mammals) is causally dependent on the structural similarity between neurally realised "cognitive maps" and target domains (Gładziejewski & Miłkowski 2017)

Of course, the appeal to the neomechanistic theory of explanation is not strictly necessary here. *Any* theory of psychological explanation that associates success conditions with psychological capacities would suffice.

What form of success is relevant to the mental models of the sort that I have outlined here? Again, if we take our lead from Craik, their core function—the central capacity for which they are responsible—is *prediction*. On this reading, then, it is our predictive capacities that are causally dependent on the structural similarity between our mental models and target regularities in the world. In the next chapter we will see research in cognitive neuroscience and machine learning that capitalizes on this link between accuracy and predictive success to explain how cognitive systems can leverage *failures* of prediction to install accurate models of the world.

Nevertheless, organisms do not try to successfully predict the world just for the sake of it. If prediction is the function of mental models, then, this must be because prediction serves other deeper functions—most fundamentally, the relevant organism's survival and reproductive success. In this way one can imagine a cascade of different levels of dependence: fundamental biological functions are causally dependent on the successful exercise of psychological capacities, which are themselves causally dependent on successful prediction, which is causally dependent on the structural similarity between mental models and worldly targets.

Exactly how to tell such a story will depend on the relevant scientific details. The point that I want to stress here is that such a story *can* be told. The basic schema whereby a representation-user exploits the structure-preserving mapping between a model and target in the service of success can be understood—in principle, at least—when it comes to *mental* models.

(5.)**6. Summary and Conclusion**

In this chapter I have argued that the core functional profile of model-based representation can be transposed into the head. In Chapter 2 I argued that one of two chief constraints on a theory

of mental representation is that it must explain—at least in broad outline—how the postulation of mental representations can contribute to the explanation of psychological capacities. It should now be clear how the basic account outlined in this chapter satisfies this constraint: if an agent's psychological capacities are causally dependent on the accuracy of its mental models, then positing such models provides a way of genuinely *explaining*—at least in part—its psychological capacities. In this way cognitive mechanisms reliant on mental models are causally indebted to the representational properties of their component parts.

Of course, so far this has been highly theoretical. At best, what I have shown in this chapter is that a vision of the mind/brain as a predictive modelling engine is *coherent*. We have not yet seen any positive reason for *endorsing* this vision, however. It is to this topic that I turn in the next chapter.

Chapter 6. Generative Models and the Predictive Mind

> "The last decade of research in machine learning and in computational cognitive neuroscience displays a pleasing convergence on an explanatory theme. That theme is the use of generative models, trained and deployed within a broad framework of prediction-driven learning and response, as a general model of cortical sensory processing" (Clark 2012, p.758).

(6.)**1. Introduction**

In Chapter 4 I argued that models should be understood as idealised structural surrogates for target domains. In Chapter 5 I argued that there is no *a priori* reason why a complex biophysical system like the brain could not command such models. That is, it is at least coherent to claim that neural networks build models of the causal and statistical structure of environmental domains in a form that can be exploited for prediction. So far this has been highly theoretical, however. My aim in this chapter is to introduce the contemporary formal ideas and empirical research that clarify and support a vision of the mind as a predictive modelling engine.

The chapter has three parts.

In Section 2 I introduce generative models and outline their computational attractions. In Section 3 I introduce probabilistic graphical models as an influential way of understanding the structure of generative models. In Section 4 I then turn to recent work in cognitive and computational neuroscience that models the neocortex as a hierarchically structured prediction machine. I postpone to the next chapter a detailed explanation of how this research maps onto the three insights that I attributed to Craik in Chapter 3.

Although I have done my best to present this material in an accessible way, this chapter is by far the most technical in the thesis. I have included it to anchor the otherwise highly abstract and theoretical discussions from other chapters in concrete research paradigms in the contemporary cognitive sciences. Nevertheless, some of the technical material could potentially be ignored without a great loss of understanding in forthcoming chapters. As such, I have concluded each section with a concise summary of those ideas that will be most relevant in what follows.

(6.)**2. Generative Models**

Although generative models are central to a growing body of work in cognitive science and machine learning, they have so far not received much attention in the philosophical literature.[11]

Conceptually, a generative model can be understood as a compact structure capable of generating a range of phenomena in a way that models (when accurate) the actual process by which those phenomena are generated. In most scientific and engineering applications, the relevant phenomena are *data*—for example, the patterns of activity at the input nodes of an artificial neural network—and the process is *causal* (Kiefer & Hohwy in press, pp.3-4). Causal generative models in this sense represent the process by which a domain gives rise to—generates—a set of observations. A paradigmatic example of such a generative model is a graphics engine of the sort used in designing video games, which renders a realistic two-dimensional image from a specification of the values of latent variables describing both the state and structure of the virtual environment (e.g. shapes, textures, positions, lighting, etc.) and the player's actions (Ullman et al. 2017).

(6.)2.1. Discriminative Models

In statistics and machine learning, generative models are typically contrasted with discriminative models (Jebara 2002; Goodfellow et al. 2016; Hinton 2007). Discriminative models represent the dependence of unobserved target variables on observed variables and are widely used for pattern recognition tasks involving classification or regression (Goodfellow et al. 2016). For example, in a classification task with images E (e.g. vectors of pixel intensities) and class labels H (e.g. "dog," "cat," "horse"), a discriminative model will determine a probability distribution over class labels given images P(H/E). Such models can be learned in purely feedforward artificial neural networks by systematically adjusting the configuration of weighted connections between neuron-like nodes to improve the network's mapping from inputs to outputs relative to a body of labelled data (i.e. a set of example input-output pairs).

In recent deep learning methods, such input-output mappings are achieved via multiple intermediate mappings in multi-level ("deep") networks (Goodfellow et al. 2016). Networks of this kind have produced highly impressive results in recent machine learning (Lecun et al. 2015). Given this, one might think that they provide an attractive model of *our* perceptual systems. Like such networks, our perceptual systems have to map proximal sensory inputs of various kinds onto appropriate categorisations of objects, our perceptual systems are

[11] Important exceptions are Clark (2016) and Kiefer and Hohwy (2017).

hierarchically structured (see below), and certain deep feedforward neural networks trained on image classification tasks have been shown to reproduce important characteristics of the response profiles of neuronal populations in the ventral stream of the visual cortex (Lecun et al. 2015, p.439).

Nevertheless, there are several important reasons why perception cannot be modelled in this way. I will mention just two.[12]

First, learning in purely feedforward discriminative models is *supervised* (Goodfellow et al. 2016). That is, they learn the input-output mapping through a gradual process in which the model's output ("prediction") for a given input is compared against the *correct* output as specified relative to a set of training examples (for example, image-label pairs). Mismatches between the network's outputs and correct outputs are thus leveraged as a training signal to adjust the model's parameters. Not only does this process typically require an enormous body of labelled data, but it is obviously not how the brain learns the structure of the distal environment. As Friston (2002, p.9) puts it, "representational learning in the brain has to proceed without any information about the processes generating inputs." As such, "when there is no *external* teaching signal to be matched, some *other* goal [i.e. other than classification] is required to force the hidden units to extract underlying structure" (Hinton et al. 1995, p.1158, my emphasis).

Second, such discriminative models cannot account for the role of top-down effects in perception.[13] There are really several issues here. Anatomically, such models offer no account of "top-down" or "feedback" synaptic connections in the neocortex. Such connections are ubiquitous, however, with connections between cortical areas almost always bi-directional (Mumford 1992). Functionally, they offer no account of the role of top-down influences in perception, despite the fact that such influences are well-documented both *within* perceptual systems and from outside of perceptual systems (see Teufel & Nanay 2017 for a review). Finally, such models also provide no explanation of endogenously generated perceptual phenomena in mental imagery, dreams, and so on, the neural implementation of which is

[12] See Friston (2003; 2005) and Hinton (2007; 2010) for other problems.

[13] It is important to note that "top-down" here does not merely mean *knowledge-driven* or *inferential*: computations in such networks and in purely feedforward models of perception more generally are typically *ampliative*, drawing on information not contained in the input data (Teufel & Nanay 2017, p.21). Instead, "top-down" refers to information from *higher levels* being used to modulate responses at lower levels.

known to overlap substantially with ordinary "online" perceptual representation (Hinton 2007; Moulton & Kosslyn 2009).

(6.)**2.2. Unsupervised Learning and Analysis by Synthesis**

Generative models have been proposed as a solution to both of these problems (see Friston 2002; 2003; 2005; Hinton 2007). Whereas discriminative models map observed variables onto unobserved variables, generative models represent the process by which unobserved variables give rise to—generate—the values of observed variables. If one thinks of the values of observed variables as given by proximal sensory inputs and unobserved variables as representations of their environmental causes, then "instead of trying to find functions of the inputs that predict the cause they [i.e. generative models] find functions of causal estimates that predict the inputs" (Friston 2002, p.231). In the context of perception, generative models can therefore seem strange. It is obvious why perceptual systems would need to implement a mapping from proximal sensory inputs onto accurate estimates of environmental states. It is much less obvious why they would need to learn the inverse mapping.

Nevertheless, several researchers have argued that generative models provide a unifying framework for understanding unsupervised learning in the brain (Friston 2005; Hinton 2007). Although there are many different applications of generative model-based learning in the machine learning literature and the technical details can sometimes be formidable, the basic idea in many applications is relatively straightforward: because generative models can be used to generate expected data, one can compare such expectations with the actual data to yield a mismatch signal that can be recruited for training the generative model. In this way generative models suggest "a sensible objective for learning: adjust the weights on the top-down connections to maximize the probability that the network would generate the training data" (Hinton 2007, p.428).

To see this, recall the problem of learning for discriminative models: because their task is to map input data onto appropriate outputs (e.g. a categorisation of the input data), they need feedback on what the correct output is. All they have access to, however, is the data. As such, the feedback must come from some other source: a teaching signal generated from a body of labelled training examples that specify the correct mapping. With generative models, by contrast, their task is quite different. To learn an accurate generative model of some training data means identifying the interactions among latent variables responsible for generating that data. A good generative model, therefore, should be able to *reconstruct* the data. Crucially, this

means that the training signal can come from the *input itself*—or, more accurately, from the mismatch between the data and the data expected by the generative model. Such mismatches or *prediction errors* provide what Clark (2016, p.19) calls a "rich, free, constantly available, bootstrap-friendly, teaching signal" in the form of the input that the network is exposed to. As Lotter et al. (2015, p.1) put it, "not only do prediction errors provide a source of constant feedback, but successful prediction fundamentally requires a robust internal model of the world." As such,

> "a critical idea… for how substantial amounts of learning can emerge from the largely passive sensory experience of babies, is that each moment can be turned into a predictive learning problem: learning to predict what visual input will arise next, given what has come before" (O'Reilly et al. 2017, p.2).

With generative models, then, the trick is to train multilayer neural networks "to generate sensory data rather than to classify it" (Hinton 2007, p.428).

In addition, generative models offer a principled account of top-down effects in perception. At the "implementational" level, it is widely supposed that they reside "in the top-down connections between cortical areas" (Hinton 2010, p.180) such that, for example, "top-down connections in the [visual] cortical hierarchy capture essential aspects of how the activities of neurons in primary sensory areas are generated by the contents of visually observed scenes" (Dayan 1997, p.43). In this way sensory cortices implement causal models of the process by which sensory inputs are generated by interactions among hidden features of the world.

Functionally, these generative models thus offer an attractive implementation of classic "analysis-by-synthesis" models of perception whereby sensory inputs are interpreted by modelling the causal process from which they are generated (Nair et al. 2008; Yildirim et al. 2015). As we will see below, because such generative models in effect encode the *likelihood* of different patterns of evidence E conditional on hypotheses about their causes H P(E/H), they can be combined with a prior distribution over states of their causes P(H) to perform Bayesian inference, a statistically optimal method of belief updating under uncertainty, thereby offering a principled account of the role of contextual and top-down effects in online perception (see Kersten et al. 2004). Finally, the *generative* character of such models offers a straightforward explanation of endogenously generated perceptual phenomena in the absence of driving sensory input (Hinton 2007). It is constitutive of a generative model that it can autonomously generate the range of phenomena proprietary to its target domain.

In this way the concept of a generative model provides an attractive—albeit so far highly schematic—framework for understanding how neural networks in the brain might reconfigure themselves to capture the hidden structure of the world responsible for generating their proximal sensory inputs and then exploit this knowledge in perception and imagination.

(6.)**2.3. Hierarchical Generative Models**

As noted above, discriminative models used in machine learning typically involve hierarchical task-decomposition, whereby the mapping from inputs to outputs is achieved via a succession of intermediate mappings. This principle of hierarchical representation is carried over to most interesting applications of generative models in machine learning and cognitive science. Hierarchical generative models implement a recursive process whereby interactions among phenomena at each level L are generated by interactions among phenomena represented at the level above L+1, such that—in the case of causal generative models, at least—the target domain is modelled "as a hierarchy of systems where supraordinate causes induce and moderate changes in subordinate causes" (Friston 2005, p.822).

Hierarchical representation of this kind is important for several reasons.

First, information processing in the neocortex is hierarchical (Felleman & Van Essen 1991; Kandel et al. 2013). To simplify greatly, sensory information first arrives in primary sensory areas from the thalamus, where neuronal populations are sensitive to simple fine-grained stimulus features at high spatiotemporal resolution. For example, the primary visual cortex processes information about stimulus features such as oriented edges, small-scale aspects of motion, and simple colour and contrast information. Primary sensory areas are then connected to higher cortical areas concerned with more abstract kinds of information and environmental phenomena at greater spatiotemporal scales. In the context of a generative understanding of perception, one can understand such higher areas as modelling those features of the environment responsible for generating the regularities registered at the level below (Friston 2005; see below). In addition, the cortex's motor system is also hierarchically organised in a similar manner, with the lowest level—the primary motor cortex—most directly connected to the spinal cord and the primary motor cortex in turn connected to a cascade of higher regions of motor cortex performing increasingly abstract motor functions (Kandel et al. 2013).

Second, hierarchical structure in which regularities are nested within regularities and parts compose wholes that in turn yield larger wholes is also a pervasive feature of the world that

cognitive systems must contend with (Clark 2016). This structure is reflected in the statistical properties of natural signals. As Lecun et al. (2015, p.439) put it,

> "Deep [i.e. hierarchical] neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences."

Finally, hierarchical representation in generative models affords important computational benefits. For example, such models typically make use of conditional independence relations between non-adjacent levels of the hierarchy (see Section 3 below), such that the activity of each level is only directly dependent on activity at adjacent levels in the hierarchy. As we will see below, this kind of architecture makes complex networks of mutual relevance computationally tractable, such that representations at the higher levels of the hierarchy need not directly generate expectations with respect to the phenomena registered at the lowest levels. Further, this hierarchical profile also allows for important top-down effects on learning from higher-level more schematic information about the structure of a domain. For example, it often happens in learning that we acquire a broad understanding of the general structure of a domain before filling in the details with respect to concrete phenomena, in a way that can be understood in terms of top-down constraints on representational learning at lower levels from higher levels of a generative model (Tenenbaum et al. 2011).

(6.)**2.4. Summary**

To summarise, the chief points of this section are as follows:

- Generative models constitute a representational structure that captures how interactions among features of a target domain generate some phenomena of interest, typically understood as data or observations.
- Generative models facilitate unsupervised learning by leveraging the mismatch or "prediction error" between expected ("synthetic") data and input data.
- Generative models provide a unifying framework for understanding top-down effects on perception. Top-down connections implement a causal model of how sensory inputs are generated that can both be combined with sensory inputs in online perception and exploited for purely "offline" perceptual phenomena.

- Most interesting generative models are *hierarchical*, where this typically involves a hierarchy *of* generative models such that the phenomena represented at each level are dependent upon interactions among phenomena represented at the level above.

Before continuing, it is worth noting three important additions to the basic story just canvassed.

First, I have spoken of sensory inputs being generated by interactions among features of the world. In the case of embodied organisms, however, *their own actions* are a crucial part of this generative process. To take a familiar example, frequent (and typically unconscious) visual saccades generate radical changes in the visual input that we receive moment by moment. As such, generative models for perception in the *brain* must also capture the way in which the organism interacts with features of the world to generate sensory evidence.

Second, I have written mostly about visual perception in this section, but of course the problem of learning and exploiting information about the causes of sensory inputs applies across all the (exteroceptive, interoceptive, and proprioceptive) sensory modalities. Further, not only does this variety of modalities put us into contact with *different* aspects of the bodily and environmental milieu, but they can also put us into contact with the *same* features of the body and environment through different channels. For example, we can pat, smell, see, and hear a dog. If there are generative models in the brain, then, they should capture the way in which features of the world reliably generate systematically correlated (mutually informative) patterns of activity across the different sensory modalities.

Finally, although I have focused on cases of perception in the form of relations among worldly states and proximal sensory inputs, the basic functions of causal generative model-based representation outlined here can be exploited whenever there is some systematic process whereby interactions among the elements of a domain generate effects that we have access to (Tenenbaum et al. 2011). As such, generative models can be useful beyond typical cases of perception. For example, in addition to the way in which interactions among features of the world generate our proximal sensory inputs, there is also a systematic process by which interactions among features of the world generate novel *states of the world*. To return to an example used above: in game engines used to design interactive video games, in addition to a generative *graphics engine* that dynamically renders a two-dimensional image from a specification of scene parameters and the player's actions, there is also typically a generative *physics engine* that generates new states of the (virtual) physical environment from its previous

states and any interventions upon or alterations to the environment (Battaglia et al. 2013; Ullman et al. 2017). I will return to this strategy of generative model-based physical prediction in the next chapter.

The discussion in this section has been highly abstract. In the next section I introduce a useful way of understanding the structure of generative models in many of their applications in contemporary machine learning and cognitive science.

(6.)**3. Graphical Models**

> "The variables describing the structures in the world are typically related in a graphical fashion, edges connecting variables which have direct bearing on each other" (Mumford 2002, p.411)

Graphical models were introduced into cognitive science and artificial intelligence largely by Judea Pearl's seminal work (Pearl 1988; see Russell & Norvig 2010). These data structures are graphical because they express the set of dependencies between the elements of a domain with a *graph*, a mathematical structure of *nodes* and *edges* that can be directed (A→B), undirected (A—B), or bidirected (A←→).

Such graphs effectively comprise compact representations of *relevance* relationships between the elements of a domain, where different kinds of graphs capture different forms of relevance (causal, statistical, informational, communicative, etc.) (Danks 2014, p.39). As Danks (2014, p.39) puts it, such models "thus address a key challenge for any cognitive agent: namely, determining what matters and, often equally importantly, what can be ignored." The graph itself captures the *qualitative* structure or gross "topology" of a domain: namely, the pattern of relevance relationships between its elements. In most cases, this qualitative structure is complemented with quantitative parameters that represent the strength of the connections between these nodes. For example, in the graphical models that will concern us, the nodes represent variables and the edges represent causal and probabilistic dependencies between them. As such, the qualitative information encoded in the graph must be complemented with a set of probability distributions that quantify how the probabilities assigned to some variables systematically depend on the values of others. For a visual illustration of these ideas, see Figure 4.

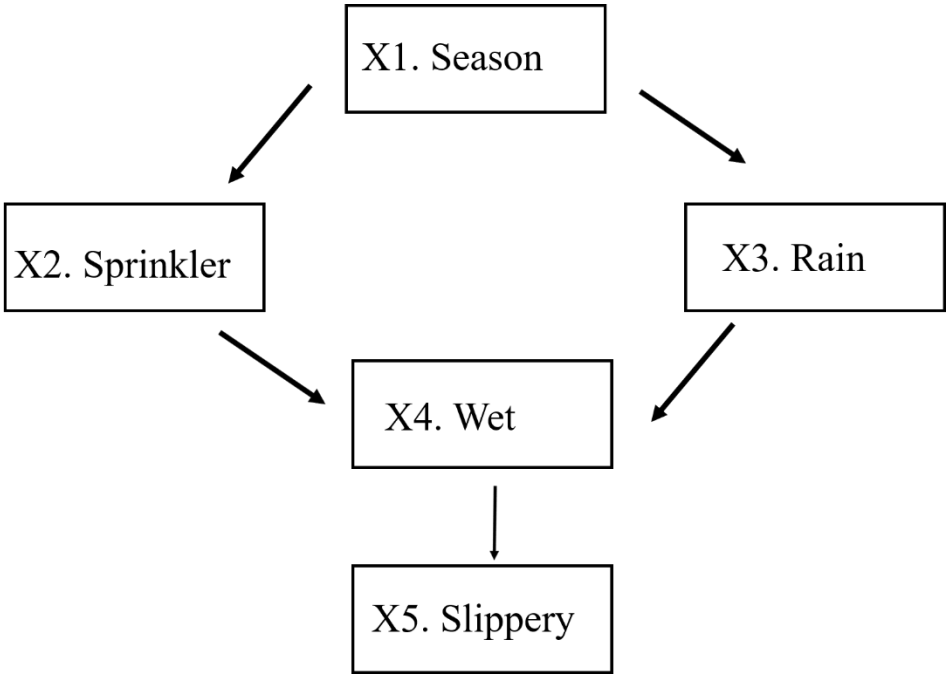

Figure 4. A visual illustration adapted from Pearl (2000) of a simple graphical model capturing the structure of a set of dependencies between (1) what season it is, (2) whether the sprinkler is on, (3) whether rain falls, (4) whether the pavement is wet, and (5) whether it is slippery. In typical models this qualitative structure will be combined with parameters that quantify the probabilistic strength of these dependencies—for example, how probable rain is given different seasons P(X3/X1).

There are several reasons why it is important to include probability in such models (see Pearl 1988). First, often we can only observe a limited part of a target system, and these observed states are consistent with a variety of different hidden states, distinguished only by their relative probability of occurrence. In perception, for example, proximal sensory inputs are notoriously consistent with many possible environmental causes. Second, observations themselves are often noisy (i.e. vulnerable to corruption by random errors and thus not perfectly reliable), which introduces uncertainty. Third, almost all models are highly selective in the manner described in chapter 4 in a way that inevitably introduces stochasticity into the model. Finally, perhaps some processes in the world at the spatiotemporal scale we model them simply are stochastic.

What makes graphical models so useful is their graphical component. Imagine, for example, that you want to model a domain that contains 100 variables—a meagre size when contrasted with the complexity of the world that we must navigate. A joint distribution over such a set of variables will thus be *100-dimensional*, which is extremely difficult both to learn and manipulate (Jacobs and Kruschke 2010, p.10). What a graph enables you to do is capture the

probabilistic dependencies between these variables in terms of *direct relevance* (i.e. dependence) (Pearl 1988). Specifically, graphical models capitalize on the fact that whilst most things are in some way informationally relevant to most other things, they are typically not *directly* relevant to most other things (Danks 2014). For example, which country someone lives in is relevant to how wealthy they are: information about the former can be used to predict the latter to some degree of precision with some reliability. But it is not *directly* relevant: once one factors in other information—their assets, the amount of money in their bank account, and so on—the knowledge of where they live becomes uninformative. In the vocabulary of probability theory, such variables are *conditionally independent* given observation of these other variables.

In a graph, this notion of direct relevance (and its inverse *conditional independence*) is captured in its edges. Roughly, if nodes are adjacent in the graph—that is, they are connected to one another by some edge between them—they are directly relevant to one another; if not, they are conditionally independent. As such, one can effectively *factorize* a high-dimensional joint probability distribution into a set of much lower-dimensional distributions in which the probability assignment of each variable depends only on the values of adjacent nodes (Jakobs and Kruschke 2010). This factorization comes with considerable representational and computational benefits (Russell & Norvig 2010).

(6.)**3.1. Bayesian Networks**

Crucially, graphical models of this kind can be used for encoding generative models (Danks 2014, pp.44-5; Goodfellow et al. 2016). An important example of graphical models for encoding causal generative models of the sort outlined above are *Bayesian networks*. These are models in which the nodes comprise random variables and the edges are *directed* (Pearl 1988; Pearl 2000).[14] An edge is directed if the dependency it captures is asymmetric. For example, whether one has a runny nose depends on whether one has the flu, but not vice versa. This makes Bayesian networks important for capturing causal structure where such asymmetries are central (see Chapter 5). Further, these directed edges typically capture probabilistic relationships: smoking increases the risk of lung cancer, but one can smoke without getting lung cancer. The qualitative component—the graph—of such graphical models thus effectively captures the way in which effects depend on causes; the quantitative component—the parameters—captures the strength of these causal dependencies.

[14] It is sometimes often held that Bayesian networks must also be acyclic and have discrete values (e.g. Danks 2014), complications I will ignore here.

Directed probabilistic graphical models of this kind have been central to much work in perceptual psychology, especially Bayesian perceptual psychology (see below). For example, they can be used to represent "how an image description I (e.g. the image intensity values or features extracted from them) is determined by the scene description S (e.g. vector with variable values representing surface shape, material reflectivity, illumination direction, and viewpoint)" (Kersten et al. 2004, p.275-6). As Kersten and Yuille (2003, p.152) put it,

> "For a given [perceptual] problem we can begin by characterizing how variables in the world influence each other and the resulting image features by representing the distribution with an influence diagram [i.e. graph] in which the nodes correspond to the variables… and we draw links between the nodes that directly influence each other."

As we will see below (Section 4), such generative models can be combined with prior expectations about the probable states of latent variables to infer the most probable explanation of incoming sensory data.

(6.)**3.2. Summary**

Graphical models have proven to be enormously influential representational structures across both artificial intelligence and the cognitive sciences. The machine learning theorist Geoffrey Hinton (2005) titles a paper "What kind of graphical model is the brain?", Dennett (2017, p.161) confidently speculates that "the fundamental architecture of animal brains (including human brains) is probably composed of Bayesian networks," and Butz and Kutter (2017, p.214) note that "various researchers…consider the brain to approximate a dynamic, highly distributed, hierarchical, and modularized Bayesian network." Importantly, the simple examples above should not mislead one about the complexity of such models if they inhere in our brains. As Hinton (2005, p.1765) puts it, "if neurons are treated as latent variables, our visual systems are non-linear, densely-connected graphical models containing billions of variables and thousands of billions of parameters."

The attractions of probabilistic graphical models are not difficult to see. They facilitate sophisticated forms of probabilistic inference, can be learned from data without labelled examples, capture complex distributions of relevance relations whilst retaining computational tractability through the clever use of conditional independence structures, and have been shown in many cases to map onto the structure of biological neural networks (see Danks 2014; Hinton 2005; Russell & Norvig 2014). As Russell and Norvig (2014, p.26) put it, graphical models combine "the best of classical AI and [artificial] neural nets."

For my purposes in this thesis, the important characteristics of graphical models are as follows:

- Graphical models capture the dependencies that exist between the elements of a domain with a structure of nodes and edges. In the models that will concern us, these nodes stand in for random variables and the edges stand in for causal (i.e. asymmetric) and statistical (i.e. symmetric) dependencies between them.
- Probabilistic graphical models such as Bayesian networks can be used for implementing generative models—for example, models of the process by which sensory data are generated by interactions among features of a domain.
- Graphical models capitalize on *conditional independence relationships* to represent complex networks of mutual influence in a way that is computationally tractable.

Importantly, although graphical models provide an especially useful way of understanding the structure of many kinds of generative models, they do not provide the only representational format for encoding generative models (Goodman et al. 2015). I will return to this point briefly in Chapter 11. In most of what follows, however, it will helpful to have in mind the structure of nodes and edges standing in for variables and causal and statistical dependencies when it comes to understanding the characteristics of generative models. By making the structure of such generative models explicit, it becomes easier to grasp what it means for there to be a structure-preserving mapping between such models and the causal-statistical structure of target domains.

(6.)**4. The Neocortex as a Hierarchical Prediction Machine**

> "Predictive coding may be applicable across different brain regions and modalities, providing a useful framework for understanding the general structure and function of the neocortex" (Rao & Ballard 1999, p.85).

So far, I have explored generative models from a conceptual and formal perspective. What relevance does this have for our contemporary understanding of the brain and the computational mechanisms underlying intelligence?

Generative models have been central to several recent developments in cognitive science. In addition to advances in unsupervised learning, for example, they are central to the research programme of "Bayesian cognitive science," which advocates "an approach to understanding cognition and its origins in terms of Bayesian inference over richly structured, hierarchical generative models" (Tenenbaum et al. 2011, p.1285). Further, generative models are also

central to Rick Grush's (1997; 2004) important "emulation theory of representation," which he derives from research in cognitive neuroscience and control theory. Although Grush does not speak explicitly of generative models, the emulator representations he characterises are just that: predictive or "forward" dynamic models capable of simulating the likely outcomes of different processes. (They are thus capable of generating mock sensory signals from a model of the underlying variables that give rise to them).

In this section I am going to focus on a theory of cortical information processing that relates to both sets of ideas: predictive processing. First, I will outline a conception of the neocortex as a general-purpose learning device (S4.1). I will then introduce hierarchical predictive coding (S4.2) and its relation to Bayesian inference (S4.3 and S4.4) before briefly summarising the chief claims of—and evidence for—predictive processing in Sections 4.5 and 4.6.

(6.)**4.1. The Neocortex as a General-Purpose Learning Device**

The neocortex is a thin sheet of neural tissue roughly 2.5 millimetres thick that envelops most of our brains. It is found in all mammals and associated with psychological functions such as perception, imagination, planning, reasoning, and—in humans, at least—much of action control (Kandel et al. 2013). Indeed, the strong relationship between mental capacities and cortical activity has led some neuroscientists to simply identify the mind with the neocortex, or the thalamocortical system more broadly (Hawkins & Blakesee 2004, p.40). Even if one rejects such "neocortical chauvinism," however, its central role in subserving human mentality is undeniable. If one wants to understand the fundamental nature of mental representation, then, the neocortex is the most important part of the brain to understand (Ryder 2004).

This task can seem formidable. The human neocortex appears bewilderingly complex, containing roughly 20 billion neurons of hundreds of different kinds, each forming thousands of synaptic connections, with signalling mediated by a staggering diversity of neurotransmitters. Further, the number of different functions it is either wholly or partly responsible for is equally heterogeneous. That is, there does not seem to be anything very *general* to say about it.

Nevertheless, many neuroscientists have sought "global" theories of the neocortex—theories in which its superficial complexity and functional diversity arise from simple underlying principles, in much the same way that the apparent complexity of planetary orbits emerges from comparatively simple (and mathematically explicable) physical laws (Bastos et al. 2012; Douglas and Martin 2007; Grossberg 2000; Mountcastle 1978; Perin et al. 2011).

There are several observations that have encouraged this search (see Hawkins & Blakesee 2004; Ryder forthcoming). I will mention just two.

First, the neocortex is radically plastic. This plasticity is evidenced both in the enormous structural changes its configuration of synaptic connections undergoes as a function of experience, the consequent extent and potential of learning that it facilitates, and the surprising capacity of cortical regions usually associated with one cognitive function to adapt to a distinct function when delivered novel kinds of input (Kandel et al. 2013). To take just one example, in a famous study in which the cortex of new-born ferrets was surgically altered so that their previously auditory regions were fed signals from their eyes, the ferrets developed functional visual pathways and capacities in that region of "auditory" cortex (Roe et al. 1992).

Of course, plasticity itself cannot motivate the search for a global theory of the neocortex. For example, it may be that such plasticity is underpinned by several distinct learning algorithms localised to different functional regions. A second—and crucial—observation, however, has encouraged many neuroscientists to search for such a theory: the cortex's surprising structural uniformity across different functional regions, both within species and across all mammals (Douglas and Martin 2007; Mountcastle 1978). Specifically, despite the wide diversity of cognitive functions subserved by the neocortex, many neuroscientists have argued that this diversity is not—as one might expect—realised by computationally different kinds of cortical networks, but rather seems to emerge from the relative input-output profile of different functional areas (Douglas and Martin 2007). Of course, it is impossible to make claims in this area that are not extremely controversial (see Marcus et al. 2014 for an alternative perspective). Nevertheless, an influential view is that "the developing cerebral cortex is largely free of domain-specific structure" (Quarts & Sejnowski 1997, p.537). That is, what fundamentally differentiates visual cortex from auditory cortex from somatosensory cortex from different regions of association cortex is simply the kinds of data to which they are exposed, and not qualitatively different information-processing architectures.

In combination with each other, then, these two observations have encouraged many neuroscientists to search for a common underlying algorithm (a "canonical cortical computation") of the neocortex—an algorithm that underlies not just learning, but also the kinds of "online" cognitive processes that facilitate real-time adaptive success:

> "One simplifying hypothesis that has existed since the time of Cajal is that the neocortex consists of repeated copies of the same fundamental circuit" (Douglas and Martin 2007, p.226).

> "All parts of the neocortex operate based on a common principle, with the cortical column being the unit of computation" (Mountcastle 1978, p.7).
>
> "The uniformity and highly specific layered structure of the neocortex of mammals suggests that some quite universal computational ideas are embodied by this architecture" (Mumford 1992, p.241).
>
> "Much of the mammalian brain might use a single algorithm to solve most of the different tasks that the brain solves" (Goodfellow et al. 2016, p.29).
>
> "The uniformity of the cortical architecture and the ability of functions to move to different areas of cortex following early damage strongly suggest that there is a single basic learning algorithm for extracting underlying structure from richly structured, high-dimensional sensory data" (Hinton 2007, p.535).

According to an increasingly influential view in cognitive and computational neuroscience, this basic cortical algorithm is predictive coding (Bastos et al. 2012; Friston 2005; Rao & Ballard 1999).

(6.)**4.2. Predictive Coding**

"Predictive coding" names an encoding strategy in which only the unpredicted elements of a signal are fed forward for further stages of information processing (Clark 2013). This provides a highly efficient mechanism for reducing redundancy by removing the predictable elements of the input signal (Huang & Rao 2011). According to what I will call "predictive processing," this information-processing strategy captures the fundamental nature of message passing in the hierarchically structured neocortex (Bastos et al., 2012; Clark, 2016; Friston, 2005; Seth, 2015).[15] Specifically, its proposal is that backwards synaptic connections from higher cortical areas attempt to predict the activity at the level below them in the cortical hierarchy, and that only the unpredicted elements of the lower-level activity—the prediction errors—are fed forward (Rao & Ballard 1999; Huang & Rao 2011). In this way each level in effect encodes a generative model of the level below (Friston 2005).

Crucially, these prediction errors provide the brain with an internally accessible quantity to minimize. According to predictive processing, it is by constantly attempting to *minimize prediction error* that the neocortex both installs a hierarchical generative model of the interacting features of the body and environment responsible for generating sensory inputs and

[15] Following Clark (2016), I use "predictive processing" to describe the use of predictive coding within the context of a hierarchically structured probabilistic generative model.

then exploits this model in the service of what is often called "perceptual inference": namely, identifying the state of the body and environment responsible for causing the brain's sensory inputs "in real time." In this way "the problem of inferring the causes of sensory input (perceptual inference) and learning the relationship between input and cause (perceptual learning) can be resolved using exactly the same principle": namely, *prediction error minimization* (Friston 2005, p.815).

Prediction-driven learning and perception thus enable the neocortex to "induce accurate generative models of environmental hidden causes by operating only on signals to which it has direct access: *predictions* and *prediction errors*" (Seth 2015, p.5). The counterintuitive upshot of predictive coding is therefore that forward connections in the neocortex function as *feedback* on the cortex's generative model, only broadcasting that aspect of the incoming sensory signal not already accounted for in systemic response (Friston 2005).[16]

Importantly, this process of prediction error minimization is modulated by the "precision-weighting" of prediction errors: the differential influence of prediction errors as a function of their estimated reliability (Feldman & Friston 2010). To understand how this works, it is useful to first show how predictive coding can be understood as an implementation of approximate *Bayesian inference*.

(6.)**4.3. The Bayesian Brain**

There has been something of a "Bayesian revolution" (Hahn 2014) in cognitive science in recent years—an "explosion in research applying Bayesian models to cognitive phenomena" (Chater et al. 2010, p.811). This revolution has been motivated both by a growing consensus that one of the fundamental problems that cognitive systems solve is inference and decision-making under *uncertainty*, alongside an appreciation of how Bayesian statistics and decision theory can be used to model the solutions to such problems in mathematically precise and empirically illuminating ways (Chater et al. 2010; Griffiths et al. 2010; Tenenbaum et al. 2011).

In the case of perception, for example, the brain must exploit the statistical patterns of activity at the organism's sensory transducers to identify the state and structure of the distal environment. This evidence, however, is notoriously "sparse, noisy, and ambiguous" (Tenenbaum et al. 2011, p.1279). Specifically, it dramatically underdetermines the complex

[16] For clarity of exposition and because it is not strictly relevant for my aims here, I ignore here the importance of lateral connections in Friston's (2005) model, which mediate comparisons between top-down predictions and bottom-up signals and decorrelate competing top-down predictions.

structured environmental causes that generate it and is transmitted via input channels that are vulnerable to corruption by random errors. Mainstream perceptual psychology understands this problem as an inference problem: the brain must infer the causes of its sensory inputs from the sensory inputs themselves (Helmholtz 1867). Because this is a paradigmatic case of inference under uncertainty, an influential contemporary tradition models this inferential process as statistical inference in accordance with Bayes' theorem (see Rescorla 2013).

Bayes' theorem is a derivation of probability theory that states the following:

(Bayes' Theorem) $P(H|E) = P(E|H)P(H)/P(E)$.

Bayesian perceptual psychology models the perceptual system as a mechanism that infers the causes of its sensory inputs in approximate accordance with this theorem (Geisler & Kersten 2002). In a typical model, the perceptual system assigns probabilities both to hypotheses estimating possible environmental states P(H) (its priors) and to different patterns of sensory evidence given such hypotheses (the likelihood P(E/H)). This likelihood is captured in a generative model of how such patterns of evidence are caused by—generated by—interactions among environmental states. When confronted with new evidence (i.e. activity at the organism's sensory transducers), the system thus exploits this generative model in conjunction with a prior distribution over hidden states to calculate P(H/E) in conformity with Bayes' theorem.

Bayesian perceptual psychology is mathematically precise and has produced empirically impressive models of a range of perceptual phenomena (Rescorla 2013). More generally, its emphasis on prior expectations and optimal statistical inference provides attractive explanations of a range of general features of perception: how brains overcome the noise and ambiguity in their sensory evidence, how they effectively integrate evidence from across the sensory modalities, how they generate perceptual constancies, and why they systematically produce specific perceptual illusions under conditions when strong prior expectations are violated (Chater et al. 2010; Geisler and Kersten 2002; Kersten and Yuille 2003; Rescorla 2013).

Despite these attractions, the Bayesian brain hypothesis confronts at least three big challenges. First, it is implausible that the brain explicitly follows Bayes' theorem. Exact Bayesian inference can be extremely slow and often computationally intractable (Penny 2012; Tenenbaum et al. 2011). Much work in statistics and machine learning is thus devoted to developing algorithms that approximate Bayesian inference (Penny 2012, p.2). Second, it needs

an account of where the relevant priors and likelihoods that power Bayesian inference come from. Although some priors might be specified genetically, the radical plasticity of cortical information processing outlined above suggests that this is unlikely for most of them. Finally, once we have a story of the aetiology of priors and the approximation algorithms responsible for Bayesian inference, how do such processes get implemented in the brain's neural networks?

(6.)**4.4. Precision-Weighted Prediction Error Minimization**

One way of understanding predictive coding and the imperative to minimize prediction error is as providing answers to these questions (see Hohwy 2017). To see this, note that Bayesian inference is in effect a matter of identifying the hypothesis P(H/E) that best predicts the evidence P(E/H) weighted by its prior probability P(H). If we assume the hypothesis space and evidence are Gaussian (and can thus by described by their sufficient statistics), one can calculate this by comparing the mean value *m* of the prior distribution with the mean value *e* of the evidence to compute a *prediction error*: the difference between these two values (Hohwy 2017; Seth 2015). In this way the prediction error varies inversely with the likelihood of the hypothesis in an intuitive way: the greater the hypothesis' likelihood—the better it predicts the evidence—the less prediction error it generates. Bayes' theorem then answers the question: how much should the prior be updated in light of the prediction error it generates—that is, in light of the difference between *m* and *e*?

To calculate this requires some means of weighting the relative reliability or uncertainty of the two sources of information. With Gaussian distributions, this can be computed from their relative *precisions*, where "precision" names the inverse of the variance of the two distributions. In effect, the precision of the two distributions stands in for one's *confidence* in them: the more precise one's priors, the more confident one is in them and the less they should therefore be updated in light of the prediction errors they generate. The ratio of these two quantities thus determines the *learning rate* in Bayesian inference: the more precise the priors are relative to the evidence, the less the agent learns about the world from the evidence—the less the prediction error influences the posterior—and vice versa (Hohwy 2017). Thus as one learns more about a domain, one's priors will become increasingly precise and prediction errors will be weighted less in driving inference—just as Bayes' theorem dictates. For a simple visual illustration of these ideas, see Figure 4:

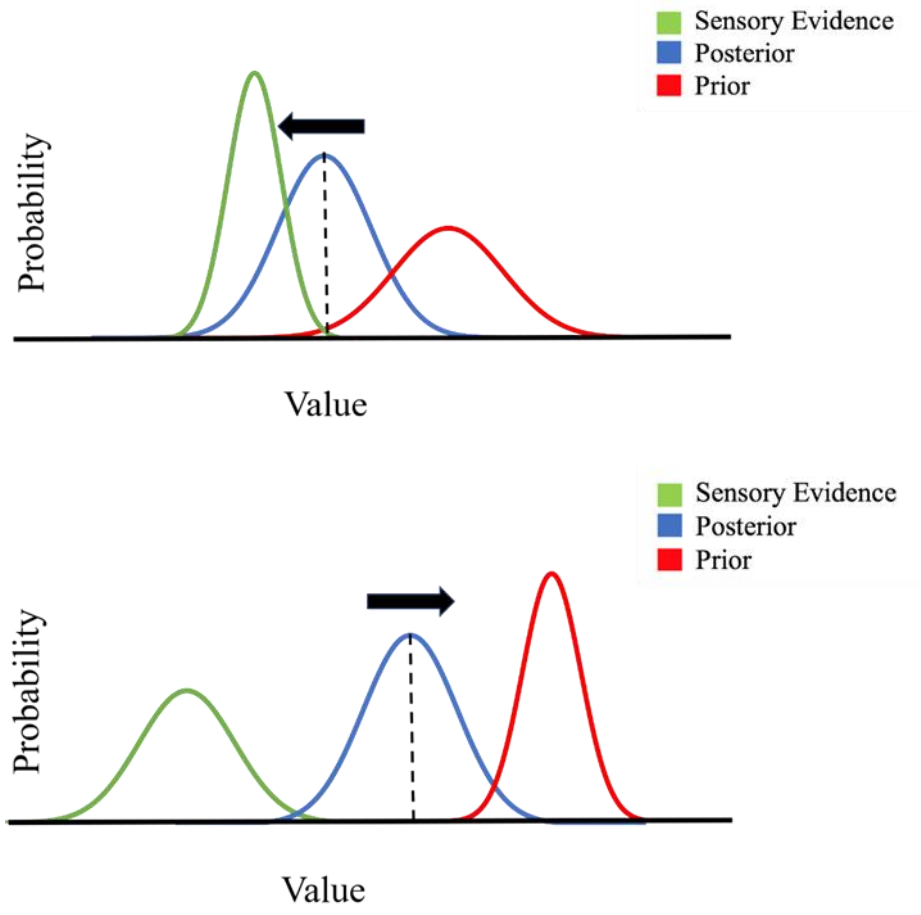

Figure 5. Posterior beliefs are a function not just of the means but also the precisions of the sensory evidence and prior distribution. On the top graph, the comparatively greater precision of the sensory evidence has a proportionately greater influence on the posterior than the prior. On the bottom, this is reversed.

As we have seen above, however, cortical representation is hierarchical, reflecting a causal process responsible for generating sensory inputs that exists at multiple levels of spatiotemporal scale and abstraction. Crucially, this cannot be captured in terms of matching a single hypothesis space against sensory evidence in the manner just described. To effectively minimize prediction error against sensory evidence, then, the brain must be responsive to this hierarchical structure. Specifically, rather than matching a single hypothesis space H against the evidence, it must build a *hierarchy* of hypothesis spaces in which each level of the hierarchy in effect encodes a generative model of regularities registered at the level below, such that hypotheses at one level L+1 provide the priors for the level below L and the evidence against which predictions errors are computed for the level above L+2 (Friston 2008; Lee & Mumford 2003). With a hierarchical model of this kind in place, the system can employ expectations

concerning regularities at slower time-scales to regulate the learning rate (i.e. the precision assigned to prediction errors) in context-sensitive ways (Hohwy 2017; Matthys et al. 2014). For example, such a system might expect highly precise visual evidence in broad daylight and highly noisy evidence during a fog, and adjust its learning rate accordingly (Clark 2016, p.58). As such, hierarchical Bayesian inference of this kind requires the system not just to form estimates of the state of the world but also estimates of the uncertainty associated with these estimates and the evidence to which they are responsive (Hohwy 2017).[17]

Predictive coding and its account of message passing between hierarchically connected levels of cortical processing can be understood as an account of how this process of hierarchical Bayesian inference gets both approximated and implemented in the neocortex. That is, the process of comparing top-down predictions of activity at one level with the level below, signalling the disparity between the two—the prediction error—and then striving to minimize this disparity over time can be understood as approximate Bayesian inference of the kind just described. At the "implementational level," an influential proposal in the literature is that top-down predictions and bottom-up prediction errors are carried via deep and superficial pyramidal cells respectively, with precision-weighting occurring through alterations to the "postsynaptic gain" of the populations of superficial pyramidal cells reporting prediction errors (Bastos et al. 2012). In turn, the mechanism responsible for altering the postsynaptic gain is thought to be at least in part the action of neuromodulators such as dopamine, serotonin, and acetylcholine (Adams et al. 2013). Further, it has been suggested that the thalamus (especially the pulvinar) also plays a crucial role in orchestrating this process of precision-weighting (Kanai et al. 2015).

(6.)**4.5. Stepping Back**

It is easy to get bogged down in the technical details here and thus miss the forest for the trees. To step back, then, predictive processing can fundamentally be understood as an answer to the following question: how do brains acquire and exploit information about the complex interacting causes responsible for generating their sensory signals given that they only have direct access to the signals themselves (Rao and Ballard 1998; Friston 2005)?

The answer that predictive processing provides is that cortical processes self-organize around prediction errors: by leveraging mismatches between internally generated top-down

[17] Crucially, these expectations concerning precision must also be learned from the overall drive to minimize prediction error, a complication I ignore here (see Hohwy 2013; 2017).

predictions and the externally generated bottom-up sensory signal, the neocortex acquires a model of the process by which incoming signals are generated. The hierarchical structure of this model reflects the fact that the generative process exists at multiple levels of spatiotemporal scale and abstraction. The model can be updated in real time by identifying those hypotheses that best predict the current sensory signal. Finally, by forming expectations about the context-dependent reliability of bottom-up and top-down information, it can modulate the relative influence of prediction errors as a function of their precision. In this way the neocortex in effect functions as a general-purpose learning device operating according to prediction error-based updating.

What is the evidence for this extremely ambitious view of cortical information processing?

Although a proper answer to this question falls beyond the scope of the current thesis, one can identify several general motivations for taking its explanatory ambition seriously (see Clark 2016 for a review).

First, there are simulations of artificial neural networks either explicitly following this approach (e.g. Chalasani & Principe 2013; Lotter et al. 2015; 2016; Rao and Ballard 1999) or closely related approaches in artificial intelligence involving generative models and prediction-based learning (Dayan et al. 1995) that have exhibited highly impressive performance and response characteristics that map onto observed cortical activity.

Second, although predictive processing is advanced approximately at Marr's (1982) "algorithmic" level of description (Clark 2016, 319), the mechanisms it requires—prolific feedback connections, hierarchical representation, functionally distinct units for predictions and prediction errors, asymmetric roles for bottom-up and top-down synaptic connections, and gating mechanisms for adjusting the influence of prediction errors as a function of their precision—are thought by at least some to map on convincingly to brain structures (Friston 2002; 2003; 2005; Gordon et al. 2017; Weilnhammer et al. 2017).

Third, as noted above, its commitment to a single principle of cortical functioning exploits evidence for the surprising uniformity and radical plasticity of the neocortex, alongside a widely held conviction that what functionally differentiates cortical areas is their input-output profile, not their computational architecture (George and Hawkins 2009; Hawkins and Blakesee 2004; Mountcastle 1978).

Finally, when these general motivations are combined with both the increasing quantity of applications of the theory to different psychological and neurocognitive phenomena (see Clark 2016) and the immense popularity of Bayesian modelling in cognitive science (Chater et al. 2010), it is not surprising that predictive processing is currently enjoying the level of interest and excitement it does.

(6.)**4.6. Summary**

To summarise, the chief lessons from this section are as follows:

- The neocortex is the most important place to look for an understanding of the nature of mental representation.
- The surprising uniformity and radical plasticity of the neocortex have suggested to many neuroscientists that it functions as a general-purpose learning device that operates according to a common algorithm.
- According to predictive processing, this common algorithm is predictive coding, whereby higher cortical areas send predictions of the activity at the level below and only the unpredicted element of that lower-level activity—the prediction error—is sent forwards.
- By *minimizing prediction error* in a way that is responsive to the context-dependent precision or reliability of prediction error signals, the neocortex learns and updates a model of the structure of the world based only on signals to which it has access.
- This process of precision-weighted prediction error minimization implements a form of approximate Bayesian inference.

Importantly, so far I have only appealed to these ideas to account for learning and perceptual inference. In the next chapter, however, we will see how this basic account of the neocortex as a hierarchically structured prediction machine can be extended to capture other psychological capacities.

Before concluding, I should note that the exclusive focus on the neocortex in this section is of course slightly distortive. As predictive processing is usually presented, for example, it is a global theory of *brain* function, not just cortical function (Hohwy 2013), and other neural structures are hypothesised to play important functional roles in the overall process of

prediction error minimization.[18] More generally, it is clear that the neocortex and its predictive modelling capacities do not float free of the organism's idiosyncratic practical interests and needs, which themselves relate in crucial ways to subcortical structures and internal physiological states. This pragmatic character of mental representation will loom large in later chapters and acts as an important bulwark against an implausible vision of the neocortex as operating independently of an embodied organism's contingent desires, emotions, morphology, and time-pressured practical engagements with its environment.

Nevertheless, these complications should not be taken to undermine the general job description of cortical functioning outlined in this section. For organisms that must coordinate their actions with a complex dynamic environment at multiple spatiotemporal scales, an informationally efficient predictive modelling engine capable of structuring itself in such a way as to reflect the complex bodily and environmental causes of the organism's sensory signals is a highly effective organ of adaptive success.

(6.)**5. Conclusion**

In this chapter I have outlined a way of understanding information processing in the neocortex in terms of generative model-based predictive processing. In the next chapter I will show how this approach to understanding cortical function both vindicates and deepens the three ideas concerning mental representation that I extracted from Kenneth Craik's work in Chapter 3.

[18] See Miller and Clark (2017).

7. Representation, Prediction, Regulation

(7.)1. Introduction

In this chapter I show how the formal ideas and empirical research outlined in the previous chapter map onto the account of mental representation that I have developed throughout the thesis. In Chapter 3 I outlined a proto-theory of mental representation that I attributed to Craik that included three core components: first, an account of mental representation in terms of idealised models that capitalize on structural similarity to their targets; second, an appreciation of prediction as the core function of such models; and third, a regulatory understanding of brain function. In Sections 2, 3, and 4 of this chapter I show how predictive processing vindicates and extends each of these ideas and conforms to the broader characterisation of model-based representation outlined in Chapters 4 and 5.

(7.)2. Generative Models as Structural Representations

> "A key conclusion—that follows from the Bayesian brain—is that the structure of a good brain will recapitulate the (statistical) structure of how sensations are caused; in short, the model resides in the structure of the brain" (Friston & Buzsáki 2016, p.501).

One of the core ideas that I attributed to Kenneth Craik in Chapter 3 is that mental representations should be understood as idealised models or elements of such models that share a relational structure with target domains. That is, mental representations should be understood as *structural representations*: representations that capitalize on a structure-preserving mapping from elements of the representation to elements of the target domain in the manner outlined in Chapter 5.

As we saw in the previous chapter, the core representational structure within predictive processing is the hierarchical probabilistic generative model (Clark 2016; Kiefer & Hohwy 2017). The idea that such models are representations that recapitulate the causal-statistical structure of target domains is widespread in the scientific literature. Seth and Friston (2016, p.3), for example, write that "neuroanatomy and neurophysiology [in the predictive brain] can be regarded as a distillation of statistical or causal structure in the environment disclosed by sensory samples," Kiebel et al (2009, p.7) claim that the brain's perceptual system "inherits the dynamics of the environment and can predict its sensory products accurately," Bubic et al. (2010, p.8) note that "the prediction of future states of the body or the environment arises from

mimicking their respective dynamics through the use of internal models," and Friston (2002, pp.237-8) writes that

> "the hierarchical structure of the real world literally comes to be "reflected" by the hierarchical architectures trying to minimize prediction error, not just as the level of sensory input but at all levels of the hierarchy."

Importantly, it may be that certain structural features of the world are embedded implicitly in our neural architecture. For example, Friston and Buzsáki (2016) suggest that the partial separation of the visual cortex into ventral and dorsal streams (the so-called "what" and "where" streams) reflects an important feature of the environment's statistical structure: the fact that an object's identity is independent of its location. (Learning *where* an object is does not predict *what* it is). More generally, it may that certain purely "anatomical constraints will sometimes bias the processing of information in a way that can be *interpreted* in terms of prior constraints" (Kersten et al. 2004, p.296, my emphasis). These architectural features might in some important sense reflect structural features of the world—and thus in some sense implicitly "model" the world—without explicitly representing those features.

When we turn to explicitly articulated generative models in the brain of the kind described in the previous chapter, however, in what sense are they structural representations? That is, what are the elements and relations implicated in the relevant structures?

In short, they are the random variables that stand in for features of the body and environment and the causal and statistical relationships between them. Generative models are a species of statistical model. Such models capture the dependencies that obtain between the values of a set of random variables and are thus accurate to the extent that the structure of dependencies they capture maps onto the pattern of dependencies among the relevant features of their target domain (Kiefer & Hohwy in press; see also Cummins 1989; 2004).

There is a sense in which most statistical models are generative models: given the values of some variables along with model parameters, the model will typically automatically generate or "fill in" the values of others (Danks 2014). The generative models within predictive processing, however, are a specific kind of generative model: models of the causal process by which sensory inputs across the different modalities are generated by interactions among hidden causes in the world. They are thus accurate to the extent that the structure of the generative model maps on to this generative process. As Kiefer and Hohwy (in press, p.3) put

it, "the degree to which a generative process is capable of modelling another generative process depends on the similarity between the two processes."

In the hierarchical architecture common to predictive processing and related theories of the neocortex, the relevant relations here exist both between variables represented at a single level and between variables at different levels. This illustrates the holistic character of such models: it is the rich networks of dependencies among variables from which they acquire their representational significance. As Friston (2002, p.6) puts it, "an essential aspect of causes is their relationship to each other (e.g. "is part of") and, in particular, their hierarchical structure." It is an interesting question just *how* holistic such representational systems are. For example, it is plausible to think that a random variable in a statistical model acquires its significance from the entire model of which it is a part. If this is right, however, the holism of the brain's representational capacities will be dependent on the question of how to individuate models. I return to this question below and in more depth in the final chapter.

Importantly, the gross network structure of a generative model captures what might be thought of as the *topology* of the model: the causal process by which interactions among hidden variables generate sensory evidence. As I noted in the previous chapter, one of the advantages of graphical models is that they make this topological structure explicit. For this reason, directed graphical models such as Bayesian networks are often referred to as "causal maps" (Gopnik et al. 2004).

To summarise, then, accurate generative models implement a process for generating a set of phenomena that recapitulates the real-world process by which those phenomena are in fact generated. In this sense there is a structure-preserving mapping between the two: the patterns of causal and statistical relationships among elements of the real generative process are mirrored in the structure of the model—when it is accurate, at least.

There are at least three ways in which the story here is more complicated than this makes it sound, however.

(7.)**2.1. Clarifications and Complications**

First, as we saw in the overview of graphical models, one must distinguish between the qualitative structure of dependencies they capture—the presence and directionality of edges between variables and the values that they can assume—and the quantitative structure that

captures what can abstractly be thought of as the "strength" of these dependencies. Comparisons of similarity must therefore capture both dimensions (Kiefer & Hohwy 2017).

Second, the patterns of dependencies captured by a generative model are combined with prior probabilities defined over the values of latent variables. That is, in addition to capturing the relationships between features of a domain and the patterns of evidence they generate (often called the "generative distribution"), a generative model also typically encodes a prior distribution over these hidden features (Friston 2005). For this reason, Gładziejewski (2015, p.572) suggests that an important aspect of the structural correspondence between generative models and target domains is that "the model's hidden variables should be distinguished from one another in a way that maps onto distinctions between prior probabilities that characterize worldly causes." In this way the "generative model should be sensitive to the probability of the system's stumbling upon a given type of object, *regardless* of the current context or current incoming sensory signal" (Gładziejewski 2015, p.572).

This is what proponents of Bayesian perceptual psychology typically have in mind when they talk of prior expectations reflecting statistical regularities in the world. For example, in addition to capturing the process by which the position of the light source interacts with other features of a scene to generate our visual input, a generative model might also encode a strong prior expectation that the light source will come *from above* (Rescorla 2013). This prior reflects a genuine feature of our environment—namely, the fact that light *does* typically come from above—and systematically biases perceptual processing. As such, when the prior is violated under ecologically atypical conditions, Bayesian perceptual systems that operate like this will systematically generate perceptual illusions (Rescorla 2013). For example, artificially lighting a concave object from below can generate the illusory inference that it is convex. The reason is that the pattern of light reflected from the object *would* signify convexity *if* the light source came from above. As such, a strong prior expectation that it *does* come from above interacts with otherwise reliable assumptions about the data-generating process to yield an inaccurate inference.

One issue that arises at this point is this: as just noted, it is widely claimed in the philosophical (Gładziejewski 2015; Hohwy 2013; Kiefer & Hohwy 2017) and psychological (Friston 2005; Kersten et al. 2003) literature that prior expectations such as these reflect regularities in the environment. As Feldman (2013) points out, however, this interpretation is difficult to reconcile with Bayesian views in the philosophy of probability, where probabilities are claimed

to reflect an agent's *uncertainty* about states of the world, not features of the world itself. For example, it is plausible to think that there is no objective chance concerning *where the light source comes from*. Instead, (setting quantum phenomena aside) the light source—like all entities—is always in a determinate position, and the introduction of probabilities merely reflects uncertainty about this position. On this reading it is thus simply a mistake to talk of prior expectations reflecting statistical regularities in the world (Feldman 2013).

I do not want to take a stand on this issue in the philosophy and metaphysics of probability here. As I noted in the previous chapter, there are in fact several distinct reasons why probabilities might be included within a model, including noisy observations, underdetermination of hidden causes by their effects, the partiality of models, and—sometimes, perhaps—the objective stochasticity of processes in the world at the level at which we model them. As such, it is plausible that a pluralistic perspective is desirable here: sometimes probabilities should be understood to reflect uncertainty, and sometimes they be should understood to reflect genuine statistical frequencies in the world. Whatever stance one takes on this issue, however, Feldman's argument is in any case very limited. The probabilities included in generative models are defined over a structure of dependence relationships that are supposed to map onto the actual causal and correlational relationships that obtain among the real features of a target domain. As such, generative models can be thought of as structural representations even if one adopts a wholly "subjectivist" interpretation of probability, because there is far more to such models than just their probabilistic character.

The third, final, and related clarification is this: I have so far been describing the structural correspondence between the model and its target domain in a manner analogous to how one might compare the pattern of spatial relations among elements on a map of London with the pattern of spatial relationships that obtain among the corresponding places in London. In addition, however, one can also exploit such representations to represent the changing state of the relevant target—for example, to index one's current location on the map as one moves around, or to update the values of variables in a generative model so that they accurately reflect the *current* state of the generative process. Under these circumstances, the model is accurate to the extent that the distribution of variable-values reflects the actual state of the domain. Crucially, in architectures like predictive coding the priors from which predictions are generated change dynamically over short time-scales to concentrate probabilities over specific values of model-variables in the service of indexing the current state of the world, which amounts to forming different hypotheses about the state of the world.

As we will see in Chapter 10, this form of representation—indexing the current state of the target domain—is *mediated* through the first form of structural correspondence and relates the model-based understanding of mental representation outlined here to the enormous body of philosophical work on mental representation focusing on concepts such as covariation and indication.

(7.)**2.2. The Targets of Generative Models**

I also argued in Chapter 5 that one must be able to distinguish between a model and its target. A model's target cannot just be whatever it happens to structurally resemble: this structure will fortuitously resemble a large set of structures in the world that have nothing to do with the representational properties of the model. Further, if a model's target is determined by what it structurally resembles, misrepresentation is obviously impossible (Cummins 1996).

I argued that one can get around this problem by specifying the target of mental models by appeal to both the *function* of the mechanism from which they are produced and the aetiology of specific models. On the assumption that the function of the mind's modelling engine is to recapitulate the causal-statistical structure of the environment, this structure is its generic target. The target of specific models (or regions of a broader model) can then be understood as that aspect of this structure causally implicated in their production.

Importantly, this basic story can be transferred straightforwardly to predictive processing. Specifically, insofar as the neocortex functions as a general-purpose learning device for constructing generative models of the process by which interactions among features of the body and world generate the organism's sensory inputs, its generic target can be understood as the causal-statistical structure of that process. For any *given* generative model, then, its target can be understood as the data-generating process from which it learned. If the model is accurate, it will thus acquire elements which map onto the functionally significant elements of that process (see Chapter 10).

Although I think that this story is basically correct, we will see below some reasons for wanting to augment it slightly (see Section 4).

(7.)**2.3. Qualifications**

For the foregoing reasons, I think that the status of generative models as genuine structural representations is secure—that common descriptions of such models as recapitulating the

causal and statistical structure of the world in the broader literature are legitimate. Nevertheless, before continuing it is worth nothing three things.

First, there is the crucial issue of idealisation. Hohwy (2013) characterises the predictive mind as a "mirror of nature," but we saw in Chapter 4 that models are typically highly selective and often distortive, and we can expect this to be the case with respect to the mind's generative models as well. I will return to this issue below and then at greater length in Chapter 9.

Second, I have described the model's structure here in a way that abstracts away from its actual implementation in the brain. This should not obscure the fact that it is ultimately states and relations among states of the brain that realise the relevant models on the view that I am defending here. That is, the causal and statistical dependencies among variables of the model, the changing values assigned to variables, and the probability density functions (encoded in terms of sufficient statistics in predictive coding) must ultimately be reflected in patterns of neural activity. Nevertheless, for my purposes here I will abstract away from these implementational details. Systems made of very different materials could implement predictive modelling engines of the sort described here. What is important are the functional relations between elements of the representational system, not their contingent implementation.

Finally, I noted above that a crucial question in the current context is *how many* generative models inhabit a predictive mind. In the broader literature it is often either explicitly claimed (e.g. Clark 2013, p.203) or simply assumed that there is just one highly complex interconnected generative model of the process by which sensory signals across the different modalities are generated, with specific regions of that model targeting specific features of that generative process at different spatiotemporal scales. I will mostly share this assumption throughout this and the next chapter and thus continue to talk of "the" hierarchical generative model. In the final chapter, however, I will raise some reasons for being uncomfortable with this talk.

(7.)**3. Prediction in the Predictive Mind**

In Chapter 3 I introduced Craik's view that the core function of the brain's models is *prediction*. Predictive processing, of course, shares this emphasis on the functional importance of prediction to intelligence. Importantly, however, it also ascribes functions to prediction that Craik himself did not identify. It is worth spelling out these functions and showing how predictive processing both vindicates and extends Craik's emphasis on the adaptive importance of prediction.

(7.)**3.1. Prediction in Learning and Perception**

According to predictive processing, the function of the brain is to *minimize prediction error*, the mismatch between externally generated sensory inputs and internally generated predictions of such sensory inputs. Importantly, "prediction" here refers to a sense of prediction that in Chapter 4 I called "synchronic prediction": predicting the "present"—in this case, the evolving sensory inputs generated by the world.[19] Of course, for this to work the brain must also exploit predictions of the evolving bodily and environmental states responsible for generating proximal sensory inputs (Friston 2005). I will return to this capacity for more forward-looking predictions in perception in the next chapter.

Craik did not foresee this role of prediction in registering the *current* state of the body and environment. Indeed, in work published posthumously he drew an explicit distinction between prediction and perception: "but this is not sense-perception; it is anticipation" (Craik 1966, p.63). With predictive processing, of course, brains identify the evolving state of the body and environment in perception by minimizing the error in their predictions of the sensory input it generates. As it is often put, they identify the multilevel hypothesis that best *explains* the current sensory signals received across the organism's various sensory transducers by minimizing the error in their predictions of such signals (Hohwy 2013).

Further, when Craik (1943, p.115) came to the question of how brains might acquire the models that facilitate this predictive capacity, he could only affirm that "it is *possible* that a brain consisting of randomly connected impression synapses would assume the required degree of orderliness [for model-based prediction] as a result of experience" (my emphasis). Of course, Craik's assumption that the brain's synaptic connections are completely *randomly* connected at birth is surely incorrect (Marcus 2004). More importantly, however, he did not appreciate the central role that prediction itself can play in the installation of the very bodies of knowledge that facilitate prediction. As we saw in the previous chapter, this trick lies at the centre of predictive processing: brains acquire the knowledge that facilitates the minimization of prediction error by minimizing prediction error.

(7.)**3.2. Prediction in Action**

Craik evidently did see the centrality of prediction to *action*. Of course, in one sense this emphasis on prediction in action is hardly surprising: without some capacity to anticipate the

[19] In fact, in some applications of predictive coding in machine learning, what gets predicted is explicitly the *next* input (for example, the next frame in a video sequence) (e.g. Lotter et al. 2017).

likely outcomes of our actions, we can have no reason to act one way rather than another. This is the sense in which "voluntary behavior in general is initiated and controlled by a representation of its expected outcomes, illustrating how anticipation lies in the foundations of goal-directed behavior" (Bubic et al. 2010, p.6). Nevertheless, there are purely model-free reinforcement learning algorithms that enable systems to adapt their behaviour to their environmental circumstances as a function of whichever actions have been rewarded or punished *in the past*. We saw in Chapter 3 that Craik believed such learning would be insufficient to account for true intelligence. For Craik (1943), thought is "constructive" (p.51), facilitating a predictive competence that releases us from the constraints of previous schedules of reinforcement. We will see below how predictive processing supports this conviction. In addition, however, it also shares with prominent views in contemporary sensorimotor psychology an emphasis on the role of prediction in basic sensorimotor control that—as we saw in Chapter 3—Craik also identified: namely, its role in overcoming signalling delays.

According to highly influential Bayesian models in sensorimotor psychology (see Rescorla 2016), fluid goal-directed movement is facilitated by two models: first, an *inverse model* that maps desired outcomes determined by an agent's behavioural goals onto motor commands that will bring them about; second, a *forward model* that maps copies of motor commands ("efference copies") onto a prediction of their likely sensory consequences ("corollary discharge"). Such forward models could in principle function as look-up tables but in fact seem to work in the manner of generative models outlined here—namely, as dynamic models with variables and covariational dependencies that map onto the relevant system the forward model predicts the behaviour of (e.g. the musculoskeletal system) (see, e.g., Grush 2004).

Forward models of this kind serve several important functions.

First, they enable the brain to overcome various signalling delays that would otherwise impede fluid sensorimotor control (Franklin & Wolpert 2011). As noted, this was one of the predictive functions that led Craik to posit internal models: by yielding rapid predictions of the likely sensory outcomes of actions, the brain can adjust its behaviour appropriately in response to such outcomes before the relevant sensory inputs conveying the information arrive.

Second, by generating fine-grained predictions of the outcomes of the organism's actions, the forward model can be exploited for disambiguating self-generated sensory inputs from externally generated ones (Blakemore et al. 2000).

Third, in online sensorimotor control, the predictions of the forward model can be compared with sensory feedback to compute an optimal estimate of the state of the relevant system being modelled (the "plant" in control-theoretic terms). In most accounts, the forward model employs a *Kalman filter* that effectively weights the prediction and sensory feedback by their uncertainty to compute this estimate. Further, the prediction errors generated by the mismatch between these two signals can be used to install and update the structure of the forward model in the manner outlined in the previous chapter (see Grush 2004; Rescorla 2016). In other words, all the central characteristics of predictive processing are found within fairly standard accounts of sensorimotor control: generative model-based prediction, prediction error-based learning, and precision-weighting.

Finally, as Grush (1997; 2004) points out, with a predictive model of this kind in place, the brain can exploit the model "offline" to generate counterfactual predictive simulations of what *would* happen as a result of merely possible actions, thereby revealing how a basic imperative towards fluid sensorimotor control might have fortuitously paved the way for intuitively "higher" forms of cognition decoupled from online sensorimotor processing (see below).

Importantly, predictive processing's account of goal-driven action shares many core features of this influential forward model-based account (Adams et al. 2012; Friston et al. 2010; see Clark 2016, ch.4 for a review). The chief difference, however, is that rather than treating the forward model as a mechanism that is distinct from the core mechanism of motor control, it claims that motor commands are in effect *implemented* by predictions of the sensory feedback of desirable actions. As such, there is no need for a separate inverse model and efference copies.

Although I cannot do justice to this approach to motor control here, the core idea is remarkably simple. It is that action can be understood in terms of "the same type of objective function as predictive coding, i.e., the minimisation of prediction errors… about sensory inputs" (Stephan et al. 2017, p.3). Specifically, whereas perception is a matter of updating internally generated predictions to match externally generated sensory signals, action can be understood as a matter of intervening upon the world to bring sensory inputs into alignment with internally generated predictions. The important predictions here are *proprioceptive predictions*, where "proprioception" names the sense that informs the brain concerning the relative locations of bodily parts and the forces being applied to them (Adams et al. 2012). By specifying desired outcomes in terms of their proprioceptive consequences, acting on the world to bring sensory

inputs into alignment with those predicted consequences provides an effective mechanism for bringing those outcomes about.

This is a radical proposal (for arguments and evidence in its defence, see Clark 2016, ch.4; for an important critique, see Rescorla 2016) and I will not attempt to adjudicate between these two approaches to sensorimotor control here. For my purposes, what is important is what they share in common: an emphasis on the centrality of predictive modelling to core mechanisms of online sensorimotor control.

(7.)**3.3. Beyond Perception and Action**

So far, the emphasis has been mostly on perception and action, albeit with the former construed broadly to include the brain's efforts to identify those features of the body and environment responsible for generating the evolving streams of proximal stimulation received by the organism, which includes the use of multimodal areas of cortex identifying common causes of correlated patterns of input across the different modalities (Friston 2005). Nevertheless, does predictive processing have anything to say about intuitively "higher" cognitive capacities?

As noted in the previous chapter, generative model-based predictive processing is useful whenever there is some systematic process by which causes give rise to effects that an agent has access to (Tenenbaum et al. 2011). Crucially, this is applicable across a wide variety of cognitive domains and problems (Lake et al. 2016; Tenenbaum et al. 2011). As such, when an organism has the capacity to model the distal causes of its proximal sensory inputs, it can exploit its access to the former to model interactions that occur purely among the distal causes themselves. One important example of this concerns the way in which novel states of the *physical world* are generated by interactions among features of the physical world. In interacting with our environments, for example, we have to generate fluid predictions about the likely effects of all kinds of possible physical interventions upon or alterations to the world. As Ullman et al. (2017, p.1) put it, "we implicitly but continually reason about the stability, strength, friction, and weight of objects around us, to predict how things might move, sag, push, and tumble as we act on them." Importantly, "this physical scene understanding links perception with higher cognition" (Battaglia et al. 2013, p.18,327).

To account for this competence, Battaglia et al. (2013; see Lake et al. 2016; Ullman et al. 2017) modelled this capacity for "intuitive physics" in terms of a generative "physics engine" similar to those used in the design of interactive video games, which can be used for running predictive simulations of the outcomes of different alterations to a scene. After providing an explicit nod

to Craik, who "first articulated… that the brain builds mental models that support inference by mental simulations analogous to how engineers use simulations for prediction and manipulation of complex physical systems," Battaglia et al. (2013, p.18,327) show how a predictive model of the environment's physical dynamics can be exploited for approximate probabilistic simulations of the likely outcomes of a large range of possible alterations to the environment. In this way we have the capacity to run quick predictive simulations to predict such things as "whether a stack of dishes will topple, a branch will support a child's weight, a grocery bag is poorly packed and liable to tear or crush its contents, or a tool is firmly attached to a table or free to be lifted" (Battaglia et al. 2013, p.18,327). Impressively, they show that the predictions of such a physics engine closely align with the predictions made by human subjects (Battaglia et al. 2013).

(7.)**3.4. Offline Cognition**

This physics engine supports rapid physical predictions via predictive simulations. In addition, however, generative models also support—in principle, at least—forms of "offline cognition" more radically decoupled from online sensorimotor processing. As many have argued, this capacity for offline simulation seems to pave an attractive path from online sensorimotor control to more sophisticated cognitive capacities (Clark 2016; Colder 2011; Grush 2004). As Clark (2016, p.84) puts it, generative model-based perception has a

> "life-enhancing spin-off… For such perceivers are thereby imaginers too: they are creatures poised to explore and experience their worlds not just by perception and gross physical action but also by means of imagery, dreams, and (in some cases) deliberate mental simulations."

With such models we can "move the world into the head, and then… run models to observe possible implications for the actual world" (Moulton & Kosslyn 2009, p.1278). In this sense, generative models give concrete expression to Craik's (1943, p.61) speculation that "small-scale models" of the world and our "own possible actions" enable us to "try out various alternatives" and "conclude which is the best of them." A generative model-based understanding of "offline" cognition thus aligns well with Moulton and Kosslyn's (2009, p.1273) contention that

> "the primary function of mental imagery is to allow us to generate specific predictions based upon past experience. All imagery allows us to answer "what if" questions by making explicit and accessible the likely consequences of being in a specific situation or performing a specific action."

As Mumford (1992, p.250) puts it,

> "In the course of reflecting about some problem, we can block out the actual stimulus being received by our senses, or we can close our eyes. At that point, all the neural machinery for sensory processing is available for thought" (Mumford 1992, p.250).

One interesting application of this idea comes from recent work in machine learning by Ha and Schmidhuber, who designed a generative artificial neural network-based agent that can learn an action policy "entirely inside of its own hallucinated dream generated by its world model, and transfer this policy back into the actual environment" (Ha & Schmidhuber 2018, p.1).

Taking inspiration from their speculation that "the predictive model inside of our brains…might not be about just predicting the future in general, but predicting future sensory data given our current motor actions," Ha and Schmidhuber (2018, pp.1-2) designed an artificial agent that combines a predictive model learned in an unsupervised way from input data (two-dimensional image frames of video sequences) with a "controller model" that learns to perform tasks "using features extracted from the world model." In this way the agent learns action policies from a compressed spatial and temporal representation of the problem domain rather than directly from the input itself. In this way prediction-driven unsupervised learning is used "to obtain a predictive world model, M, that a controller, C, may use to achieve its goals more efficiently, e.g., through cheap, "mental"…trials, as opposed to expensive trials in the real world" (Schmidhuber 2015, p.5).

After demonstrating that this strategy can be used to enable the artificial agent to successfully learn to play a car racing video game, they note that "since our world model is able to model the future, we are also able to have it come up with hypothetical car racing scenarios on its own" (Ha & Schmidhuber 2018, p.6). Impressively, they show that the artificial agent can "*learn inside of its own dream*, and transfer this policy back to the actual environment" (Ha & Schmidhuber 2018, p.6, my emphasis).[20] Specifically, after "learning only from raw image data collected from random episodes, it learns how to simulate the essential aspects of the game" in a form that is wholly offline (i.e. detached from any externally provided data) (Ha & Schmidhuber 2018, p.8). Crucially, it can then exploit such offline predictive simulations—what the authors call its "dream world"—to learn action policies, *which can be successfully transferred back to the "actual" environment.*

[20] They demonstrate this not in a car racing game but in a version of the game "Doom," in which the aim is avoid fireballs shot by enemies from another side of a room (Ha & Schmidhuber 2018).

Of course, this is a rather radical form of learning via offline simulation. Nevertheless, it provides an impressive computational demonstration of a more general point: the installation of a predictive model acquired via unsupervised learning in online interactions with a domain can be exploited for highly adaptive but purely "offline" predictive simulations, the fruits of which can then be exploited in real-world action.

(7.)**3.5. Summary**

An understanding of mental representation in terms of generative model-based predictive processing vindicates the adaptive significance that Craik attributed to prediction. Predictive processing situates prediction at the centre of unsupervised learning, perception, and sensorimotor control, and it suggests an attractive path to more sophisticated forms of cognition through the capacity of generative models to be used "offline." Of course, the exploration of this latter domain of higher cognition here has been skeletal, ignoring—given space constraints—many applications of generative model-based predictive processing in the literature to this domain, including influential work on social cognition (e.g. Friston et al. 2011). Further, we will see in the final chapter some reasons for scepticism about whether theories such as predictive processing are able to capture some of the more rarefied aspects of distinctively human thought. Nevertheless, this brief overview should serve to illustrate how useful a capacity for generative model-based prediction can be across different psychological domains. In the next chapter I will argue that prominent nonrepresentational approaches to perception, action, and cognition neglect this importance of prediction.

(7.)**4. The Cybernetic Predictive Mind**

> "…the living brain, so far as it is to be successful and efficient as a regulator for survival, must proceed, in learning, by the formation of a model (or models) of its environment" (Conant and Ashby 1970, p.89).

The third insight that I extracted from Craik's work in Chapter 3 was a commitment to a *regulatory* understanding of brain function. This regulatory perspective reflects Craik's position as an important pre-cursor to the intellectual tradition of cybernetics (see Boden 2006; Gregory 1983). As noted in Chapter 3, cybernetics was focused on explaining seemingly adaptive or teleological (goal-directed) behaviour in terms of regulatory mechanisms or—in the case of organisms—self-regulating mechanisms that can maintain their internal states and structural integrity in the face of external perturbations (Pickering 2010). Adaptive systems are

thus defined by a set of essential variables that must be kept within certain bounds. As Ashby (1952, p.64) puts it,

> ""Adaptive" behaviour is equivalent to the behaviour of a stable system, the region of the stability being the region of the phase-space in which all essential variables lie within their normal limits."

Insofar as a system is configured in such a way as to reliably trigger actions that reduce the difference between its current state and its "goal" state as defined by these "normal limits," it will thus appear to pursue goals. As Craik (1966, p.12) argues, systems which regulate their internal state in this way "show behaviour which is determined not just by the external disturbance acting on them and their internal store of energy, but by the relation between their disturbed state and some assigned state of equilibrium." By exchanging matter and energy with their environments to restore this equilibrium, living organisms "show a degree of spontaneity, variability, and purposiveness unknown in the non-living world. This," Craik (1966, pp.12-13) claims, "is what is meant by "adaptation to environment.""

Summarising this focus, Seth (2015, p.8) writes that

> "[the] fundamental cybernetic principle is to say that adaptive systems ensure their continued existence by successfully responding to environmental perturbations so as to maintain their internal organization."

At least as it was first conceived, then, cybernetics fundamentally sought to understand—and construct—regulatory systems. At a minimum, such systems involve processes of negative feedback whereby information concerning deviations from acceptable values of essential variables is exploited to adjust behaviour to restore equilibrium (Rosenblueth et al. 1943).

Importantly, this focus on regulatory systems distinguished cybernetics from the understanding of "cognitive systems" that emerged from the cognitive revolution of the late 1950s. For the most part, the interdisciplinary field of cognitive science that emerged at around that time modelled intelligent systems as computational systems that were highly *reactive*—that were designed to reliably execute a set of instructions that yield a desired output in response to specific inputs (see Bechtel 2008; Cisek 1999). By contrast, self-regulating mechanisms are configured to actively *bring about* certain inputs: those that signify the stability of their internal states (Cisek 1999). In this sense such systems are fundamentally *active*, not *reactive*, and thus provide a much better model for understanding biological systems than, say, a laptop. As Bechtel (2008, p.210) puts it,

> "Insofar as some of its mechanisms are designed to maintain a constant internal environment despite changes in the external environment, a living system can appear as an active system doing things that resist its own demise."

As the cyberneticists saw, then, the focus on self-regulating systems bears a direct relevance to neuroscience and the origins of adaptive success in biological agents (Ashby 1952; 1956; Rosenblueth et al. 1943). All living systems must maintain themselves far from thermodynamic equilibrium, resisting the tendency towards disorder described by the second law of thermodynamics by exchanging matter and energy with their environments to maintain their internal organisation (Anderson 2014; Bechtel 2008; Cisek 1999). In the late 1900s, Bernard (1878, p.121) had speculated that "all the vital mechanisms, however varied they may be, have only one object, that of preserving constant the conditions of life in the internal environment." The physiologist Walter Cannon (1928) coined the term "homeostasis" (from the Greek words for "same" and "state") to describe this tendency of all living things to maintain the stability of their internal conditions in the face of external disturbances (see Bechtel 2008, p.210). This stable state varies between organisms and is dependent on a large number of interacting variables, such as body temperature and blood sugar level, which collectively constitute optimal functioning for the organism.

As Craik (1966, pp.13-4) put it,

> "the device of using energy derived from outside to restore equilibrium is, I think, one of the first and most important features of living organisms to appear… [It] is so important that it opens, to living organisms, an immense realm of possible reactions which, if it were not for the external supply of energy, would be thermodynamically "uphill" and therefore unable to occur."

In this sense biological systems are a prime example of the teleological mechanisms that were the focus of cybernetics. As the cyberneticists saw, this perspective on adaptive success suggests a fundamental job description for the brain: "to regulate the organism's internal milieu" (Sterling & Laughlin 2015, p.xvi; see Ashby 1952). Further, they recognised and embraced the potentially deeply counter-intuitive picture of the mind that this account of brain function implies, in which even "cognitive processes are grounded in fundamental evolutionary imperatives to maintain physiological homeostasis" (Seth & Friston 2016, p.2) and mental mechanisms are understood as "further accoutrements that enable… systems to better maintain themselves" (Bechtel 2008, p.238).

This control-theoretic or cybernetic understanding of the brain has become increasingly influential in recent years (L.F. Barrett 2017a; Sterling & Laughlin 2015; see Williams and Colling 2017 for an overview). It constitutes a theoretical approach to understanding the brain that might be called "bottom-up functionalism" (Williams and Colling 2017). Rather than starting with ideas or prejudices (often derived from folk psychology) about the tasks or functions performed by putatively distinct cognitive systems, or seeking to understand cognition primarily based on low-level mechanistic details of neural circuitry, bottom-up functionalism derives a general job description for brains from theoretical considerations drawn from statistical physics and control theory, and then seeks to understand psychological capacities in terms of their contribution to this core regulatory task (L.F. Barrett 2017a; Bechtel 2008; Pezzulo & Cisek 2016).

(7.)**4.1. Regulation and Modelling**

In Chapter 3 I argued that this regulatory understanding of brain function provides a fundamentally pragmatic context within which to understand the predictive models that Craik posits. Nevertheless, so far I have not said anything to connect this regulatory understanding of brain function to the concept of predictive modelling.

In fact, as I noted in Chapter 3, most early work in cybernetics did not focus on explicit representation or modelling (see Pickering 2010). One important exception to this, however, was Ashby's (1952; 1956) work. In 1970 Ashby put forward the famous "good regulator theorem" with Roger Conant (Conant & Ashby 1970), developing his earlier views on the "law of requisite variety" in control theory (see Seth 2015). The good regulator theorem attempts to formally demonstrate that "any good regulator of a system must be a *model* of that system" (Conant & Ashby 1970, my emphasis). Specifically, it claims that an optimal regulator must exploit a structure-preserving mapping between elements of its control system and the system being regulated. To get an intuitive feel for this claim without wading into the specific details of the proof, consider a maximally simple example of a regulatory system such as a typical thermostat tasked with regulating the temperature in a room. Such systems exploit a structural correspondence between the height of mercury in a thermometer and the ambient temperature, such that variations in the latter are mirrored by variations in the former (O'Brien 2015). In this way the level of mercury in the thermometer can be described as "modelling" the room's temperature.

Importantly, this theoretical connection between regulation and modelling sits at the centre of the "free energy principle" as described by Friston, which provides the theoretical foundation for the most ambitious versions of predictive processing (see Friston 2010). I have neither the technical competence nor space to attempt a proper overview of the free energy principle here, and we will see shortly that it is in any case slightly limited from the perspective that I am defending. Nevertheless, at its core is a relatively simply idea, which it is worth briefly spelling out.

The free energy principle starts with the observation outlined above: namely, that "the defining characteristic of biological systems is that they maintain their states and form in the face of a constantly changing environment," thereby somehow violating "the fluctuation theorem, which generalizes the second law of thermodynamics" (Friston 2010, p.1). One way of describing this process of homeostasis is this: of all the possible states that a living system *could* be in, it must remain within a very narrowly circumscribed subset of them. If one defines a probability distribution over this set of possible states, then, the probability that it will be in a tiny fraction of these states is very high relative to other possible states *insofar as it is alive*. As such, deviations from homeostasis can be described in terms of occupying increasingly improbable—"surprising"—states, where "surprise" (technically "surprisal") here roughly describes the improbability of an outcome relative to a probability distribution (Friston 2010; Hohwy 2013; Seth 2015).

Friston's (2010) bold proposal is that a task that a biological system can perform that approximates the minimization of "surprise" in this sense is the minimization of "variational free energy," an information-theoretic quantity that—under some simplifying assumptions—translates to *long-term prediction error*. In this broader theoretical context, then, the fundamental problem that prediction error minimization solves is homeostatic regulation, such that the former "is, essentially, a tool for self-organisation" (Gładziejewski 2015, p.563).

(7.)**4.2. Regulation and Predictive Modelling**

Both the good regulator theorem and the free energy principle provide an important theoretical link between regulation and modelling. Nevertheless, there is also a sense in which these ideas are quite limited from the perspective that I am defending here. Specifically, both apply to *all* forms of regulation—from thermostats to single-celled organisms to primates like us—and the sense of "modelling" they point to is in many cases quite different from the explicit predictive modelling described in previous chapters. A typical thermometer, for example, is a highly

reactive system that is deeply coupled to ambient temperature. Beyond exploiting a structure-preserving mapping between the level of mercury and the "target" (the temperature of ambient air), then, it bears little similarity to the generative models outlined here, which are fundamentally *predictive* in character and capable of being used for purely "offline" purposes.

Despite this, both the good regulator theorem and the free energy principle point towards a fundamental and ultimately commonsense insight: namely, that self-regulation in biological systems requires that such systems exploit mechanisms that reliably track those aspects of the world *relevant* to their core regulatory function (Bechtel 2008). Whether or not we think it is appropriate to call these tracking mechanisms in simple biological systems "models," these theoretical ideas thus suggest that the explicit predictive modelling outlined in this and previous chapters is ultimately just a more complex elaboration of the core world-involving strategy demanded by successful homeostasis (L.F. Barrett 2017a; Seth 2015). That is, they suggest that the difference between the explicit predictive modelling outlined here and the "modelling" found in simpler control systems reflects the greater *difficulty* and *complexity* of the regulatory task, not a difference in the task itself.

There are various ways in which this core task of self-regulation might give rise to the predictive modelling outlined here. I will briefly mention five that are highly salient in the context of this thesis.

First, unlike thermometers, we have seen that complex adaptive systems like the brain must often surmount severe noise and uncertainty in sensory signals to both *learn* and then *identify* those features of the environment responsible for generating them. Under such conditions, as Seth (2015, p.9) puts it, "it will pay to engage explicit inferential processes in order to extract the most probable causes of sensory states, insofar as these causes threaten the homeostasis of essential variables."

Second, we saw above that sophisticated control often requires the construction of a forward or predictive model of the system being regulated (the "plant" in control-theoretic terms) (Grush 1997; 2004). Thermostats are an example of "closed-loop control systems" in which the controller is exclusively reliant on continual feedback about the state of the plant as the control process unfolds. This is the sense in which they are highly reactive. There are many engineering contexts in which this process is inadequate, however, and the control system must exploit not just a controller coupled to the behaviour of the system being regulated but "also a subsystem whose job is to mimic the operation of the plant" (Grush 2003, p.69). As we saw

above, such forward models can be crucial for overcoming feedback delays and allow the control system to optimally combine its prediction of the state of the plant with its sensory feedback in approximate conformity with Bayes' theorem (see Rescorla 2016).

Third, optimal regulation requires energetic efficiency (L.F. Barrett 2017a; Sterling & Laughlin 2015). As we saw in the previous chapter, predictive coding is especially attractive in this regard. Originally developed as a strategy for data compression, it ensures that the only aspect of an incoming signal a system needs to transmit is *unexpected* information: information not already accounted for in systemic response (Huang & Rao 2011). In this way predictive modelling can be exploited to make energetically efficient use of incoming information (L.F. Barrett 2017a).

Fourth, insofar as homeostasis is cast as a reactive process whereby systems are designed to *respond* to deviations from homeostatic set-points only once they have occurred, several theorists have pointed out that it does not seem to characterise the behaviour of complex regulatory biological systems. Specifically, such systems seem to be fundamentally *predictive*, anticipating needs *before* they arise and acting to avoid them. To describe this anticipatory character of self-regulation in complex biological systems, the term "allostasis" has been introduced into the literature (L.F. Barrett 2017a; Sterling & Laughlin 2015).

Finally, it is obvious that in creatures like us successful regulation of our internal state demands not just that we coordinate our behaviour effectively with our immediate environment at any given time, but that we engage in processes such as planning in which we coordinate our behaviour and choice with affairs distal in space and time—affairs that we are not in any immediate causal contact with (Clark & Toribio 1994; Corcoran & Hohwy 2018).

The upshot of such considerations is that the predictive modelling described in this thesis need not be understood independently of an organism's ultimate pragmatic concerns, but rather can be—as Craik claimed they *should* be—understood in terms of their "survival value," facilitating the complex ways in which we coordinate our behaviour with features of the world relevant to our survival.

Crucially, this theoretical context should fundamentally reconfigure how we think about the *target* of the mind's predictive modelling system. Specifically, it suggests that the ultimate function of this system is not simply veridical representation of the world's regularity structure in the service of prediction but rather that aspect of this structure *relevant* to the specific

organism's survival. As L.F. Barrett (2017a, p.16) puts it, it suggests that "the brain models the world from the perspective of its body's physiological needs," such that

> "the content of any internal model is species-specific…including only the parts of the animal's physical surroundings that its brain has judged relevant for growth, survival and reproduction… Everything else is an extravagance that puts energy regulation at risk."

In this way a cybernetic understanding of brain function suggests that representation of the external world "emerges as a *consequence* of a more fundamental imperative towards homeostasis and control" (Seth 2015, p.8).

(7.)**4.3. Summary**

To summarise, in this section I have sought to show how the predictive model-based account of mental representation outlined and defended so far can be situated within the context of a regulatory understanding of brain function championed by Craik and other cyberneticists. This cybernetic perspective provides a deeply pragmatic context within which to understand mental representation that should fundamentally reconfigure how we conceive of the target of the mind's models and the ultimate aim of predictive modelling outlined in this and previous chapters.

Before concluding, however, I should stress that the predictive model-based account of mental representation defended here is conceptually independent of this broader cybernetic understanding of brain function, and I am aware that more would need to be done to extract the explicit links between self-regulation and predictive modelling than the rather schematic remarks on the topic that I have offered here. Nevertheless, situating the account of mental representation defended here within this regulatory context has several attractions in the context of this thesis that make it worthwhile: it suggests a plausible pragmatic rationale for the representational capacities I have described, it honours Craik's broadly cybernetic approach to the mind, and—as we will see in future chapters—it allows me to show in more transparent terms how the account of mental representation that I have defended can either accommodate or avoid some of the most important challenges to representation-heavy accounts of the mind in recent cognitive science.

(7.)**5. Conclusion**

In this chapter I have argued that the three insights extracted from Kenneth Craik's work in Chapter 3 can be situated within an increasingly influential conception of the neocortex as a

prediction-driven general-purpose learning mechanism, reconfiguring its patterns of synaptic connections and internal dynamics to recapitulate the causal-statistical structure of the body and environment in a form that can be exploited for successful prediction.

With this account of mental representation in hand, in the next three chapters I will put it to work, favourably contrasting it with influential trends in both philosophy and psychology that seek to minimize the importance of mental representations in cognitive processes.

Chapter 8. The World Is Not Its Own Best Generative Model

> "Although suggestions regarding the importance of predictive mechanisms originate from very early phases of both psychology and neuroscience, until recently they have not been strongly advocated in the majority of frameworks of cognitive and neural processing" (Bubic et al. 2010, p.2).

(8.)**1. Introduction**

In this and the next two chapters I argue that the account of mental representation outlined and defended in previous chapters can answer and expose what is wrong with three challenges to representationalist accounts of the mind as advanced by a motley crew of anti-representationalists over many years: pragmatists, behaviourists, phenomenologists, reductionists, and those in the tradition of "4E" (embodied, embedded, extended, and enactive) cognition. Specifically, I consider the following three challenges: (1) that embodied interactions with the environment replace the need for mental representations in sensorimotor processing (this chapter); (2) that the pragmatic character of cognition undermines the prospects of any similarity-based account of the mind-world relation of the sort defended here (Chapter 9); and (3) that representational content is so metaphysically problematic that it should be eliminated from cognitive theorising (Chapter 10).

By addressing these challenges, I hope to showcase not just the attractions of the general theoretical account of mental representation that emerges from previous chapters relative to broadly anti-representationalist views in the philosophical and scientific literature, but also to favourably contrast this account with other prominent theories of the nature and functions of mental representations.

First, however, I should stress the following clarification: I do not claim that the three challenges that I address in these chapters exhaust the sources of anti-representationalism either in philosophy or psychology. The claim is rather that they capture what I think are the three most important and historically influential critiques of representationalist views of the mind within the context of cognitive science and philosophical naturalism. Some authors seem to derive anti-representationalist conclusions from epistemological or phenomenological considerations that are wholly divorced from the explanatory concerns of cognitive science, for example (e.g. McDowell 1996). I will not address such arguments here.

(8.)**2. The Replacement Hypothesis**

> "If the relevant features are reliably present and manifest to the system (via some signal) whenever the adjustments need be made, then they need not be represented" (Haugeland 1991, p.62).

The challenge to representationalist accounts of the mind that I will address in this chapter derives from what Shapiro (2010, p.4) calls the "replacement hypothesis," the hypothesis that "an organism's body in interaction with its environment replaces the need for representational processes thought to have been at the core of cognition." The replacement hypothesis comes from work in what (following Shapiro) I will call "embodied cognition," although my use of the term is very broad and intended to include research and ideas from ecological psychology, dynamical systems theory-approaches to cognition, enactivism, and certain work in robotics, as well as insights from American pragmatism and phenomenology that predate and provide the theoretical foundation for much contemporary work in embodied cognition.

Importantly, the replacement hypothesis is just one strand of embodied cognitive science, sometimes called "radical" embodied cognitive science (e.g. Chemero 2009). There is a large body of work within the tradition of embodied cognition that is not hostile to the explanatory importance of mental representations in cognitive processes (e.g. Clark 1996). As I will stress below and as has been argued elsewhere (e.g. Clark 2016; Grush 2003; Williams 2017), many of the insights from this work are consistent with a predictive model-based understanding of mental representation of the sort that I have defended here. As such, my target in what follows is not embodied cognition *as such*, but rather those who have drawn on ideas from embodied cognition to bolster and support the replacement hypothesis, which *is* inconsistent with the account of mental representation that I have defended in previous chapters.

The replacement hypothesis itself comes in stronger and weaker forms. In its most extreme form, it claims that embodied interactions with the environment replace the need for internal representations in *all* cognitive processes. On this view, the "only representations with any substantial or real implications for human cognition are to be found *outside the head*" (Malafouris 2013, 31, my emphasis). In a weaker form, the replacement hypothesis primarily targets "online" sensorimotor processing, where I intend "sensorimotor processing" to be understood as a vague term that applies broadly to capacities of "online" perception and intentional action (Anderson 2014; Chemero 2009). As will be clear, having a broad and ill-defined name of this kind for such phenomena is useful in describing a controversy that itself concerns the nature of the relevant capacities and the mechanisms underlying them. On this weaker manifestation of the replacement hypothesis, internal representations are not "central

and foundational to cognition and action" but rather "peripheral and emergent" (Anderson 2014, p.142) (see Brooks 1991; Chemero 2009; Van Gelder 1995). As Chemero (2009, p.115) puts it, "explaining how we write novels or plan vacations might require invoking something like a representation… But perception never does."

In what follows I will mostly focus on this latter form of the replacement hypothesis. It is more plausible, more influential, and its falsity would obviously imply the falsity of more radical nonrepresentational positions. It is important to stress, however, that sensorimotor processing here is not itself supposed to be understood as a peripheral aspect of intelligence. Chemero (2013, p.149), for example, claims that radical embodied cognitive science has been most successful in illuminating what he calls (after Brooks 1991) "the *bulkiest* parts of intelligence: perception, action, motor control, and coordination" (my emphasis). One of the central features of embodied cognition is its methodological turn away from intuitively "higher" cognitive functions that had preoccupied much of classical cognitive science—logical reasoning, parsing the syntactic structure of a sentence, playing chess, and so on—towards these more fundamental psychological capacities exhibited in "everyday coping" (Dreyfus 2002) or "basic cognition" (Hutto & Myin 2012). This methodological conviction reflects the plausible view that the primary evolved function of nervous systems is sensorimotor control, not abstract reasoning. As Anderson and Chemero (2016, p.16) put it, "the brain evolved to guide action, not write sonnets." Further, it also reflects the hope that by focusing on these more basic sensorimotor processes we might ultimately gain a deeper understanding of higher-level cognitive capacities and the way in which they depend on phylogenetically older processes (Anderson 2014; Brooks 1991).

My aim in this chapter is to defend two related claims:

1. Proponents of the replacement hypothesis neglect the importance of prediction to sensorimotor processing;
2. A predictive model-based understanding of sensorimotor processing can avoid the problems confronted by traditional representational accounts of sensorimotor processing and accommodate the phenomena that motivate the replacement hypothesis.

To this end, I structure the chapter as follows. First, I identify three core elements of what I will call the "classical" model of sensorimotor processing (S3). Second, I highlight those reasons that have led many to embrace the replacement hypothesis in response to alleged deficiencies of the classical model (S4). Third, I argue both that proponents of the replacement

hypothesis neglect the predictive character of sensorimotor processing and that a predictive model-based account of mental representation can accommodate the phenomena that motivate nonrepresentational views (S5). Finally, I briefly address recent arguments that one can embrace a heavily prediction-based understanding of sensorimotor processing without embracing representationalism (S6).

(8.)**3. The Classical Model**

Proponents of the replacement hypothesis are most staunchly opposed to what I will call the "classical" model of sensorimotor processing (see Fodor 1983; Marr 1982; Pinker 1997). This model is classical in two senses.

First, it is most strongly associated with early work in what is often called "classical" or "standard" cognitive science that emerged from the so-called "cognitive revolution" of the late 1950s and 1960s (Shapiro 2010). Second, the model is mostly a thing of the past: even among those who work within a broadly "cognitivist" tradition in contemporary cognitive science, there is a widespread recognition that perception does not work in exactly the way that the classical model describes. Nevertheless, focusing on the classical model will serve to highlight certain assumptions about sensorimotor processing that most obviously motivate the replacement hypothesis.

The classical model contains three core assumptions which concern, respectively, the *function* of perception, the chief *problem* that it must solve, and the kind of *mechanism* by which it solves it.

First, then, the classical model assumes that the function of perception is to map proximal sensory inputs onto a veridical representation of their environmental causes in a form that can be passed onto cognitive mechanisms of belief fixation and decision-making that mediate between perception and action. For example, the visual system must translate patterns of light stimulating retinal cells into a description of the three-dimensional structure of the environment. With this description in hand, downstream cognitive processes can combine it with stored conceptual knowledge and beliefs and goals for the purpose of action-planning (Marr 1982; Pinker 1997). In this way systems of action-planning and action-execution operate both independently of and posterior to perceptual mechanisms, such that "cognition is postperceptual—even in some sense aperceptual—representation-rich and deeply decoupled from the environment" (Anderson 2014, p.164). Once perception does its job, the overall

cognitive system can thus "ignore the actual world, and operate in the model" (Brooks 1999, pp.136-137).

Second, the chief *problem* that this perceptual task confronts is *underdetermination*: proximal sensory inputs radically underdetermine their environmental causes (Marr 1982). For example, there are many (in fact, an infinite number of) possible scenes consistent with any given retinal input. As such, the perceptual system must rely on a substantial body of knowledge ("implicit assumptions") about the structure of the world to narrow down the range of possible perceptual representations and accurately infer the correct one from the impoverished sensory data it receives. In this way the classical model conforms to the basic constructivist understanding of perception that goes back at least as far as Helmholtz (1867), according to which perception is a matter of knowledge-driven inference.

Finally, the classical model delineates a kind of mechanism that can solve this problem and thus enable the perceptual system to reconstruct the state and structure of the distal environment: a *computational* mechanism in which computations are defined over states that *represent* or *stand in for* features of that environment. This appeal to computational operations over representational states provides a defence against the criticism that constructivist models imply an inner homunculus performing the relevant inferences (Pinker 1997). Although in traditional applications of the classical model computation is conceived along the lines of rule-governed operations defined over an arbitrary symbolic code (e.g. Marr 1983), in principle it could also be implemented by alternative kinds of computation. For example, many models of image classification involving vector transformations in hierarchically structured feedforward neural networks (see Lecun et al. 2015) can be understood in terms of the classical model.

(8.)**4. Replacing Mental Representations**

Criticisms of the classical model in the psychological and philosophical literature are varied and complex. Here I will focus specifically on those criticisms that have been most important in motivating the replacement hypothesis. These criticisms derive principally from three related sources: pioneering work by Gibson (e.g. 1979) and the subsequent tradition of ecological psychology that he in large part initiated, work in dynamical systems-theory approaches to understanding cognition (Van Gelder 1995), and certain work in robotics (Brooks 1991; 1999) (see Shapiro 2010). Rather than attempting anything like an exhaustive summary of this enormous body of literature, I will instead focus on just those grievances that emerge most

clearly in the context of challenges to the three core assumptions of the classical model just outlined.

(8.)**4.1. Perception Guides Action**

First, many critics of the classical model deny its assumption that the function of perception is the veridical reconstruction of the distal environment in a form that can be passed onto cognitive mechanisms that mediate between perception and action. Instead, a widespread claim is that the function of perception is to guide or control *action*:

> "Perception is not reconstructive; representing the environment is not what our brains evolved to do. Our brains evolved to control action" (Anderson 2015, p.7).

> "Gibson's ecological approach to psychology… takes as its starting assumption that animals' perceptual systems are geared for action in the world—foraging, finding shelter, avoiding predators and the like—and not creating a replica of the world inside their heads" (L. Barrett 2011, p.155).

Unfortunately, despite the frequency with which such claims are made (e.g. Anderson 2014; Chemero 2009; Engel et al. 2013), it is not at all clear how to interpret them. Someone who endorses the classical model need not deny that the *ultimate* function of perception is to guide action, for example. The claim is rather that perceptual reconstruction of the sort described by the classical model is instrumentally necessary for effective action-guidance. For example, Marr (1983, p.3)—the bogeyman of much work in embodied cognition—explicitly argues that vision science must explain how the visual system extracts "from images the various aspects of the world that are *useful* to us" (my emphasis). Thus when Dawson (2014, p.60) asks, "what if the purpose of perception is not to build mental models, but is *instead* to control actions on the world?" (my emphasis), the question presupposes a tension between these two ideas that need not exist. Or—rather—if there is a tension, it is not one of logical inconsistency.

I will return to this pragmatic perspective on brain function in more depth in the next chapter. For now, I think that there are at least two ways of interpreting the ubiquitous slogan that the function of perception is to guide action. On the first, it embodies a substantial *architectural* claim about sensorimotor processing, in which there are fairly direct links or mappings from sensory inputs to actions (e.g. Brooks 1991). I will return to this suggestion below. On the second, the slogan signals a commitment to a variety of arguments against the view that the classical model identifies a plausible *instrumental* function of perception in biological systems insofar as its ultimate function is to guide action. Three such arguments stand out:

- Insofar as the function of perception is to guide action, one would expect perception to be much more *selective* than the classical model describes.
- Insofar as the function of perception is to guide action, one would expect sensorimotor processing to be much more sensitive to *temporal* factors than the classical model describes.
- Insofar as the function of perception is to guide action, one would expect the organism to employ different sensorimotor *strategies* than the classical model describes.

As we will see, proponents of embodied cognition argue that each of these expectations is in fact vindicated by the evidence.

(8.)**4.2. The Pseudo-Problem of Perception**

Second, proponents of the replacement hypothesis typically deny that perception confronts an underdetermination problem (Anderson 2014; L. Barrett 2011; Chemero 2009). Importantly, they do not deny that, say, retinal images underdetermine their environmental causes, because it is demonstrable that they do. Rather, they claim that the inputs to perception are not static sensory inputs but evolving patterns of sensory inputs across different modalities as the organisms moves around and explores its ambient environment. As Anderson (2014, p.168) puts it,

> "Our perceptual apparatus has evolved and developed as part of a moving body embedded in the world. The visual input to the brain is not like the snapshot of a camera but involves multiple, mutually informing information flows from the eyes, the ears, the limbs, and the rest of the acting, sensing body."

In the case of vision, for example, Gibson (1979) famously rejected the idea that the visual system must infer the hidden state of the world from highly ambiguous retinal images because he denied that the retinal image is the proper starting point for vision. Instead, the inputs to vision are *invariant* aspects of structured light as found in the "ambient optic array" as the animal moves around, where the "optic array" names the light present as a result of diffusion and reflection as it converges on an observer (Shapiro 2010, p.31). Gibson argued that by taking into consideration patterns in the optic array that remain invariant in the face of the organism's movements, the inputs to vision are sufficient to specify their sources—the shape and structure of the environment—without requiring supplementation by stored knowledge (see also Anderson 2014; 2017). In this way perceptual systems are not best understood as in-the-head

inferential devices mapping static sensory inputs onto descriptions of their causes but rather parts of an embodied and exploratory organism. As Gibson (1966, p.5) puts it, "the active senses cannot simply be the initiators of *signals* in nerve fibres or *messages* to the brain; instead they are analogous to tentacles or feelers."

(8.)**4.3. Nonrepresentational Sensorimotor Processing**

This leads directly to the rejection of the third assumption associated with the classical model. According to proponents of the replacement hypothesis, once we acknowledge that the task of perception is to guide action and that this task does not require supplementation by stored knowledge or "implicit assumptions," the door is open to denying that sensorimotor processing requires any internal states to re-present, stand in for, or model the environment. Indeed, the door is open to arguing that internal represnetations would positively *get in the way* of sensorimotor processing (Brooks 1991). Again, there are several broad considerations used to support this radical conclusion in the literature. I will just focus on two of the most influential.

(8.)**4.3.1. Perceptuo-Motor Loops**

The first was implicit in discussion of Gibson's views: proponents of the replacement hypothesis deny that sensorimotor processing involves a sequential process in which sensory inputs are first converted into a detailed representation of the distal world before that representation is passed onto downstream cognitive mechanisms of action-planning and -execution. Instead, they stress the complex ways in which perception and action work together in cyclical processes such that "sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition" (Varela et al. 1991, p.173). As Anderson (2014, p.170) puts it,

> "No organism ever *starts* with sensory stimulation, for each and every stimulation was itself preceded (and is generally accompanied) by an action—action and perception are constant, ongoing, and intertwined."

In this way a defining feature of work in embodied cognition is its stress on "the constant engagement of the world by cycles of perceptuo-motor activity" (Clark 2016, p.1).

A classic example of this dynamic is the "outfielder" problem in which a person's goal is to catch a "fly ball" in baseball (see Chemero 2009, p.114; Clark 1999, p.346; Wilson & Golonka 2013, p.5). Whilst the so-called "sense-model-plan-act" (Brooks 1991) understanding of intelligent response associated with the classical model might lead one to believe that this

problem is solved via the transformation of visual input into a rich internal model that can be exploited to predict the ball's trajectory, evidence suggests that a very different solution is used: players simply *run* in such a way that the baseball's movement through the visual field occurs at a constant speed. When successful, this movement cancels apparent changes in the ball's "optical acceleration," enabling the player to end up where the ball hits the ground (Wilson & Golonka 2013, p.6). Action thus does not occur after perceptual processing has delivered an internal model of the problem domain. Instead, it plays an active role in determining the information to which perception is responsive, which in turn targets only those aspects of that information relevant to the agent's current behavioural goals—a complex cyclical process that (it is argued) characterises the relationship of perception and action more generally (Clark 2008).

A more mundane example of this general phenomenon is the use of frequent (typically unconscious) saccades to relevant aspects of the environment during visually guided tasks (see Clark 2008, p11-4). Specifically, rather than translating retinal inputs into an enduring and detailed three-dimensional model of the world that downstream cognitive processes can operate on, there is compelling evidence that vision is much more selective, exploiting constant eye movements to extract task-relevant high-resolution information from selected foveated locations, often just in time for use (Ballard 1991; Churchland et al. 1994; Clark 1996). As with the outfielder example, then, perception and action appear to conspire to put us into contact with just those features of the environment relevant to our goals and interests at a given time, often exploiting the environment itself as a kind of external source of working memory (Anderson 2014; L. Barrett 2011; Clark 1996).

(8.)**4.3.2. Dynamics**

The second influential argument for nonrepresentational views of sensorimotor processing focuses on its *dynamic* character: the fact that sensorimotor control is something that essentially happens *in time* (Chemero 2009; Van Gelder 1995; 1998). In fact, this emphasis on the dynamics of sensorimotor processing has three related manifestations in the broader literature that are worth teasing apart.

First, some theorists have argued that computational models of cognition are *intrinsically* ill-suited to capturing its temporal characteristics (Van Gelder 1998). According to this argument, computational models are restricted to describing a *sequence* of instructions where time plays no essential role. Specifically, time features within such models only insofar as the sequence

of such operations must conclude within some specific time *limit* determined by the nature of the task, and even this latter consideration is often missing from classical computational models (see below). As proponents of the replacement hypothesis rightfully point out, however, there are many systems where timing plays an essential role in the process itself, not just as a constraint on that process. Given that cognitive systems are such systems, the argument goes, computational models are unable to adequately represent their behaviour.

As a *general* argument against computational or representational accounts of cognition, this is not very convincing (Bechtel 2008; Clark 1996). Specifically, although it is true that temporal considerations do not feature in many classical computational models of cognition and that this is surely a deep problem (see Van Gelder 1998), there seems to be no in-principle reason why such considerations cannot be integrated into such models or why computational and dynamic analyses cannot complement each other (Bechtel 2008). Nevertheless, the general argument feeds a more specific argument that has more bite.

According to this second manifestation, there are specific *kinds* of systems that exhibit a dynamic profile that is not illuminated by concepts such as computation or representation. These are systems or interactions between systems that feature what Clark (2008, p.24) calls "continuous reciprocal causation (CRC)," in which "some system S is both continuously affecting and simultaneously being affected by activity in some other system O." Under these conditions the parts of the system or two systems are *coupled*, such that the "mathematical description of the behavior of one must include a term that describes the behavior of the other" (Shapiro 2010, p.118). A classic example of this dynamic is when two pendulum clocks are placed closed together such that their motions become synchronised (Bruineberg et al. 2016). To model their behaviour, the relevant differential equation describing the dynamics of each individual pendulum must also include a term describing the other.

Crucially, proponents of dynamical systems theory-approaches to cognition argue that this relation of coupling characterises the kind of reciprocal causal interactions exhibited between brains, bodies, and environments, such that—just as with the pendulums—one cannot understand the brain without also understanding its bodily and environmental context (Bruineberg et al. 2016; Van Gelder 1995). As L. Barrett (2011, pp.125-6) puts it,

> "As the nervous system, body, and environment are simultaneously changing and influencing each other in a continual cycle of adjustment (they are "dynamically coupled"), we should

> properly consider a "cognitive system" to be a single, unified system that encompasses all three elements and doesn't privilege the brain alone."

Further, they argue that this kind of interaction is not well-understood in terms of concepts like representation or computation. When two systems or parts of systems are coupled to each other in this way, they argue, there is no need for anything to re-present or stand in for anything else:

> "Perception is always a matter of tracking something that is present in the environment. Because animals are *coupled* to the perceived when they track it, there is never need to call upon representations during tracking" (Chemero 2009, p.115, my emphasis).

A final manifestation of the emphasis on the dynamics of sensorimotor processing points to the importance of timing as a *constraint* on cognitive processes. Specifically, proponents of the replacement hypothesis argue that the classical model describes a form of perception that is insufficiently sensitive to the organism's time-pressured practical engagements with its environment. A classic argument of this sort is given by Brooks (1991; 1998), who argues that that the sequential sense-model-plan-act understanding of intelligent response described by the classical model renders cognitive systems insufficiently sensitive to *change* in their environments (see also Anderson 2003). Specifically, routing all adaptive responses through a central planner operating on an internal model of the world gives rise to what Brooks (1998) calls a "representational bottleneck" that blocks a necessary sensitivity to the way in which environments change over time—often suddenly, and often in task-relevant ways.

Famously, this argument—as well as and the relative lack of success in building robust intelligent systems under the influence of the classical model—led Brooks to build robots that exhibited a very different sensorimotor architecture to that described by the classical model. Specifically, Brooks (see 1991; 1998) designed robots with a "subsumption architecture," in which the overall system is decomposed into semi-autonomous activity-producing "layers" that in most cases connect "perception to action directly" (Brooks 1998, p.90).

Because such robots seem highly dated now, I will not provide a review here (see Clark 1996 for a review; for more recent work in a similar vein, see Pfeifer & Bongard 2007; Pfeifer et al. 2014). Nevertheless, they illustrate a perspective on adaptive success that has been highly influential within work in embodied cognition (see L. Barrett 2011; Shapiro 2010). The direct relations between perception and action they embodied showed how complex forms of adaptive behaviour could emerge without central planners operating on internal models (Clark 1996). Rather than exploiting sensory input to reconstruct the distal environment and then generate

actions conditional on the state of this reconstruction, such systems displayed "the adoption of highly reactive models of perception, bypassing the representational level altogether" (Anderson 2003, p.97). Brooks and many others since contend that this eschewal of internal modelling in favour of close connections between sensing and action provides a superior framework for understanding adaptive response than the one enshrined in the classical model (Anderson 2014; L. Barrett 2011; Chemero 2009). Given the premise that sensorimotor processing constitutes the bulk of what nervous systems do, such theorist conclude that "representation is the wrong unit of abstraction in building the bulkiest parts of intelligent systems" (Brooks 1991, p.14). For our basic interactions with our environments, it is "better to *use the world as its own model*" (Brooks 1991, p.141, my emphasis).

(8.)**4.4. Summary**

The aim of this brief overview of certain themes and ideas from embodied cognition has not been to exhaustively summarise this enormous literature but rather to give a taste for the kinds of considerations—both theoretical and evidential—that have motivated the replacement hypothesis. Specifically, the foregoing remarks should illustrate a stark difference between two competing accounts of the origins of adaptive success in intelligent systems.

In the classical model, organisms coordinate their behaviour with the world by interacting with an internal reconstruction of the world built from impoverished sensory data and substantial prior knowledge. By contrast, work in embodied cognition points to a more selective and action-oriented understanding of the core mind-world relationship, whereby perception and action interact in complex cyclical ways to put organisms into contact with just those features of the world relevant to their time-pressured practical engagements. Although many have embraced this alternative perspective without renouncing the importance of mental representations (e.g. Clark 1996; 2008), proponents of the replacement hypothesis contend that these complex cycles of perceptuomotor activity replace the need for mental representations in sensorimotor processing altogether. As Brooks (1991, p.141) famously put it, "explicit representation and models of the world *simply get in the way*" (my emphasis).

(8.)**5. The World Is Not Its Own Best Generative Model**

If the account of mental representation that I have defended in previous chapters is along the right lines, the replacement hypothesis is mistaken. That much is clear. In this section, however, I want to argue for two more contentious claims: first, proponents of the replacement hypothesis systematically *neglect* the predictive character of sensorimotor processing; second,

a predictive model-based account of mental representation of the sort that I have defended can accommodate the phenomena that motivate the replacement hypothesis.

First, however, I should stress that that the predictive phenomena I point to here are by no means the *only* problems for a nonrepresentational understanding of sensorimotor processing. For example, there are several other powerful critiques of the replacement hypothesis in the literature that have received a large amount of attention. These include that:

- the replacement hypothesis and ecological views drastically underestimate the problem of perception and the noise and uncertainty that it must surmount (Kersten & Yuille 2003);
- nonrepresentational theories either cannot or do not account for the invariance in our perceptual responses in the face of radical changes in proximal stimulation (Burge 2010);
- nonrepresentational theories either cannot or do not account for our capacity to detect highly abstract properties that do not map on to ordinary physical properties (Clark & Toribio 1994);
- nonrepresentational theories either cannot or do not account for or illuminate systematic perceptual illusions (Rescorla 2013).
- nonrepresentational theories either cannot or do not account for the cognitive and behavioural flexibility exhibited in sensorimotor processing (Wilson 2002).

With the exception of the last of these claims, I will not focus on these arguments in what follows, all of which have received substantial attention elsewhere. Instead, I will focus on an aspect of the replacement hypothesis that is highly salient in the context of this thesis: its neglect of the predictive character of sensorimotor processing.

(8.)**5.1. Reactive Models of Perception**

Proponents of the replacement hypothesis neglect the predictive character of sensorimotor processing. This neglect is largely explicit in their work. Gibson (1979; p.243), for example, explicitly denies "that expectation …. [has] an essential role to play in perceiving," Brooks encourages "the adoption of highly reactive models of perception, bypassing the representational level altogether" (Anderson 2003, p.97), and Chemero's (2009) influential description of radical embodied cognitive science treats perception as a matter of *tracking* whereby internal neural states become strongly coupled to external states.

In this sense there is an important similarity between the classical model and the replacement hypothesis: both frameworks treat perception as fundamentally a matter of detecting or tracking what is out there in the world. The chief difference between the two concerns *what* gets tracked and how the tracking is achieved: among proponents of the replacement hypothesis, what gets tracked is much more selective than assumed by the classical model and it is achieved via perceptuomotor cycles through which the organism continually engages the world in action.

By contrast, the account of sensorimotor processing that emerges from previous chapters treats perception as founded on *predictive simulation*. Although the upshot of this prediction-based processing is something functionally akin to tracking (see Chapter 10), it is internally generated predictions that play the crucial role in the generative model-based account of mental representation that I have defended. Insofar as perception involves *predicting*—not just reacting to—the world, it involves resources that are fundamentally decoupled from environmental and bodily processes (Grush 2003).

In the next chapter I will argue that this emphasis on predictive modelling can accommodate the selectivity of perceptual response championed by proponents of the replacement hypothesis. First, however, it will be helpful to revisit those areas where prediction plays central functional roles in sensorimotor processing according to theories such as predictive processing. As will be clear, standard nonrepresentational accounts have no resources to accommodate these predictive functions. By contrast, a predictive model-based account of mental representation can accommodate the phenomena that motivate nonrepresentational views.

(8.)**5.2. Predictive Learning**

First, then, none of the work outlined above in relation to the replacement hypothesis addresses the crucial issue of *learning*: namely, how perceptual systems configure themselves to be able to identify bodily and environmental contingencies in the first place. As we have seen, however, one of the central roles of generative model-based predictive processing is to account for this capacity of nervous systems to capture those aspects of the causal-statistical structure of the body and world responsible for generating the organism's evolving sensory signals.

Crucially, in one important sense predictive processing therefore shares with the ecological approach a strong commitment to the informational richness of the proximal sensory inputs an organism receives, because it is from the spatiotemporal patterns in such proximal inputs that the organism's internal models of the world are learned (Kiefer and Hohwy 2017). Further, generative model-based unsupervised learning is not only free to make use of but almost

certainly *requires* temporally evolving patterns of input in order to be successful in perception (George 2008). Nevertheless, as an enormous body of work from machine learning attests, extracting the relevant information about distal causes from the statistical patterns of input received by a neural network is a highly complex and difficult task (Goodfellow et al. 2016). Predictive processing offers at least a schematic account—with some powerful computational demonstrations of its feasibility—of how this task is achieved. Nonrepresentational views do not.

Strangely, in recent work Gallagher (2017) appeals to the importance of "neural plasticity" as an argument *against* representational views in favour of embodied ones:

> "On the enactivist view, neural plasticity mitigates to some degree the need to think that subpersonal processes are inferential [or representational]. The neural networks of perception are set up by previous experience… Whatever plastic changes have been effected in the visual cortex, or in the perceptual network constituted by early sensory and association areas, such changes constrain and shape the current response to visual stimuli" (Gallagher 2017, p.115).

Without an explanation of how this occurs, however, the appeal to "neural plasticity" here is empty. At best it is a placeholder for an explanation of what underlies that neural plasticity, not an explanation itself. By contrast, prediction error-based updating offers an explanation—one which focuses on the gradual installation of a generative *model* of the distal causes of sensory stimuli.

(8.)**5.3. Predictive Perception**

In addition, we have seen that prediction error-based updating plays a crucial role not just in learning but also in online perception. According to predictive processing, the online estimation of the state of the body and environment is a matter of minimizing the mismatch between externally generated sensory signals and internally generated top-down reconstructions of that sensory signal from a generative model of its bodily and environmental causes. Crucially, this prediction-driven processing can answer worries about the representational bottleneck raised by Brooks and others (see Clark 2016, p.311). Predictive coding ensures that the only incoming information that the system must respond to—that gets propagated forwards—is *unexpected* information: that is, information not already accounted for in systemic response. In this sense modelling is not *posterior* to sensing but *prior* to it, with incoming sensory inputs simply providing feedback on the brain's endogenously generated top-down predictions. Thus when Anderson (2003, p.96) complains of model-based views of sensorimotor processing that "once

the system has built its model, it will not notice changes in its actual environment," this is exactly wrong: the only aspect of incoming information that gets "noticed" is changes.

Importantly, the core case of predictive coding applies to predicting the "present": the evolving streams of proximal stimulation as they arrive. For this to work, however, the system must acquire an understanding of how features of the body and environment are likely to evolve over time, thereby generating flexible expectations about what is going to happen *next*. The fact that forward-looking expectations of this kind play a role in perception is evident both phenomenologically (think of expecting the next note in a piece of music or the suite of multimodal sensory information one expects on turning a car's steering wheel) and well-documented empirically (Lange et al. 2018). For example, there is evidence that in well-learned tasks people pro-actively attend to locations in the environment that they have learned to *expect* task-relevant information to materialise (see Clark 2016, pp.66-9). Crucially, this applies to saccades, which, as Tatler et al. (2011, p.16) point out, "are often proactive, that is, they are made to a location in a scene *in advance of an expected event*" (my emphasis). In this sense, the very utility of pro-active saccades in visually guided tasks pointed to by work in embodied cognition often *depends* on flexible predictive capacities. Indeed, there are some "time-critical behaviours such as driving and ball sports" where "given neural delays between the eye [that is, visual input] and cortex… action control *must* proceed on the basis of predictions rather than perceptions" (Tatler et al. 2011, p.16).

(8.)**5.4. Precision-Weighting**

Moreover, as we have seen in previous chapters, the installation of a predictive modelling engine of the sort described by predictive processing allows for the system to flexibly adjust the influence of bottom-up sensory feedback and top-down predictions according to context. In this way precision-weighting allows systems that employ predictive processing to place a greater emphasis on top-down prediction under conditions where incoming sensory information is noisy or otherwise unreliable and reverse this processing profile when it is better to let that incoming information play more of a driving role. As Geisler and Kersten (2002, p.509) puts it,

> "Different sources of information do not always keep the same relative reliability, and hence a rational perceptual system should adjust the weights that it assigns to different information sources contingent upon their current relative reliabilities."

Without integrating predictions, there is no way for the Gibsonian approach to accommodate this attractive strategy. More importantly, *by* integrating predictions and precision-weighting in this way, predictive processing can effectively subsume highly reactive models of perception as one end of a processing continuum in which top-down predictions are assigned low or no precision, thereby ensuring that perception is maximally "data driven" (Clark 2016, p.251-2; Hohwy 2014).

(8.)**5.5. Predictive Sensorimotor Control**

Finally, we have also seen the crucial role of predictive models in guiding sensorimotor *control*, where a forward model is exploited to generate predictions of the sensory feedback produced by the organism's actions. As several authors have pointed out (Clark 1996; Grush 2003), the use of a forward model in these contexts solves an essentially *temporal* problem: a capacity to flexibly predict the sensory outcomes of one's own actions allows one to exploit those predictions in guiding one's actions *before the actual sensory feedback is available* (see Chapter 7).

Even more interestingly, acquiring a predictive model of the relationship between the organism's actions and its sensory inputs provides an explanation of an organism's knowledge of "sensorimotor contingencies" that lie at the foundation of enactivist accounts of perceptual experience (see Noë 2004; O'Regan & Noë 2001). Those enactivist accounts stress the centrality of "implicit knowledge" or "implicit mastery" of sensorimotor contingencies in perceptual experience, where sensorimotor contingencies are the complex regularities relating the organism's behaviour to specific patterns of sensory stimulation.

Importantly, not only is this emphasis on sensorimotor contingencies consonant with the generative model-based account of mental representation defended here, but it seems to positively demand such an account. The complex predictive capacities that underlie the mastery of sensorimotor contingencies must be explained. Pointing to "implicit knowledge" or "implicit mastery" is—again—at best a *placeholder* for that explanation. In general, a highly flexible predictive competence of the sort highlighted by enactivists does not come for free. As Seth (2014; 2015; see also Clark 2012) notes, the generative models at the centre of predictive processing seem especially well-suited to provide that explanation, capturing the complex regularities relating the organism, its movements, and its sensory inputs in a format that enables that organism to reliably predict the likely sensory consequences of its behaviour.

As we saw in Chapters 6 and 7, however, a generative model-based account of mental representation need not restrict itself to describing how brains model the process by which interactions among features of the world—including the organism's own actions—generate proximal sensory stimulations. Once an organism has acquired a generative model of the world and its dynamics, it can also use that model to generate flexible predictions about how interactions among features of the world—including its own actions—generate *novel states of the world*. This was at the centre of the generative model-based account of intuitive physics outlined in Chapter 7, where a capacity to run predictive simulations facilitates a highly flexible ability to manipulate and predict one's physical environment. Again, this is a crucial feature of "everyday coping" that the replacement hypothesis neglects. As Ullman et al. (2017, p.649) puts it,

> "Human perception cares not only about "what is where" but also about "where to", "how", and "why". We implicitly but continually reason about the stability, strength, friction, and weight of objects around us, to predict how things might move, sag, push, and tumble as we act on them."

To flexibly generate the predictions that underlie our "implicit reasoning" about such physical dynamics is—once more—something that cries out for explanation. Further, the highly flexible nature of our physical predictions in this domain—that is, our ability to reliably generate off-the-cuff predictions for an extremely large number of possible physical events—strongly suggests that this competence is not accounted for in terms of a large list of memorised rules, which would very quickly confront a combinatorial explosion (see Lake et al. 2016), and certainly not in terms of any direct perception-to-action links.

This last point is important. One of the most persuasive critiques of nonrepresentational views of the sort proposed by Brooks that postulate fairly direct mappings from perception to action is that they cannot account for the cognitive and behavioural *flexibility* of sensorimotor processing in humans and many other animals. Specifically, any architecture founded on direct sensory input-motor output mappings "suffers from a combinatorial explosion" as the number of tasks and operations the system can perform with the relevant perceptual information increases: "we must posit an input-output program for each task in which humans draw intelligent inferences" (Goodman & Tenenbaum 2016, p.1). By contrast, with a general-purpose veridical representation of the world, an intelligent organism can exploit it for a large number of *different* operations or tasks (Wilson 2002). The kind of predictive flexibility exhibited in sensorimotor processing thus demands an explanation of exactly the sort that

positing structured generative models provides. By self-organizing around prediction error minimization rather than seeking to learn any specific input-output mapping (see Clark 2016, p.270), predictive processing installs an accurate (see Chapter 9) body of information about the structure of the world that can be flexibly used for a variety of predictive functions.

Perhaps most importantly, with a predictive modelling engine of this kind, a system can take full advantage of the phenomena identified by proponents of the replacement hypothesis and other views in embodied cognition. That is, a flexible predictive modelling engine can underpin an organism's mastery of sensorimotor contingencies, guide exploratory actions that put perceptual systems into contact with expected high-reliability information, and identify those contexts in which different sensorimotor strategies are apt (Hohwy 2014). In this sense there is an important asymmetry here: whereas a generative model-based account of mental representation of the sort that I have defended can subsume the kinds of psychological phenomena that motivate the replacement hypothesis, nonrepresentational views do not have the resources to accommodate the flexible predictive and knowledge-driven aspects of sensorimotor processing that theories such as predictive processing can account for.

(8.)**5.6. Summary**

The upshot of these considerations can be stated in the form of a slogan: *the world is not its own best generative model*. Insofar as sensorimotor systems are concerned with predicting—and not simply reacting to—the world, they are importantly decoupled from the processes that they interact with. By relying on predictive modelling in this way, they are one step ahead of perceptual systems simply involved in *tracking*. They can acquire new bodies of information about the structure of the world, remove redundancies in incoming sensory inputs, and flexibly generate a range of highly adaptive predictions: of what is going to happen next, of the sensory information they are likely to receive, and of the outcomes of their interventions on the world.

As we saw above, proponents of the replacement hypothesis often argue that internal representations or models *get in the way* of dynamic sensorimotor processing. Why coordinate with a model of the world when one can coordinate directly with the world itself? Nicely articulating (but not endorsing) this perspective, Godfrey-Smith (2006a, p.57) notes that model-based views of cognition of the sort that I have defended try

> "to explain intelligence by giving the mind access to something with the same structure as its target. Call this structure S. If the mind's problem is dealing with things that exhibit S, how does it help to

> put something with S inside the head? The mind still has to detect and respond to S, just as it did when S was outside."

Insofar as perception is exclusively involved in *tracking* or *detecting* what is out there, this is a legitimate worry. Insofar as it is fundamentally *predictive*, however, the world itself is insufficient: by exploiting an internal model of salient aspects of the world's causal-statistical structure, predictive sensorimotor systems acquire a highly adaptive competence that nonrepresentational views cannot account for. By contrast, such predictive systems *can* accommodate the phenomena that motivate the replacement hypothesis: for example, perceptuomotor loops, exploiting action for disambiguation, and an acute sensitivity to change.

How might a proponent of the replacement hypothesis respond to the claim that nonrepresentational accounts neglect the predictive character of sensorimotor processing?

One obvious response is to deny that prediction does perform the functions that I have outlined in this and previous chapters. Of course, this is an empirical question and fully adjudicating it would fall beyond the scope of this thesis. Nevertheless, it is looking increasingly difficult to sustain a non-predictive view of sensorimotor processing (see Bubic et al. 2010; Clark 2016; de Lange et al. 2018). I hope that I have done enough in illustrating many of the powerful functions that prediction-based processing can perform to illustrate why.

Another possible response embraces the predictive character of sensorimotor processing but denies that it requires any internal modelling of the sort that I have described here. I will conclude this chapter by responding to two recent arguments of this kind.

(8.)**6. Prediction and Representation**

I have argued that proponents of the replacement hypothesis neglect the predictive character of sensorimotor processing. Several theorists have recently argued that one can accept the truth of theories such as predictive processing without being committed to internal representations or models, however (Anderson 2017; Anderson & Chemero 2013; Orlandi 2014). I will conclude by addressing two arguments for this claim.

(8.)**6.1. Prediction or "Prediction"?**

First, several theorists have argued that the sense of "prediction" involved in predictive processing is nonrepresentational. For example, Anderson and Chemero (2013, p.203) argue that representational interpretations of predictive processing conflate "different senses of "prediction" that ought to be kept separate." One sense of "prediction"—what they call

"prediction$_1$"—"is closely allied with the notion of correlation, as when we commonly say that the value of one variable "predicts" another," and is "essentially model-free"; another sense ("prediction$_2$"), by contrast, "is allied instead with abductive inference and hypothesis testing," and is "theory laden and model-rich" (Anderson & Chemero 2013, p.203). At most, they argue, the evidence for predictive processing is evidence for the ubiquity of prediction$_1$ in cortical activity. Conceptualising such activity in terms of prediction$_2$ is a "theoretical choice not necessitated by the evidence" (Anderson & Chemero 2013, p.204).

Orlandi (2018) advances a similar argument. She argues that representational interpretations of predictive processing should be replaced with interpretations according to which neural states are involved in predicting *other* neural states, not states of the world. "High-level states of predictive systems," she argues, "do not concern what is to be expected *out there*. They concern what is to be expected *in the brain*. Such states are not representations" (Orlandi 2018, p.18). Similarly, O'Regan and Degenaar (2014, p.131) write that "there may exist brain processes that can be viewed as hierarchically organized, with one layer functioning as if it "predicts" the activity of a lower layer. But this should not be taken to say that the higher levels represent *external causes*" (my emphasis; see also Downey 2017 for a similar argument).

There are two problems with this suggestion.

First, it neglects the question of *how* predictive brains acquire the successful anticipatory dynamics that these theorists draw attention to. According to predictive processing, it is by reconfiguring cortical networks to recapitulate—model—the causal-statistical structure of target domains. Predictive success is not accidental. It requires an explanation. It is the role of accurate *models* of target domains within predictive processing to provide that explanation.

Further, as noted in Chapter 5, brains are not trying to predict sensory inputs just for the sake of it. Successful prediction underlies the ability of organisms with predictive brains to engage the *world* in perception, action, and imagination. In the context of predictive processing, such successful engagements are causally dependent on the predictive success that *models* of the world afford. Just as one would not be able to recognise a map as a map if one's focus were exclusively on the piece of paper, one will also not be able to recognise neural structures as models if one's focus is exclusively on the neural structures. Instead, one needs to focus on the way in which the relevant representation enables a representation-user to successfully coordinate its behaviour with a specific target—which, in the case of predictive processing, (typically) exists *outside of the brain*.

(8.)**6.2. Growing Structures**

Orlandi (2014) has also advanced another argument against representational interpretations of predictive processing. She argues that—when stripped of its common representational interpretation—predictive coding merely explains how neural networks "grow" structures that enable differential responsiveness to environmental conditions, and that the internal states that theorists call "priors," "likelihoods," and so on are really better thought of as mere "biases." "In predictive coding accounts," she argues, "priors, hyperpriors and likelihoods do not look like contentful states of any kind. They are not map-like, pictorial or linguistic representations… [R]ather [they] look like built-in or evolved functional features of perception that incline visual systems toward certain configurations (Orlandi, 2014, p.25). Echoing Gibson, Orlandi (2014, p.16) argues that prediction error minimization thereby explains how perceptual systems become "attuned" to the structure of the environment without representing it.

As I argued in Chapter 7, in some contexts this non-representational interpretation might be apt. Nevertheless, there are two problems with Orlandi's suggestion as a general interpretation of predictive coding.

First, the argument that priors and likelihoods do not perform functional roles similar to ordinary public representations such as maps neglects their roles with broader representational structures, which—as I have been at pains to argue—*do* function analogously to public representations.

Second, Orlandi's claim that what predictive coding describes as expectations and predictions are really just "biases" in the circuitry of neural networks that "attune" organisms to the structure of the environment is deeply unsatisfactory here. Again, it focuses on the explanandum, not the explanans: *how* do such "biases" attune organisms to their environments and facilitate the successful prediction of its sensory effects? *All* possible perceptual systems are biased. According to predictive processing, the ones with biases that *successfully* attune the organism to the structure of the environment are those that build internal models that re-present the way in which interactions among features of that structure generate the organism's sensory stimuli.

(8.)**7. Conclusion**

Certain proponents of embodied cognition contend that embodied interactions with the environment in which organisms continually engage their worlds in perceptuo-motor cycles replace the need for mental representations. Among the many problems that this suggestion confronts, I have argued that it neglects the flexible predictive capacities that underlie the way in which creatures like us engage our worlds.

Chapter 9. Modelling the Umwelt

> "One of the basic features of living organisms seems to be their power to utilize, and bring out the possibilities of, parts of the natural world which in the absence of life lie inactive…" (Craik 1966, p.67).

(9.)1. Introduction

The second challenge to representationalism that I will consider is less precise than the previous one, but it has nevertheless had a significant influence in motivating anti-representationalism, especially in traditions such as pragmatism and embodied cognition. Further, it provides an opportunity for addressing a number of interesting phenomena from the perspective of the account of mental representation that I have defended in this thesis.

The challenge has two parts. The first is an emphasis on the pragmatic nature of intelligence. The second is an inference often drawn from this pragmatic understanding of intelligence: that the concept of representation with its implication of re-presenting an independently identifiable external reality fails to capture the many ways in which idiosyncratic properties of the organism contribute to the construction of its experienced world (Anderson 2014; Dewey 1925; Engel et al. 2013; Varela et al. 1991). Importantly, this emphasis on the "organism-relativity" of perceptual experience is especially threatening in the current context because it directly challenges the idea that *similarity* provides the appropriate concept for describing the relationship between what happens within the brain and what happens outside of it—an idea that is seemingly central to any structural representation-based account of the mind-world relation of the kind that I have defended here.

I structure the chapter as follows. In Section 2 I briefly review motivations for a pragmatic understanding of brain function and its alleged implication that perceptual experience is deeply rooted in idiosyncrasies of the organism. In Section 3 I describe the challenge to similarity-based accounts of mental representation that this organism-relativity of experience poses. In Sections 4 and 5 I then show how two of the characteristics of models outlined in Chapter 4 can address part of this challenge: namely, their selectivity (Section 4) and idealisations (Section 5). Finally, in Sections 6 and 7 I argue that a generative model-based account of mental representation can also accommodate our representation of two especially difficult classes of phenomena for similarity-based accounts of the mind-world relation: response-dependent properties (Section 6) and affordances (Section 7).

(9.)**2. The Pragmatic Brain**

The idea that we should understand the mind primarily in terms of its role in guiding action is one of the defining commitments of American pragmatism as it emerged at the turn of the twentieth century in the work of Peirce and then James and Dewey (see Williams 2018b for a review). Both James and Dewey drew on a Darwinian perspective to argue that the mind should be understood in terms of the organism's adaptation to its environment. For example, Dewey (1925) rallied against what he labelled the "spectator theory of knowledge" in which psychological processes are decoupled from practical ends, a view which (he argued) characterised most work throughout the history of Western philosophy. In its place, he advocated an understanding of the mind as an organ for "coping, not copying," in the slogan popularised by Rorty (1989).

As we saw in the previous chapter, this "action-oriented" perspective on cognition is central to work in embodied cognition and ecological psychology. Classic slogans in that tradition, for example, include that "perceptions are written in the language of actions" (Michaels and Carello 1981, p.42) and that "the brain evolved to guide action" (Anderson and Chemero 2016). According to Engel et al. (2015, p.1), the growing appreciation of this fact has led cognitive science to experience a "paradigm shift" in the form of a "*pragmatic turn* away from the traditional representation-centred framework" towards a view that understands cognition "as subserving action."

(9.)**2.1. Sources of the Pragmatic Brain**

There are different sources of this emphasis on the pragmatic character of intelligence in the psychological literature.

One source we have already seen: evolution. According to this view, brains evolved because of their contribution to the organism's survival and reproduction. As such, their functional properties must ultimately be subservient to this wholly pragmatic end:

> "Looked at from an evolutionary point of view, the principal function of nervous systems is to get the body parts where they should be in order that the organism may survive… Truth, whatever that is, definitely takes the hindmost" (P.S. Churchland 1987, p.548).

Another—highly complementary—source comes from a broadly control-theoretic perspective on brain function as a fundamentally regulatory organ of the sort that I attributed to Craik in Chapter 3 and outlined in Chapter 7. Cisek (1999) has developed this view in interesting and

influential ways, drawing on insights from Dewey, cybernetics, and perceptual control theory. As per the discussion in Chapter 7, he notes that living systems are distinguished from non-living systems in acting upon their environments to regulate their essential variables and thus maintain internally optimal states. In this way they effectively self-organize and thus "actively, if temporarily, resist entropic dissolution" (Anderson 2014, p.183). He then takes this fact to imply a fundamental job description for brains: "to exert control over the organism's state within its environment" (Cisek 1999, 8-9) and thus "maintain organism-relevant variables within some desired range" (Anderson 2016, p.7). Viewed in this light, one sees that

> "an organism controls its own behavior, and hence its own fate, by its actions in the world. Its "purpose" is to defend its internal states (i.e., to sustain homeostasis) and the external state of the perceived world, so that it remains within certain limits that are conducive to its survival" (L. Barrett 2011, p.100).[21]

In both cases (the evolutionary and control-theoretic), however, the idea is the same: the function of the brain—including its perceptual systems—is pragmatic. As such, any account of perception or intelligence more broadly should make clear its relevance to the organism's time-pressured practical engagements with its environment.

(9.)**2.2. Organism-Relativity**

As we saw in the previous chapter, some theorists attempt to draw a fairly direct link from this pragmatic or action-oriented understanding of brain function to anti-representationalism. As I argued there, any such link is deeply suspect: there are many contexts in which pragmatic success is dependent on accurate representation. There is another line of argument that I did not explore in the last chapter that is more interesting, however. According to this argument, a pragmatic perspective on brain function should lead one to expect a thoroughly "organism-relative" understanding of perceptual experience that is inconsistent with the concept of representation and its putative implication that an organism's internal states mirror or copy an independently identifiable external world. As Anderson (2017, p.8) puts it,

> "The world properties it is important to pick out for the purpose of reconstruction are not the same as those that best support interaction, and psychology has tended to (mistakenly) focus on the former class of properties to the exclusion of the latter."

[21] For clarity, I distinguish L. Barrett (Louise Barrett) from L.F. Barrett (Lisa Feldman Barrett) in the body of text.

The idea that perceptual experience is radically organism-relative of course has a long history. It is embodied in James's (1907/2000, p.68) claim that "the trail of the human serpent is…over everything," for example, and in Dewey's (1948, p.14) contention that "the organism—the self, the "subject" of action—is a factor within experience and not something outside of it to which experiences are attached as the self's private property." In the tradition of embodied cognition, these ideas are often expressed by drawing on von Uexküll's (1957) highly influential concept of the "Umwelt," which captures the idea of "the world as it is experienced by a particular organism":

> "Even though a number of creatures may occupy the same environment, they all have a different umwelt because their respective nervous systems are designed by evolution to seek out and respond only to those aspects of the environment that are relevant" (L. Barrett 2011, p.80).

Further, an emphasis on organism-relativity also underlies Gibson's (1979) foundational concept of *affordances* as the proper targets of an organism's perceptual sensitivities. Although there is a good deal of controversy over how to understand this concept in the literature (see Section 7 below), a relatively widespread view is that affordances are

> "possibilities for action that can be engaged in with respect to some object, event, place, etc. These are… organism-referential properties of the environment, and because of that they are "ready-made" meaning" (Michaels & Palatinus 2014, p.20).

Finally, the emphasis on the organism-relativity of experience is embraced by the psychological tradition of enactivism, which—in one popular manifestation, at least—rejects the idea that "the world out there has pregiven properties" that "exist prior to the image that is cast on the cognitive system, whose task is to recover them appropriately" (Varela et al. 1991, p.172). "The criterion for success of cognitive operations," write Engel et al. (2013, p.202), "is not to… construct a veridical representation of the environment. Instead, cognitive processes *construct the world* by bringing forth action-relevant structures in the environmental niche" (my emphasis). As Thompson (2010, p.59) puts it, "instead of internally representing an external world in some Cartesian sense… [organisms] *enact* an environment inseparable from their own structure and actions" (my emphasis); that is, they "constitute (disclose) a world that bears the stamp of their own structure."

(9.)**3. The Challenge to Similarity**

As noted above, this organism-relative understanding of experience has been widely recruited to challenge representationalist accounts of the mind. However, the link between the two—

between organism-relativity and anti-representationalism—is far from clear (Shapiro 2010). Why can't representations reflect idiosyncrasies of the representer? Just think of the infinite number of possible paintings that different artists might produce of the same landscape, each reflecting idiosyncratic perspectival takes on the same scene.

In the current context, however, I think that the challenge can be made precise. At the core of the generative model-based account of mental representation that I have sought to defend is the concept of structural similarity: a structure-preserving mapping between the elements of the model and its target domain. If perceptual experience is as radically organism-relative as the foregoing suggests, however, it is difficult to make sense of this concept of similarity. Similarity of structure implies the existence of two domains that can be characterised independently of one another: the representing world and the represented world (Palmer 1978). In Rorty's (1979) famous phrase, it thus seems to imply an image of the mind/brain as a "mirror of nature." One way of understanding the appeal to organism-relativity is as a direct attack on this idea (Anderson 2014; Price 2011). Specifically, it challenges the idea that there are representational systems in the brain that mirror the structure of the world as it exists independently of such systems and the organisms of which they are a part.

In fact, there are really two issues here. The first consists of this direct attack on representationalism from considerations of organism-relativity. Strictly speaking, however, any attack on the idea that concepts such as accuracy or mirroring are relevant to our understanding of the mind-world relation poses a problem for the account of mental representation that I have defended here, where structural correspondence between generative models and target domains is supposed to be causally relevant to the successful exercise of our psychological capacities (see Chapter 5). As we will see below, then, even those who criticise the assumption of accuracy *within the context* of representationalist accounts pose a problem for the *specific* account of mental representation that I have defended.

One of the major problems in this context, however, is the lack of clarity on what exactly is being claimed when theorists point to organism-relativity or a lack of accuracy. Although I gestured above to some ubiquitous terms and slogans in the literature in embodied cognition and enactivism—"affordance," "Umwelt," "enacting a world," "the brain evolved to guide action"—it is often difficult to understand exactly what is being claimed by theorists advancing these ideas.

As I hope to make clear, the account of mental representation that I have defended helps to clarify these issues. Further, I will argue that this account can accommodate certain phenomena that seem most problematic for theories of the mind-world relation founded on relations of copying or mirroring: namely, response-dependent properties and affordances. Nevertheless, I will also argue that the most radical claims concerning organism-relativity should be rejected: although our mental models are selective, idealised, and deeply involved in representing the world as it relates and matters to *us*, there is no reason to deny—and there are powerful reasons to accept—that they provide largely accurate recapitulations of "real patterns" (Dennett 1991) in the objective world.

(9.)**4. Selectivity**

As I stressed in Chapter 4, models are subject to at least two important forms of idealisation: they are often highly selective, abstracting away from many features of their target domain, and they are typically not completely accurate with respect to those phenomena that they do represent. As Weisberg (2013, p.113) puts it, "models often fall short of veridicality with respect to their targets. To put this point in model-oriented language, models are often not highly similar to their targets." It is no part of the account defended here, then, that the brain's generative models capture all parts of the ambient environment with perfect accuracy. As I stressed in Chapter 4, models are always interest-relative because interests determine these two species of idealisation: which features of the target domain the model represents, and how much accuracy is desirable for the task at hand (Giere 2004).

The first thing to point out, then, is that there is nothing in the account that I have defended that implies that mental models cannot be selective. Of course, in one sense this emphasis on selectivity is trivial. Nobody has ever suggested that our brains contain representations of the position of each of the subatomic particles in the universe, for example. Nevertheless, once one posits structural models of the world inside our heads, how does anything less than this exhaustive representation get motivated? As Haugeland (1987, p.92) puts it,

> "One thing that's frightening about ''mental scale models'' is that there's no obvious end to them: Why not just recreate the entire universe, monad-like, inside each individual brain? Well, because it's manifestly absurd, that's why. But what could motivate, or even delineate a more modest scheme?"

In Chapter 7 I suggested that the concept of regulation motivates such a scheme: organisms model only those features of the world relevant to their core homeostatic imperative to survive.

This provides a principled restriction on their scope (Seth 2015). Insofar as different features of the world are relevant to the survival of different organisms, then, a straightforward kind of organism-relativity falls out of the account of mental representation that I have defended here quite directly. L.F. Barrett (2017a; 2017b) introduces the helpful concept of the "affective niche" to capture this selectivity of prediction error minimization when it is situated within a broader regulatory understanding of brain function. "Anything outside of your affective niche," she writes, "is just *noise*: your brain issues no predictions about it, and you do not notice it" (L.F. Barrett 2017b, p.73).

Nevertheless, L.F. Barrett then goes on to make a mistake that permeates discussions of organism-relativity in the psychological literature. Describing homeostasis as the maintenance of one's "body budget," she writes:

> "From the perspective of your brain, anything in your affective niche could potentially influence your body budget, and nothing else in the universe matters. That means, in effect, that *you construct the environment in which you live*" (L.F. Barrett 2017b, p.83).

This does not follow, however. The fact that a map selectively attends only to a subset of features of a terrain does not mean that the map—or the map-user—"constructs" those features.

Sometimes the only thing that proponents of an organism-relative understanding of perceptual experience seem to have in mind is this focus on selectivity. As Clark (1996, p.25) puts it, "biological cognition is highly *selective*, and it can sensitize an organism to whatever… parameters reliably specify states of affairs that matter to the specific life form" (my emphasis). Likewise, L. Barrett (2011, p.81) elaborates on the concept of the Umwelt by noting that "while the environment is full of all kinds of objects and various kinds of stimulus energy, an organism is sensitive only to those that are relevant to its needs." Crucially, however, this selectivity is not inconsistent with the idea of structural similarity: there can be a structural correspondence between the layout of a map and a target terrain, even if the map does not exhaustively capture every aspect of that terrain.

This selectivity can manifest itself in various ways at multiple levels. At the most basic level, there is the obvious fact that different organisms are responsive to different kinds of stimuli. Humans are not sensitive to ultraviolet light or ultrasonic sound, for example, whereas other animals are (L. Barrett 2011, p.81). Further, there is the level of granularity at which we represent the world. As noted above, we represent the physical world mostly in terms of middle-sized objects rather than, say, in terms of the positions of—or a probability distribution

defined over positions of—subatomic particles (Ullman et al. 2017). At faster time-scales, we saw in the previous chapter some reason to think that *online* perception is also highly selective, focusing only on those features of the environment relevant to a given task (see Clark 1996; 2008). Finally, we will also see below some more interesting examples of the way in which what we represent is strongly tied to our idiosyncratic responses and behavioural profiles.

The upshot of this is that the generative model-based account of mental representation defended here can happily embrace at least one manifestation of the claim that we engage the world from the perspective of our contingent interests. Nevertheless, I think it is also important to push back against some of the more radical claims of selectivity or partiality in the literature on this topic. Often these arguments are supported by reference to the selective response of creatures such as ants (L. Barrett 2011, p.80), ticks (Von Uexkull 1934), or jewel beetles (Hoffman et al. 2015). These organisms have very little behavioural flexibility, however, and nothing like the predictive modelling engines realised in hierarchically structured—and highly malleable—neocortices of the sort that I have described in this thesis. Predictive processing describes a processing routine whereby brains unravel the objective causes of patterns of proximal stimulation at multiple spatiotemporal scales in a form that can be exploited for a very large number of possible predictions and interventions (see Chapter 8). Although there is an interesting theoretical perspective from which we can note that our representation of the world is similarly nevertheless partial and interest-relative, it is important not to overplay this similarity.

(9.)**5. Accuracy**

Now consider the second species of idealisation in models: not only are they selective, but they are rarely if ever perfectly accurate. If there are models of the world inside our heads, then, they will no doubt be comparatively low-resolution models that fail to be strictly isomorphic to worldly structures (Cummins 1994). In Craik's (1943, p.53) words, they will be "working models" that are "bound somewhere to break down…"—that is, to fail to capture reality with perfect accuracy. In this respect, however, they are no different from many other models. Again, we saw in Chapter 4 that a map can be highly inaccurate and yet still useful as a proxy for a domain given certain practical ends, capturing only very coarse-grained characteristics of the spatial structure of the relevant terrain.

Of course, any discussion of accuracy has to address the question of what is being represented. I will set aside both a strangely influential view in the literature according to which the only

"real" entities and properties are those described by physics (see, e.g. Feldman 2015, p.1524), as well as traditional forms of philosophical scepticism that question whether there *is* an external reality (e.g. Varela et al. 1991). As noted in Chapter 5, I assume a bland form of realism for the purpose of this thesis: the universe in general is not constitutively dependent on activity within the nervous systems of a particular species of primate (i.e. us), and it can be accurately described at multiple levels of spatiotemporal scale and abstraction.

Given this, are there any reasons for denying the claim that our successful predictive abilities and the psychological capacities dependent on them are underpinned by generative models that recapitulate the causal-statistical structure of target domains? The first thing to say is that there seem to be some compelling reasons for *endorsing* this claim. For example, appealing to structural resemblance offers at least a schematic explanation of predictive success, prediction error minimization conforms to approximately Bayes optimal ways of combining prior expectations with sensory evidence, and it delivers a body of information not geared to any specific behavioural tasks or input-output mapping—a kind of flexibility strongly suggestive of veridicality (Wilson 2002). As such, I think that there is a strong presumptive case in favour of the accuracy of our generative models. Because I have dealt with these considerations in previous chapters, however, I will not elaborate on them here.

Nevertheless, the fact that there is a strong presumptive case in favour of the view that our internal models are broadly accurate should not be confused for the claim that they perfectly mirror the structure of their targets. As with all models, the models inside of our brains are answerable to a host of pragmatic considerations—for example, simplicity, the generation of fast predictions, the minimal use of energy expenditure—that likely result in all kinds of idealisations that systematically bias them away from strict isomorphism to their targets. As Craik (1966, p.68) puts it, "often the reproduction is not identical with the original—it is intentionally different in order to be more convenient for the purpose in hand."

A concrete example of this use of idealisations comes from the work by Battaglia et al. (2013) that models intuitive physics in terms of a probabilistic scene simulator (see Chapters 7 and 8) (see also Lake et al. 2016; Ullman et al. 2017). According to such theorists, although the generative model that underlies intuitive physical predictions in our brain "represents the world with a reasonable degree of fidelity," it is still "inherently approximate," relying on all sorts of idealisations that "favour speed and generality over the degree of precision needed in engineering problems" (Battaglia et al. 2013, p.18,328). As such, "relative to physical ground

truth, the intuitive physical state representation is approximate and probabilistic, and oversimplified and incomplete in many ways" (Lake et al. 2016, p.10). For example, these researchers speculate that our generative models for physical predictions may "dramatically simplify objects' geometry, mass density distributions, and physical interactions" (Battaglia et al. 2013, p.18,328), and "rely on numerous shortcuts and hacks to efficiently simulate approximations to Newtonian mechanics for complex scenes in real-time" (Ullman et al. 2017, p.1).[22]

The point to stress here, however, is that such idealisations do not undermine the view that such models represent real features of the objective world's causal structure. As I noted in Chapter 4, one of the chief attractions of focusing on structural *similarity* (rather than isomorphism or homomorphism) is that it characterises accuracy as a graded notion (see Giere 2004; Weisberg 2013). A commitment to the existence of mental "working models" thus does not entail that such models literally mirror the structure of target domains. It just requires that they recapitulate worldly structure with a sufficient degree of fidelity to be practically useful for flexible predictions.

Of course, some models are just completely wrong-headed (Weisberg 2013). Is there any reason to suppose that this would be the case with respect to our mental models?

(9.)**5.1. Challenging Accuracy**

One classic example of scepticism about whether our perception of the world maps onto objective structure concerns colour perception. According to a long tradition in philosophy and psychology, colour perception reflects idiosyncrasies of human perception, not real structure in the world (see, e.g., Hardin 1988). This debate is fraught with controversy, however, given that some aspects of colour perception do seem responsive to objective structure in the world of the sort described by the natural sciences—for example, surface reflectances—whereas others do not—for example, the similarity structure of our colour space (e.g. the fact that red and purple seem highly similar despite occupying opposite ends of the surface reflectance spectrum) (Hardin 1988). As such, one might think that this debate would benefit from a graded notion of accuracy rather than a binary (e.g. "true/false") one (Churchland 2012).

Nevertheless, whatever stance one takes on this issue, colour perception occurs at the lower levels of representational hierarchies in the neocortex and seems to be highly influenced by the

[22] For a thorough review of these potential hacks, which include explicitly excluding certain objects from the simulation and treating objects at rest and at motion differently, see Ullman et al. (2017).

structure of our own visual systems. As one ascends cortical hierarchies and finds representations that reach deeper into the world's causal structure, it is thus plausible that such forms of idiosyncratic or "subjective" representation disappear (see Ryder forthcoming). This is why the relevant metaphysical controversies concern *colours* and not, say, dogs or middle-sized physical objects. Given this, any theory that bases its general understanding of the mind-world relation on the peculiarities of colour perception (e.g. Varela et al. 1991) is extremely dubious (Shapiro 2010, p.55). As such, is there any reason for thinking that radically idiosyncratic forms of representation or perceptual response generalise beyond this example?

Several theorists have recently sought to challenge the appeal to the explanatory relevance of accuracy *in general* in the specific context of predictive processing or the Bayesian brain hypothesis. For example, in a treatment that otherwise has much in common with the account of mental representation that I have defended in this thesis, L.F. Barrett (2017a, p.6) claims that

> "modelling the world "accurately" in some detached, disembodied manner would be metabolically reckless. Instead, the brain models the world from the perspective of its body's physiological needs.

The problem with this claim, however, is that it fails to disentangle different dimensions of representation. For example, it is plausible that modelling the world *exhaustively* would be metabolically reckless, as noted above. Without further elaboration, however, it is difficult to see how modelling the world *inaccurately* would be more adaptive or energetically efficient than accurate representation. L.F. Barrett thus seems to fall into two traps: first, conflating how exhaustive a representation is with how accurate it is; second, setting up a false contrast between accuracy and pragmatic success.

In recent work Feldman (2013, 2015) has advanced a positive argument that attempts to explicitly undermine the link between accurate generative models and pragmatic success. Feldman's argument is directed specifically at what he takes to be the widespread and erroneous commitment in discussions of Bayesian models of perception to what he calls the "Lord's Prior," the view that "the optimal Bayesian observer is correctly tuned when its priors match those objectively in force in the environment" (Feldman 2013, p.15). Although Feldman's argument does not challenge representationalism, it does explicitly challenge the commitment to a structural correspondence between internal models and target domains that I have advocated here. According to Feldman (2013; 2015), it is just a mistake to assume that

Bayesian mechanisms inside the head require accurate internal representations of the reality outside:

> "Bayesian inference does not require—nor, indeed, in any way involve—the literal truth of any hypotheses. All that is needed is that selection of hypotheses guide action "effectively."... *Veridicality is a red herring*" (Feldman 2015, p.1525, my emphasis)

Clark (2016, p.272) concurs, arguing that the

> "deep problems [with the concept of the Lord's Prior] emerge as soon as we reflect that active agents are not, at root, simply trying to model the data so much as to come up with recipes for acting appropriately in the world."

As we have seen, however, setting up a strict contrast between pragmatic success and accurate representation is confused; there are many contexts in which the latter is instrumental to the former. What positive argument, then, do Feldman and Clark provide for thinking that the two come apart in the case of mental representation? Both focus on the dangers of *overfitting* a model to a set of observed data:

> "Generally to avoid overfitting, one must be willing to allow the data to be fit imperfectly, leaving some variance unexplained… In any realistic situation, one does not really want to fit the data perfectly, because some of the data are noise—random fluctuations unlikely to be repeated. Overfitting thus inevitably leads to poor generalization" (Feldman 2013, pp.24-5; see also Clark 2016, p.272).

This argument is not convincing, however.

First, note that the point is not specific to the models inside our heads; it captures a constraint on *all* statistical modelling. As such, there is no *specific* problem for mental representation here.

Second, the argument is self-defeating: the very claim that "observations are a mixture of reliable and ephemeral factors" (Feldman 2013, p.25) *presupposes* that the task is to exploit the data to capture the *real* data-generating process. Data are noisy or unreliable when they do not reflect this process. The reason one does not want to set one's priors to exactly match observed frequencies is thus precisely that this would result in *inaccuracy*.

Third, the contents of generative models are not restricted to priors defined over possible world states (see Chapter 7). They also capture the complex causal and statistical *relationships* among worldly features, and Feldman's argument does not address this.

Finally, Feldman expresses his argument in terms of scepticism about "the literal truth of… hypotheses" and the exact matching of prior distributions to statistical regularities (Feldman 2015, p.1525). As noted above, however, representational success in models is not a *binary* phenomenon. The question is whether the models inside our heads capture real patterns that structure the body and environment. At best, Feldman's argument just shows that strict isomorphism is not required—and perhaps not even desirable—for such models. To reiterate, however, *all* models are answerable to pragmatic considerations that often systematically bias them away from strictly mirroring the structure of their targets.

(9.)**5.2. Summary**

To summarise, then, there are two characteristics of the generative models that I have described in this thesis that can accommodate at least some aspects of the organism-relativity of perceptual experience: the fact that they are *selective*, abstracting away from those features of the world irrelevant to the organism's survival, and the fact that they are answerable to a host of pragmatic considerations that bias them away from strictly mirroring their targets. Claims that accuracy is instrumentally *irrelevant* to these models' functioning should be rejected, however. These claims either pick on examples that do not generalise well, fail to distinguish between accurate representation and exhaustive representation, or set up a false contrast between accurate representation and pragmatic success.

Nevertheless, there is a sense in which the considerations outlined so far are quite limited when it comes to the most serious challenges to a similarity-based understanding of the mind-world relation posed from considerations of organism-relativity. These challenges draw on features or putative features of our experienced world that do not seem to map on *at all*—no matter how selectively or imperfectly—to properties of the objective "mind-independent" world.

I think that there are two important kinds of phenomena here, and it is worth disentangling them: response-dependent properties and affordances. I will consider them in turn.

(9.)**6. Response Dependence**

Consider the following properties: cuteness, disgustingness, loveliness, grossness, beauty, and attractiveness. Properties such as these seem to saturate our experienced world. Further, they seem to be properties that we straightforwardly *respond* to: we recoil from the disgusting smell coming from the bin, our attention is drawn to the attractive face from across the room, we can't help but pat the cute puppy. Such regularities appear to be counterfactual-supporting: if

the bin did not smell and the person and puppy were hideous, we—and others like us—would have acted differently. Nevertheless, unlike other features of our experienced world—the shape and weight of objects around us, for example—these properties do not seem to be objective. Instead, as many philosophers have argued, what we intuitively think of as features of the environment that we *respond* to in such cases are probably better thought of as an *effect* of our contingent evolutionarily endowed or acquired responses to the world (Price 2011). To use Hume's (1751/1998, Appendix 1.21) famous phrase, in such cases the mind "gilds and stains" the world. As such, these "response-dependent" properties do not seem to be features of the world's structure that mental models might recapitulate. They do not seem to be features of the world's structure at all.

This is too quick, however. To see this, it is helpful to draw a distinction. As we saw in Chapter 6, for the brain to be a prediction machine it must be able to anticipate the sensory inputs generated from the environment. Crucially, such sensory inputs are ultimately just objective patterns of proximal stimulation across the organism's sensory transducers. Although there is a sense in which the causal process by which such patterns of stimulation are generated is idiosyncratic to the organism—it is the organism's sensory transducers, after all, with their own peculiarities (i.e. sensitivity to just a specific range of inputs)—this process itself is straightforwardly objective. As such, we should expect the brain's generative models of this causal process to literally reflect—albeit no doubt highly selectively and in an idealized manner—the causal process and states implicated in that process that are parts of the real organism-independent environment.

To function effectively as a prediction machine, however, brains must also be able to reliably anticipate not just the sensory inputs generated from the environment but also the organism's contingent *responses* to such sensory inputs, where the concept of responses here should be understood liberally to include not just behavioural dispositions but also internal physiological responses (Clark 2018; Dennett 2013).[23] As Dennett (2013, p.210) puts it,

> "[A]mong the things in our *Umwelt* that matter to our well-being are *ourselves*! We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will *expect* next! And we do."

[23] Note that Dennett and Clark's primary concern here is to explain the appearance of *qualia*, a theoretical aim which I do not sure in the current context.

Crucially, this task of self- (rather than mere sensory-) prediction requires that brains identify those features of the world that *elicit* such responses and behaviours (Clark 2018). Exactly analogous to the case of sensory inputs, then, this will require that they model not just those features of the world and the interactions among them implicated in generating sensory inputs but those features and interactions responsible for eliciting the organism's responses to them. Crucially, this relationship between environmental contingencies and *responses* is just as objective—just as much part of the real causal structure of the world—as the relationship between environmental contingencies and patterns of proximal stimulation. The same story will thus apply: *cuteness* and *loveliness* are just latent variables—or, more likely, complex networks of latent variables—implicated in the brain's generative model of how specific kinds of responses are elicited from the organism.

If this is right, it suggests that the way in which we model the world is at least in part a function of our own contingent behavioural dispositions and response profile. The concept of structural similarity still seems to be equally fitting, however: response-dependent properties may not be features of the objective world considered independently of the organism, but they feature in regularities implicating the objective environment-organism relationship of the sort that mental models can recapitulate in the service of prediction.

It is an interesting question to what extent this concept of response dependence generalises beyond the obvious cases above. For example, Dennett (2013) suggests that even our representation of colours is grounded in our idiosyncratic suite of responses and behavioural dispositions, suggesting an additional complication to the brief treatment of colour in Section 5, and other theorists have sought to extend the concept of response dependence so that it penetrates much deeper into our experienced world than the examples listed above (e.g. Price 2011). Nevertheless, I will not pursue this interesting question here.

(9.)**7. Affordances**

The second species of organism-relative properties often identified in the literature are *affordances*. The term "affordance" was introduced by Gibson and is central to much work in ecological psychology and embodied cognition. He defined affordances as follows:

> "The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in

> a way that no existing term does. It implies the complementarity of the animal and the environment" (Gibson 1979, p.127).

Affordances, Gibson noted, have a peculiar ontological status:

> "[A]n affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer" (Gibson 1979, p.129).

For many theorists, the idea that perception is fundamentally perception *of* affordances is intended to replace the idea that perception is representational (Anderson 2014; Chemero 2009; Gibson 1979). As Anderson (2014, p.162) puts it, "many if not most of our neural states will not be usefully understood as representations, and here the concept of affordances will have important work to do for us."

Unfortunately, as these quotes from Gibson reveal, there is very little clarity—and much disagreement—on how to understand the concept of affordances in both contemporary psychology and philosophy. In some treatments (e.g. Dennett 2013; 2017), it is clear that the term is used to express characteristics of perception that we have largely already dealt with: an organism's perceptual sensitivities are highly selective, likely distortive in certain ways, and as much focused on those features of the environment that reliably elicit its contingent behaviours and responses as generate its sensory inputs.

Other authors intend by the concept something potentially more ontologically radical, however. To see this, it is useful to focus on one of the most influential case studies used in the literature: the so-called "size-weight illusion" (see Anderson 2014, p.173). This names the robust finding that when a human is given two objects of the same weight but different sizes, she will judge the smaller object to be heavier. There are many attempts to explain this finding in the literature. One of the most influential is the so-called *sensory mismatch hypothesis*, according to which the effect arises from the expectation that the smaller object will be lighter being combined with the experience of the actual weight of the object—a kind of prediction error that results in an inaccurate weight judgement (Ross 1969).

Many proponents of a broadly ecological perspective on psychology argue that such explanations are misguided. Specifically, they argue that this alleged "illusion" reflects the fact that people are actually tracking—and *accurately* tracking—something other than weight:

throwability (Zhu & Bingham 2011). Crucially, throwability is not a mere function of an object's weight but rather various characteristics of the object (size, weight, density), various characteristics of the *thrower* (hand size, strength, etc.), and the relationship between these things. There is some limited evidence in favour of this alternative interpretation. For example, subjects will typically report as the same weight all objects that they are presented with that are optimally throwable even when their weights differ, suggesting that what has been labelled the experience of "heaviness" is really the experience of "throwability"—a characteristic of the object that is (as Gibson would put it) "equally a fact of the environment and a fact of behavior" (see Anderson 2014, pp.173-7).

This proposal is speculative and controversial, but it serves to illustrate a broader point emphasised by ecological psychologists: insofar as brains evolved to guide action, we should expect organisms to be primarily sensitive not to what Anderson (2014, p.174) calls "observer-neutral" properties but rather "ecologically relevant properties such as whether an object is edible, reachable, graspable, movable, liftable, or throwable." Although some of these can probably be subsumed within a broad category of response-dependent properties, others pertain to our *abilities*—to the kinds of behaviours and actions that we are *able* to produce as specific kinds of embodied organisms. Unlike a characteristic like shape or weight, these abilities determine affordances that are features of the environment only insofar as it matters to *us*. From this fact Anderson (2014, p.175) extracts an anti-representationalist conclusion:

> "Our brains are not primarily in the business of constructing observer-independent models of the world. We will only rarely have native neural sensitivities to the kinds of reconstructive properties… *that would be best for model building*" (my emphasis).

The challenge here should by now be familiar. Mental models are supposed—according to the argument—to capitalize on structural similarity between neural structures and environmental structures. Insofar as perception is primarily sensitive to affordances, however, it is not primarily sensitive to extra-neural environmental structures *as such*. Rather, it is sensitive to features of the environment inextricably determined by relations *between* the organism and environment. As such, an affordance-based understanding of perception contradicts a model-based one. As Anderson (2015, p.7) puts it,

> "Because perception is both active and in the service of action, much of the information to which organisms are attuned is not objective information of the sort one might need for model-building, but rather *relational* information that is more immediately useful for guiding action in the world" (Anderson 2015, p.7).

(9.)**7.1. Affordance Models**

Although this line of reasoning has evidently been seductive to many psychologists and philosophers, I think that it should be rejected.

First, the fact that affordances are party determined by contingent features of the organism—specifically, that they are a function of both environmental and bodily variables—does not mean that they are hallucinated. As Anderson (2014, p.181) himself puts it, a commitment to the centrality of affordances to perception does not mean

> "giving up on objective in favour of subjective perception or of pitting accurate against inaccurate perception; reachability (for instance) is an objective, relational, scalar property of objects in an agent's environment, and it is quite clear that agents generally perceive this property accurately."

As such, affordances are features of the objective causal structure that relates organisms to their environments. Therefore there is no in-principle reason why the regularities that they are implicated in could not be recapitulated in neurally realised models. For this reason, the contrast that Anderson draws between "relational information" on the one hand and "objective information of the sort that one might need for model-building" on the other is confused. Anderson seems to assume that the only relational structures that an organism could be concerned with modelling are purely environmental structures. But—again—there is no a priori reason why brains could not model the networks of complex dependencies that obtain *between* features of the organism and its environment. As Horst (2016, pp.187-8) puts it,

> "When reliable connections exist between features of the environment and particular things we need or do, affordance detectors pick out real patterns…They are simply real patterns that involve *us*, with our particular set of needs, wants, and capacities, rather than patterns that do not depend on us."

Of course, Anderson might respond that there is no reason to think that we *do* build models of affordances—that there is no reason to build models of affordances when one can just interact with them "directly." Consider a core reason that we have seen for exploiting internal models in the case of sensorimotor processing, however: their predictive capacities. Such predictive capacities do not become obsolete in the case of affordances. In fact, quite the contrary: insofar

as affordances are those features of the environment relevant to our actions, they constitute an area where prediction would seem to be essential (see Chapter 7; Jeannerod 2001).[24]

For these reasons, attempts to draw a distinction between affordance-based perception and model-based perception are confused. Affordances can feature as variables or networks of variables in generative models. The fact that such models do not capture causal structure of the sort that might be recovered from an impartial description of the organism-independent environment merely reflects the fact that their target is the complex web of dependencies between organisms and environmental parameters, not that they are not models.

Nevertheless, I think that it is also important—once again—to push back against claims to the effect that *all* we experience are affordances, at least if this is meant to suggest that we never represent "observer neutral" objects and attribute properties to them (as per Gibson 1979). Insofar as the brain is tasked with the challenge of predicting the sensory signals generated by the environment and the outcomes of interactions among those hidden causes, it will pay to reconstruct those objective characteristics of the environment—including the organism—responsible for generating those sensory signals. This lies at the heart of standard descriptions of predictive coding in the neuroscientific literature, where perception is described as a matter of inverting the process by which sensory signals are generated (e.g. Friston 2005; Huang & Rao 2011; Rao & Ballard 1999). The advantage of reconstructive—albeit idealised—representation of this kind is that it facilitates an adaptive flexibility that more myopic kinds of perceptual sensitivity preclude (Battaglia et al. 2013; Lake et al. 2016). By installing largely veridical models of the objective process by which sensory signals are generated, predictive minds acquire a body of information that is importantly "purpose-neutral" (Wilson 2002, p.622)—that can be exploited both for identifying affordances relevant to our immediate physical actions, but also for a range of flexible predictions that do not pertain to our immediate actions (Battaglia et al. 2013). That is, for organisms *like us*, many of the "ecologically relevant" properties *are* "observer neutral" properties—properties of the sort that many *different* kinds of organisms would latch onto insofar as they have to coordinate their actions with the same physical environment at multiple spatiotemporal scales.

(9.)**8. Structural Representation and the Manifest Image**

[24] As I have argued elsewhere (Gadsby & Williams 2018), there is evidence from cognitive neuropsychiatry that individuals with anorexia nervosa systematically *misperceive* the affordances of their environment precisely because the predictive models on which their capacity to evaluate potential actions is based rely on distorted information about the shape and size of their bodies.

The predictive modelling engines inside our heads construct and exploit generative models that are selective, idealised, and interest-relative, modelling the world as it relates and matters to us as specific kinds of organisms. Nevertheless, our "Umwelt" is much less parochial than many proponents of views in embodied cognition claim, and we model it in a way that recapitulates—albeit in an idealised manner—its objective regularities and our place within them. Pragmatic success is not an alternative to veridicality and the fact that we do not model the world exhaustively does not mean that we do not model it accurately. If you want to survive in a hostile and highly complex world whose regularities span multiple spatiotemporal scales, it pays to have mental models that recapitulate the objective patterns relevant to this goal.

10. Generative Intentionality

> "The deepest motivation for intentional irrealism derives…from a certain ontological intuition: that there is no place for intentional categories in a physicalistic view of the world; that the intentional can't be *naturalized*" (Fodor 1987, p.97).

(10.)**1. Introduction**

The third and final challenge against representationalism that I will consider targets representational content. It has two premises: first, that an essential property of mental representations is content, typically understood in terms of satisfaction conditions of some kind; second, that there is no such thing as representational content, or at least no such thing as content of the sort posited in contemporary cognitive science. The second premise is typically justified by appeal to what Hutto and Myin (2012) call the "hard problem of content": the challenge of identifying in a non-circular way those natural properties and relations in virtue of which a putative mental representation expresses the content it does and not some other content or none at all. Among the many characteristics of content that have generated challenges for this reductive project, perhaps the most recalcitrant property has been its apparent normativity: the fact that representations can *mis*represent—that they can be in *error* relative to how things are (see Ryder 2009b for an overview).

At least since Quine's (1960) infamous attack on the scientific credentials of meaning, an influential trend in philosophy contends that the hard problem of content is insoluble (Hutto 2017; Hutto & Myin 2012; Kemp 2012; Kripke 1982; Rosenberg 2015). As such, content attributions should either be eliminated from scientific theorising or at best understood in terms of a purely instrumental *stance* on human and animal behaviour (Dennett 1978). As Rosenberg (2015, p.544) puts it,

> "Advances in neuroscience will eventually reveal the wrongness of intentional content attributions, as science has revealed the wrongness of color attributions, folk physics, and other adaptively useful fictions."

If the views that I have defended in this thesis are correct, Rosenberg's anti-representationalist prophecy is mistaken. Far from undermining the status of internal representations, promising work in contemporary neuroscience positively vindicates a conception of the neocortex as a specific kind of representational mechanism: a predictive modelling engine. Nevertheless, I have not addressed most mainstream philosophical debates about content determination and naturalistic psychosemantics in this thesis. Why not?

The chief reason is this: in Chapter 2 I argued that the overwhelming focus on questions concerning content determination in the philosophy of mental representation neglects the fact that mental representations are first and foremost structures within the brain that *do* important things. As such, the most fundamental constraint on an account of mental representation is that it should clarify the functional profile that makes mental representations so interesting to the project of cognitive science. Once we have good reason to believe that an internal structure genuinely functions as a representation, ontological worries about content determination become less pressing. As Ramsey (2007, p.30) puts it, "if theorists can develop accounts of cognitive processes that posit representations in a way that reveals just how the posit plays a representational role in the system, then the explanation of content can wait."

I have argued that the hierarchically structured generative models posited by theories such as predictive processing genuinely function—unsurprisingly—as *models*: as idealised structural surrogates for target domains that are exploited by the brain for prediction and prediction-based cognitive functions. I think that this is sufficient to certify their representational status—to certify *that* they are representations—without wading into the messy controversies concerning the topic of content determination (Williams 2018b; 2018d).

In this respect, Ramsey's appeal to "waiting" is important. As noted in Chapter 2, there is a standard view in contemporary philosophy of mind according to which questions concerning content determination can be pursued largely independently of research in the cognitive sciences. As such, it is reasonable to expect a full-blown theory of content determination *now*. Insofar as we do not have one, philosophers thus take this to license scepticism about mental representations (Hutto & Myin 2012; Hutto 2017). I regard this as a mistake. As I argued in Chapter 2, insofar as there is a project of integrating content into a scientific metaphysics, it should primarily be a job for *science*, not metaphysics. Although there is a useful role for philosophers to play in this project, this role should consist of carefully clarifying and scrutinising the implications of viable theories in the cognitive sciences, not holding those theories answerable to vindicating the truth of our folk psychological and semantic ascriptions. As such, there is no reason to think that questions concerning content determination can be meaningfully addressed in independence of the details of ongoing scientific research.

In this light, the problem with assuming that all questions concerning content determination should be settled *now* is that we do not currently have anything like a detailed scientific understanding of how the brain works and the nature of the mechanisms—some of them non-

neural (Clark 2008)—that underlie intelligence. In previous chapters I have outlined a highly schematic theory of mental representation founded on generative model-based prediction, and I have drawn on an ambitious theory in contemporary neuroscience that offers some support for this theory. Nevertheless, this work leaves numerous questions relevant to the topic of content determination unanswered. For example, exactly what is represented in different cortical areas? How determinate are such contents? How exactly should the psychological capacities that draw on internal generative models be individuated and characterised? To what extent are contents shared between agents with different learning histories? Perhaps most importantly, in what ways do the structured sociolinguistic practices of communication and evaluation characteristic of human life augment or transform the representational capacities and properties of our internal models? Questions concerning content determination are ultimately dependent on highly complex and fundamentally empirical questions like these. As such, we should reject a prominent strain of thought in contemporary philosophy of mind according to which all fundamental issues concerning representational content can and should be adjudicated independently of such scientific developments.

Nevertheless, it is also true that a kind of proto-theory of content determination has emerged from previous chapters—one in which content is ultimately grounded in the relation of structural similarity between generative models and their targets. In Section 4 I will attempt to flesh this story out somewhat, focusing especially on the vexed issue of misrepresentation and those characteristics of theories such as predictive processing that should genuinely change how we view this complex theoretical landscape.

First, however, it will be instructive to address a different issue: insofar as questions concerning content determination in contemporary philosophy have been pursued in large part within the context of an understanding of mental representation motivated by folk psychology, how does the theory of mental representation that I have outlined and defended in previous chapters diverge from this understanding?

(10.)**2. The Fodorian Model**

Much of the philosophical work on questions concerning content determination in recent decades has taken place within the context of what I will call the “Fodorian model” of mental representation after the unparalleled influenced of Jerry Fodor (1975; 1983; 1987) in its formation, although it extends far beyond Fodor’s own work (see, e.g., Rey 1997; Schneider

2011; Sterelny 1990).[25] The Fodorian model encompasses three ideas (see Cummins 2010; O'Brien & Opie 2015).

The first is a staunch realism about that aspect of folk psychology involved in attributing intentional states (beliefs, desires, intentions, and so on), combined with the orthodox philosophical interpretation of such intentional states as *propositional attitudes*—as mental states that involve taking up an attitude (e.g. believing, desiring) towards a proposition (e.g. that Paris is the capital of France). Given a widespread commitment to physicalism, this realism about folk psychology implies that propositional attitudes must be realised by states of the brain. The challenge, then, is to find a way of understanding the brain such that its states and operations map on to the mind as described by propositional attitude psychology (see Fodor 1987).

The second idea comes from Davidson's (1967) pioneering work on truth-conditional semantics, drawing on work from Tarski (1956). According to that work, the meanings of declarative sentences are truth conditions, which are constructed from the referential relations that their constituent expressions stand in to individuals and properties in the world. This form of truth-conditional semantics is now the orthodox way of understanding semantics throughout much of contemporary analytic philosophy, at least in the philosophy of mind and language (O'Brien & Opie 2015). Many philosophers were thus led to appropriate its account of propositional content for declarative sentences to provide an account of how our thoughts express their propositional contents: namely, via an internal combinatorial language-like system with syntactic forms amenable to truth-conditional semantics—that is, a *language of thought*, or "Mentalese" (Fodor 1975). Our concepts thus come to be identified with words of this mental language and our thoughts come to be identified with the relations that we stand in to the mental sentences that they compose.

Finally, these two ideas combined with the emergence of classical cognitive science in the late 1950s and 1960s to provide a hypothesis about the brain's information-processing architecture: mental words are formally individuated symbols, the mental sentences that express the contents of our attitudes are molecular composites of such symbols, and thinking is a matter of syntax-sensitive inferential transitions between these molecular composites—"a three-tiered picture of the elements of our cognitive architecture" involving "word-sized concepts, sentence-sized

[25] It is important to note that even Fodor has strong doubts about the truth of the Fodorian model (Fodor 1983; 2008).

intentional states, and argument-sized inferences" (Horst 2016, p.5; see Fodor 1975; Schneider 2011). That is, the first two trends combined with the view that the brain is a rule-governed symbol manipulator—a digital computer.

In this way the Fodorian model amounts to a merger between propositional attitude psychology, truth-conditional semantics, and the foundational commitments of classical cognitive science. Importantly, this constellation of commitments is so widespread in many quarters that it is often simply described as *the* "representational theory of mind" (Fodor 1987; Rey 1997; Sterelny 1990).

Crucially, this merger gives rise to a fundamental challenge. The referential relations that underlie the truth-conditional semantics of public language are presumably to be explained in terms of some combination of conventions and the mental states of language-users. This story obviously cannot work as an explanation of the referential relations that underlie the truth-conditional semantics of our mental sentences, however. What is needed, then, is an account of which natural properties and relations determine the reference mapping from the primitive symbols of Mentalese to individuals and properties in the world (Fodor 1987). The most influential strategy for providing such an account appeals to some species of causal or covariational relation between mental words and their referents:

> "The proponent of LOT has a naturalistic story about the aboutness, or intentionality, of symbols... Symbols refer to, or pick out, entities in the world, in virtue of their standing in a certain causal or nomic relationship that exists between the symbols and property tokens/individuals in the world" (Schneider 2009, p.283).
>
> "Thoughts are sentences in the head. They represent or misrepresent the world, in virtue of causal relations of some kind between chunks of these sentences and individuals and types of individual in the world" (Sterelny 1990, p.x).

(10.)**3. Contrast with the Fodorian Model**

There are many differences that one could highlight between the account of mental representation that I have defended and the Fodorian model. I will highlight just a few salient ones.

Most fundamentally, the account that I have defended is not motivated by an attempt to explain the truth of propositional attitude psychology. Warfield and Stich (1994, p.4), for example, claim that

> "mental representations are theoretical postulates invoked by philosophers and cognitive scientists in an attempt to analyse or explain propositional attitudes, such as belief and desire, *that play a central role in commonsense psychology*" (my emphasis).

This is emphatically not how the account of mental representation that I have outlined and defended should be understood. In Chapter 2 I argued that mental representations can be understood as functional analogues of public representations and in Chapter 3 I drew on Kenneth Craik's speculation that the relevant analogue for understanding mental representation is the model, a structure that capitalizes on relational similarity to representational targets in the service of successful prediction and control. Since then, I have sought to relate this speculation to advances in the contemporary cognitive sciences, where generative and forward models are introduced to account for a range of capacities that propositional attitude psychology appears to be silent on: unsupervised learning, overcoming signalling delays in sensorimotor processing, overcoming the noise and uncertainty in neural processing, filtering redundancies in incoming information, and so on. Nowhere in this research is there the motivation of identifying the truth-makers for propositional attitude ascriptions.

Second, reasoning within the Fodorian model is typically understood in terms of truth-preserving logical inference, often modelled on first-order logic (Fodor 1975). By contrast, representation and inference within the kinds of models that I have drawn upon are probabilistic, explaining how nervous systems become adapted to the noise, uncertainty, and ambiguity of sensory signals (Kersten et al. 2004). Further, given the large number of variables and complex relationships among them implicated in such models, the relevant probability distributions will likely not be "semantically transparent" (Clark 2001). That is, the contents of such generative models are unlikely—for the most part, at least—to map on in any straightforward way to the familiar word-sized concepts that typically feature in public language (Clark 2015).

Third, the representational power of generative models is not grounded fundamentally in causal or informational relations between atomic mental words and worldly items but rather the structural similarity that obtains between their variables and parameters and the patterns of causal and statistical relationships among the bodily and environmental features that constitute their targets. Indeed, as we will see shortly, predictive processing and related views appeal to this holistic structural representation to *explain* the informational relations between nervous systems and worldly features that most incarnations of the Fodorian model take for granted. The upshot, as Clark (2016, p.107) puts it, is that predictive processing affords a "structured

grip" on the world "that consists not in the symbolic encoding of quasi-linguistic "concepts" but in the entangled mass of multiscale probabilistic expectations used to predict the incoming sensory signal."

Perhaps most fundamentally, a theory of mental representation founded on generative models offers a principled account of *learning*. Although most proponents of the Fodorian model reject Fodor's (1975) original claim that the primitive vocabulary items of Mentalese are innate, it has always proven difficult within that framework to explain how it is that we acquire the capacity to represent new phenomena. In principle, at least, generative models overcome this problem, explaining how brains reconfigure their patterns of internal connections to recapitulate the causal-statistical structure of the environment as it can be recovered from the higher-order correlations among elements of the incoming sensory signal. Indeed, unsupervised learning is so central to predictive processing that Kiefer and Hohwy (2017, p.10) argue that it just is "a specific proposal about how unsupervised learning occurs in the brain."

Crucially, there is a sense in which *misrepresentation* is at the very foundation of this process of prediction error-driven unsupervised learning, such that an agent's structured representation of the world ultimately emerges from a computationally tractable means of ridding itself of detectable *error*—a profound shift in the way in which we understand mental representation that I return to below.

These brief remarks only scratch the surface of the differences between the account of mental representation defended here and the Fodorian model. Nevertheless, I have included them to showcase the fact that the formal and empirical developments drawn upon in previous chapters should fundamentally reconfigure the way in which we think about content determination for mental representations. Of course, these departures from the Fodorian model raise the question of how to accommodate propositional attitudes and the attractions of a highly flexible domain-general mental language within a generative model-based understanding of mental representation. I return to this question briefly in the next and final chapter.

In the remainder of this chapter, however, I turn to the question of content determination within probabilistic generative models.

(10.)**4. Content Determination in the Predictive Mind**

As noted above, I do not have a full-blown theory of representational content determination to offer here, and I think that any attempt to provide such a theory at this point would likely be

premature. Nevertheless, I want to sketch in very broad strokes how I think the issue looks from the perspective of mental representation that emerges from previous chapters.

(10.)**4.1. A Simple Example**

Once more, it is useful to creep up on this topic by first focusing on the maximally simple example of a cartographic map. As I noted in Chapter 4, such maps give rise to at least two forms of representational evaluation (Churchland 2012).

First, one can compare the structure of the map with the structure of its target domain. For this, one needs to be able to determine what the target domain *is*. As I have argued, this cannot itself be specified by structural similarity: not only is structural similarity *as such* ubiquitous, but such a proposal would evidently rule out the possibility of misrepresentation. Once we have determined a target, however, a map's accuracy is a function of its structural similarity to that target. For example, the map might indicate that point A is closer to point B than point C. Is this pattern of distance relations replicated among the corresponding features of the terrain? As I have stressed throughout this thesis and will return to below, accuracy so understood is a matter of degree and almost never perfect.

Second, one can evaluate an attempt to index one's current location on the map, i.e. "We are *here*." Importantly, this second dimension of evaluation is mediated through the first. There are two related reasons for this. First, such judgements are only intelligible once one has a broadly accurate map of the target domain in the first place. (Attempting to locate oneself in London using a map of Paris falls under the category of "not even wrong.") Second, *what* one's judgement indicates—its content—is relative to the broader configuration of elements and patterns of relations among them specified by the map. If I index that we are *here* on a given map, you cannot evaluate this claim by looking exclusively at my finger or our current location. Instead, you have to look at the position of the finger *relative to the broader layout of the map*, and compare this to our current location. If I put my finger on Blue Street when in fact we are on Red Street, the reason why my judgement indicates Blue Street is because it has indexed the part of the map that, as it were, "plays the Blue Street *role*"—that occupies that position in the broader context of relations exhibited in the map.

(10.)**4.2. Content in Generative Models**

It will be helpful to keep this maximally simple example in mind when now turning to a case of representation that is many orders of magnitude more complex.

First, just as one can evaluate a map for its two-dimensional similarity to the spatial layout of a terrain, one can evaluate generative models for their similarity to the causal-statistical structure of target domains in the world. In this case, the question is whether the model's variables and parameters map onto the actual patterns of causal and statistical dependencies that obtain between features of the relevant domain. In Chapter 5 I argued that the *target* of such generative models is a function of that part of the environment causally implicated in their production. Most abstractly, insofar as there is a function that maps causes onto effects (e.g. proximal sensory inputs) that the system has access to, the task of a causal generative model is to recreate *that* function (Friston 2005; Gładziejewski 2015; Rao & Ballard 1999). Of course, it would be wrong to think that mental models recapitulate the structure of the world in all its complexity (see Chapter 9), but they can still be evaluated for whether they capture real patterns that structure the body and environment that predictive brains are tasked with controlling.

Second, just as we can index our location on a map, we can do the same with models of the causal-statistical structure of a domain. Think of a simple supply/demand model used in economics. On the one hand, we can evaluate the degree to which it captures the covariational dependencies that obtain between supply, demand, and price for a given product, but we can also use the model to index the current state of the market (e.g. price = x, supply = y, demand = z). Likewise for generative models in the brain. In the case of perceptual inference, at least, the task is to exploit the generative model to estimate the current state of the body and environment—or, as the point is often phrased, to identify the multilevel hypothesis that best *explains* the specific patterns of proximal stimulation received by the brain. The addition of probability in this context can be taken to reflect uncertainty in this estimate.

This in effect seems to be the interpretation of generative model-based representation endorsed by Gładziejewski (2015) and Kiefer and Hohwy (2017): in perceptual inference, we compare the possible world indexed by the distribution of variable-values with the state of the actual world. In the case of *mis*representation, as Gładziejewski (2015, p.577) puts it,

> "a hypothesis has been selected…whose location within the model's structure does not correspond to the location of the cause of the incoming signal within the causal-probabilistic structure of the environment (i.e. the system selects wrong variables as a basis for its sensory predictions)."

As Kiefer and Hohwy (in press) point out, the two forms of representation (and associated kinds of misrepresentation) just outlined relate to each other in the following way: accurate

representation in the second sense—that is, correctly identifying the state of the world actually responsible for generating the relevant patterns of proximal stimulation the organism is exposed do—is dependent on accuracy in the first sense, such that perceptual "inferences will tend to yield accurate results in proportion to the extent to which the model captures the structure of the target domain" (Kiefer & Hohwy in press, p.21).

(10.)**4.3. Maximizing Mutual Information**

Although I think that this view is correct as far as it goes, one might wonder what it really means in "naturalistic" terms to say that a vector of variable-values correctly indexes the state of the world. Here I think it is possible to make a deeper connection to phenomena such as *covariation* and *indication* that have played such a central role in philosophical theories of mental representation (Dretske 1988; Fodor 1987; Millikan 1984). Recall from Chapter 5 that "mutual information" roughly names correlation: the degree to which observation of the values of one variable systematically reduces uncertainty about—predicts—the values of another. As Friston (2002a, p.229) points out, in acquiring a generative model "one wants to find the parameters that maximize the mutual information or statistical dependencies between the dynamics of representations and their causes" (see also Hohwy 2013). That is, one wants to install a generative model in which the states of its variables and the features of the world that they map onto systematically covary with one another in perceptual inference. In this way maximizing the mutual information between internal representations and their causes in effect boils down to covariation: one wants to indicate the presence of a cat in perception if and only if a cat is in fact present.

For this reason, the generative model-based account of representation defended here is able to recover a generation of philosophical work that tries to illuminate representation by appeal to relations such as indication or covariation. From the perspective of this account, however, we can see that most of this work gets things backwards: the complex informational sensitivity that nervous systems bear to environmental contingencies here falls out as an *effect* of successful representation, not an explanation of it (Churchland 2012). Indication does not explain representation on this view; it is explained *by* accurate representation. As such, it is also possible to partially legitimise the widespread practice in cognitive neuroscience of identifying the contents of internal states by appeal to what they systematically covary with (e.g. Friston & Price 2001; O'Reilly and Munakata 2000, p.25): in cases of veridical perceptual

inference, the states of model variables will systematically covary with the state of the world that constitutes their referent.

It is important to be careful here, however.

First, the contents of internal states are not *determined* by what they indicate or covary with. In misrepresentation, for example, a generative model indexes a state of the world that does not in fact obtain, which would be unintelligible if representation is a function of indication. Several theorists (Eliasmith 2000; Usher 2001; Hohwy 2013) try to get around this problem by appealing to what the relevant internal state indicates as averaged across all possible stimulus conditions. This is deeply unsatisfactory, however. First, insofar as an internal state X reliably indicates Y as averaged across all possible stimulus conditions, this is the very fact that needs to be explained. Second, there is an important sense in which misrepresentation is *systematic* within the predictive mind: in a system that relies upon expectations for veridical representation, misrepresentation will occur whenever the unexpected happens—hence the set of reliable illusions in which strong prior expectations are violated (Geisler & Kersten 2002; Rescorla 2013). Finally, we can engage with the contents of our generative models "offline," simulating what would happen under merely possible conditions, such that the states of model variables will not covary with what is currently in the surrounding environment (see Chapter 7). As per the discussion above, then, what makes it the case that an internal state X represents Y is not that X indicates or covaries with Y but that X plays a similar role within the structure of the relevant model to the one that Y plays within the target domain. When this model is sufficiently accurate and when it is used in perceptual inference, X will reliably indicate Y as averaged across all possible stimulus conditions—but that is an *effect* of successful representation, not an explanation of it.

Second, the covariational relations that states of the nervous system bear to bodily and environmental conditions is fraught with layers of complexity if predictive processing and related theories are correct. Not only are there populations of neurons encoding prediction errors in addition to populations encoding representations, but the response profiles of the latter are such that the "responses evoked by bottom-up input should change with the context established by prior expectations from higher-levels of processing" (Friston 2002b, p.21). Further, this situation is rendered even more complex by the role of precision-weighting in modulating the influence of top-down and bottom-up information. The upshot is that "the representational capacity and inherent function of any neuron, neuronal population, or cortical

area in the brain is dynamic and context sensitive" (Friston 2002a, p.247). The *purpose* of this dynamic, highly context-sensitive processing, however, is precisely to maximize the mutual information between internal representations and worldly causes, such that reliable indication is a predictable *effect* of this strategy.

(10.)**4.4. Digging Deeper**

The foregoing remarks offer what might be thought of as a "God's eye view" of this topic. That is, they approach the phenomenon of content and content determination from a perspective that includes the organism, the environment, and the relationship between the two. Misrepresentation is thus understood in a way analogous to how one might look down from the sky on someone using a map of a terrain, and notice—as an external observer—either that the structure of the map fails to adequately capture the structure of the terrain, or that the agent has indexed a location on the map that fails to correspond to her current location.

Of course, misrepresentation so described might be wholly *epiphenomenal* for an agent. One of the central characteristics of predictive processing, however, is the importance that it ascribes to the capacity for the brain to *detect* representational errors—or at least a reliable proxy for those errors. Specifically, there is an important sense in which system-accessible *misrepresentation* is fundamental within predictive processing: rather than starting with a theory of representation and then trying to explain representational error, it starts with a description of a certain kind of detectable error—the mismatch between internally generated predictions and the externally generated sensory signal—and then seeks to explain the emergence of successful representation from that. As we saw in Chapter 6, this is why prediction errors are such a useful signal for unsupervised learning and representational updating: they can be conjured entirely from ingredients to which the system has direct access.

Insofar as other philosophers have sought to explain how systems are in a position to detect representational error, they have typically appealed to failures in *action* (e.g. Anderson & Rosenberg 2008; Bickhard 2004). Specifically, such theorists reason that insofar as successful actions are dependent on accurate representation, failures of the former kind can be used to indicate failures in the latter. Of course, predictive processing can accommodate this connection between successful action and accurate representation, insofar as successful action is dependent on prediction and successful prediction is dependent on accurate representation. Nevertheless, appealing to action failures alone is not theoretically plausible: it offers no explanation of how systems get into the business of representing and acting on the world in the

first place, and action failures do not offer fine-grained information concerning the nature of the misrepresentation over and above the fact that a misrepresentation has likely occur occurred. By focusing on prediction error-based processing, theories such as predictive processing solve both of these problems, whilst preserving the important link between behavioural success and accurate representation.

Of course, prediction error is not an infallible proxy for misrepresentation. Nevertheless, it may be the best signal that an actual cognitive system can access short of jumping outside of its own skin and assuming an impossible God's eye view from which to evaluate the correspondence between its internal models and their target domains.

(10).**5. Conclusion**

In this chapter I have tried to outline in very broad strokes how content determination should be understood within the predictive mind. At the core of this story lies the relation of structural similarity between generative models and target domains and the capacity of the brain to exploit prediction errors to gauge the extent of that similarity. I think that this proto-theory should be sufficient to ward off prematurely pessimistic claims that theories such as predictive processing have no viable way of legitimising their appeal to representational content (e.g. Hutto 2017). Nevertheless, I am also aware that these sketchy remarks constitute at best initial forays into a highly complex theoretical domain that warrants much more investigation in future work.[26]

[26] For constructive work on questions concerning content and content determination in predictive processing, see Kiefer and Hohwy (2017; in press), Hohwy (2013), Gładziejewski (2015), and Williams (2017; 2018b).

Chapter 11. Conclusion

> "…only this internal model of reality—this working model—enables us to predict events which have not yet occurred in the physical world, a process which saves time, expense, and even life" (Craik 1943, p.82).

(11.)**1. Introduction**

In this final chapter I want to briefly summarise the theory of mental representation that emerges from previous chapters and then identify three important areas for future research.

(11.)**2. The Mind as a Predictive Modelling Engine**

At the core of the theory of mental representation defended in this thesis are three insights that I extracted from the work of Kenneth Craik in Chapter 3.

The first insight was that our capacity to mentally represent the world is underpinned by neurally realised models that share an abstract structure with their representational targets. This view is one instantiation of one of the oldest attempts to explain the mind-world relation in philosophy, according to which mental phenomena literally re-present the properties of worldly phenomena. This similarity-based understanding of mental representation was "all but universally abandoned" (Sterelny 1990, p.65) throughout the twentieth century. I have argued that it should be rehabilitated. Following Craik, I have argued that our capacity for mental representation should be understood in terms of neurally realised functional analogues of the kinds of public models that are familiar from everyday life, science, and engineering. In Chapter 4 I sought to clarify the core functional profile of such models: models, I argued, can be understood as idealised structural surrogates for target domains.

If one wants to transpose this functional profile into the brain, one needs an account of how to understand both the concept of a representational target for the brain's models and the kind of structure that they allegedly replicate. In Chapter 5 I sought to provide such accounts: the generic target of the brain's mental models—the structure they are designed to recapitulate—is world's "regularity structure," the highly complex networks of causal and statistical relationships among features of the body and environment involved in generating the organism's sensory information. In subsequent chapters this idea was augmented: their target is that aspect of the causal-statistical structure of the world that bears on the organism's contingent practical interests. Nevertheless, the basic idea has remained: on the view defended here, our core psychological capacities are causally dependent on the accuracy of our mental

models—on their capacity to recapitulate the causal-statistical structure of the body and environment. Relating this schematic framework to work in the contemporary cognitive sciences, I have argued that such models can be understood as multilevel probabilistic generative models: models that recapitulate the causal process by which interactions among features of the world generate the data to which organisms have access—most fundamentally, the patterns of stimulation at their sensory transducers.

The second insight that I extracted from Craik's work is that the core function of such models is not detection, pattern recognition, or abstract logical reasoning but *prediction*. For Craik, prediction is the key that renders intelligence not merely reactive but constructive, facilitating various species of anticipation that underlie the adaptive significance of thought. One of the chief aims of this thesis has been to reveal the ways in which Craik's emphasis on the functional centrality of prediction to intelligence has been both vindicated and deepened by recent work in the contemporary cognitive sciences. To this end, I have drawn on research in neuroscience that models the neocortex as a hierarchically structured prediction machine—an engine of adaptive success where model-based prediction underlies unsupervised learning, perception, motor control, and the kinds of exploratory engagement with counterfactual possibilities central to "offline" cognition.

The third insight that I attributed to Craik is that the predictive modelling engines inside our heads are fundamentally geared to *pragmatic* success. The brain's models, Craik (1943) argued, are answerable to norms of "cheapness, speed, and convenience" (p.52), and should be understood in terms of their "survival-value" (p.60). By meshing a predictive model-based account of the mind-world relation founded on structural similarity within a pragmatic brain, I have tried to show how the resulting account of mental representation can either accommodate or avoid some of the most powerful challenges to representationalist accounts of the mind advanced over recent decades. Our psychological capacities, I argued, are causally dependent on models that are highly selective, sometimes imperfect, and simplified in various ways, as much concerned with the body and environment as it relates and matters to idiosyncratic features of *us*—our contingent interests, responses, morphology, sensory apparatus, and so on—as they are with the world as it exists independently of us.

As I noted in Chapter 1, I hope that the upshot of this theory is a schematic framework for understanding mental representation—a framework fundamentally founded on the practical utility of generative models that recapitulate the organism-relevant causal and statistical

structure of target domains in a form that can harnessed for prediction-based cognitive functions.

(11.)**3. Three Directions for Future Research**

Any exhaustive summary of the many open questions, objections, and challenges that this predictive model-based account of mental representation gives rise to would be impossible in the space remaining. Instead, I want to conclude by briefly identifying what I think are the three most interesting areas for future research in light of the contents of this thesis.

(11.)**4. How Many Models?**

If the mind is a predictive modelling engine, how many models does it build? More generally, insofar as the chief representational structure that I have appealed to in this thesis is the hierarchical generative model, how many such generative models does, say, an adult human command? One can imagine a continuum of possible answers to this question. On one end of that continuum sits the view that we each install just one highly complex but ultimately unified hierarchical generative model of the world. On the other end sits the view that Horst (2016) describes (and advocates) as "cognitive pluralism," the thesis that we each command a large number of distinct, largely autonomous models of different content domains.

Many in the predictive processing literature opt for the first of these extremes, according to which "there is always a *single* hierarchical generative model in play" (Clark 2013, p.203, my emphasis) and "the entire neuronal architecture within the brain becomes *one* forward or generative model with a deep hierarchical architecture" (Allen & Friston 2016, p.2464, my emphasis). Of course, reference to a single model here could just be a loose description of all the information an organism commands about the structure of its world, no matter how this information is organised. Proponents of predictive processing seem to have something much more substantive in mind than that, however. Specifically, I think that the commitment to a single such model can be understood to include at least two related commitments, both of which are problematic.

First, it is widely held in the broader literature that it is meaningful to talk of "the" cortical or inferential hierarchy, such that one can talk *in general* of the lower levels and higher levels of this hierarchy (e.g. Clark 2013; Friston 2005; Hohwy 2013). For example, a popular view among proponents of predictive processing is that what we (folk psychologists) call "perceptual experiences" and "beliefs" really just track ill-defined regions of this hierarchy, with the former

occupying the lower levels and the latter occupying the higher levels (Clark 2013; 2016; Hohwy 2013). As Corlett and Fletcher (2015, pp.96-7) put it,

> "In a system that is arranged hierarchically, we may perhaps *choose* to refer to the inferences at the lower levels as perceptions and the inferences at the higher levels…as beliefs, but we suggest that it is important to consider that similar processing obtains at all levels of the hierarchy…Although it is possible—and sensible—at one level of analysis to distinguish beliefs from perceptions…at another level of analysis—the one that we think is more useful—*no distinction is called for*."

I have argued elsewhere that this widespread commitment to a single representational hierarchy of this kind confronts serious obstacles. For example, it cannot both be the case that beliefs occupy higher (deeper) levels of this hierarchy and that going higher up (deeper into) the hierarchy corresponds to representations of worldly phenomena at greater spatiotemporal scale (as claimed by, e.g., Clark 2012; George & Hawkins 2009; Hohwy 2013), because we can form beliefs and engage in highly abstract thought about small phenomena implicated in fast-moving regularities (Vance 2015; Williams 2018a; 2018c). More generally, conceptual thought seems to be profoundly *domain general*. We can form beliefs and reason about phenomena irrespective of its spatiotemporal scale or abstraction. As such, it is difficult to see how such beliefs could be localised to the specific regions of a hierarchy, at least if the levels of this hierarchy are ordered according to their contents—according to *what* they represent (see Williams 2018c).

A second way of fleshing out the commitment to a single hierarchical generative model is to focus on the lack of what Fodor (1983) calls "informational encapsulation" between the different bodies of information that we command. That is, even though the single overall model might—as Clark (2013, p.203) puts it—support "many levels of processing which distribute the cognitive labor by building distinct "knowledge structures" that specialize in dealing with different features and properties," these distinct knowledge structures are nevertheless in principle informationally relevant to one another, such that they are subsumed within a broader model of the world's overall structure that is substantially unified.

I have argued that this vision of our cognitive architecture is also difficult to accept (Williams forthcoming). Specifically, it seems that we navigate the world with *distinct* generative models of different content domains, which are partially dissociable from one another, represent their targets as operating according to different principles, and can sometimes license contradictory predictions about the behaviour of worldly entities. For example, an influential view in

cognitive psychology is that human commonsense reasoning is structured around intuitive theories of different content domains such as psychology and physics, which "cleave cognition at its conceptual joints," such that each domain is organised by a different "set of entities and abstract principles relating the entities" (Lake et al. 2016, p.9, my emphasis; see also Goodman et al. 2015). More generally, it seems plausible that the nature of reality at the spatiotemporal scale that we must interact with it does not have a particularly unified structure. In the special sciences, for example, we do not use just one model. Rather, we triangulate the world with multiple distinct models of different regularities and structures, some of which actively contradict each other, and some of which are incommensurable (Horst 2016; Weisberg 2013; Williams forthcoming). It would not be surprising if the representational strategies exploited by the predictive mind are similar.

Ultimately, I am not sure what to say about either of these issues. They rest on complex questions concerning modularity, the individuation of models, and the representational capacities of hierarchical generative models that would need to be clarified to a much greater degree of precision in order to address these questions properly. Nevertheless, these considerations do at least suggest that a commitment to a conception of the mind as a predictive modelling engine should not be taken to imply a commitment to a single predictive model, and they point towards important work to be carried out in the future.

(11.)**5. Explanatory Scope**

I noted in Chapter 1 that the claim that the mind is a predictive modelling engine should not be confused for the claim that the mind is *only* a predictive modelling engine. Nevertheless, I am acutely aware that without some more precise account of the limits of what can be achieved within the generative model-based account of mental representation that I have defended here, there is a legitimate worry that the account constitutes something of a moving target: whenever there seems to be some feature of our representational capacities that the account does not capture, I can always point out that it was not intended to capture everything in the first place.

Unfortunately, I do not have any precise account of these limitations. Instead, I will briefly sketch three very broad psychological domains that I have not addressed in this thesis and that pose at least prima facie challenges for the predictive model-based theory of mental representation that I have defended.

First, for the most part I have not addressed distinctive features of human *thought* or what is sometimes called "higher cognition" in this thesis, including our capacities for long-term

planning, symbolic reasoning, and decision making. Specifically, I have only gestured towards two ways of scaling the predictive model-based account of mental representation beyond its obvious home in domains such as perception, imagination, and motor control: first, by appeal to the fact that generative model-based predictive processing is applicable whenever there is some systematic process by which causes give rise to effects, which applies across *multiple* cognitive domains beyond perception and motor control (Goodman & Tenenbaum 2016; Tenenbaum et al. 2011); and second, by appeal to the capacity of certain organisms to harness their generative models *offline* for the purposes of endogenously controlled simulation (Clark 2016; Grush 2004; Pezzulo 2017). Nevertheless, it is a crucial question for future work to what extent these kinds of predictive model-based strategies can scale to capture the complexities of higher cognition.

To take just one example, I have argued elsewhere (Williams 2018c) that multilevel probabilistic graphical models are unable to capture the profoundly compositional abilities that underlie distinctively conceptual thought in human beings (see also Goodman et al. 2015; Russel & Norvig 2010). Like propositional logic, graphical models are restricted to *factored representations*, in which a represented state of the world comprises a vector of variable values (Russell & Norvig 2010, p.58). (By contrast, predicate or first-order logic decomposes representations of world states into representations of individuals, properties, and existential quantifiers). Such expressive limitations have motivated some in Bayesian cognitive psychology (Goodman et al. 2015; Lake et al. 2016) and machine learning (Ghahramani 2015) to abandon graphical models as the format of the mind's generative models in favour of more structured symbolic representational formats. Navigating this highly complex domain is a crucial task if we want to understand the nature of mental representation and the limits of probabilistic generative models.

Second, and relatedly, some of our representational capacities are directed at abstract logical or mathematical phenomena that do not play any role within the causal order: the number *3*, *Bayes' theorem*, mathematical functions, and so on. Whilst it may be that a structural similarity-based understanding of representation is plausible for our representation of such phenomena (see O'Brien & Opie 2004), an understanding of mental representation founded on generative models of the world's causal and statistical structure evidently is not. Of course, the capacity to flexibly direct our attention to highly abstract elements, relations, and operations of this kind might be a uniquely human ability, and thus perhaps not central to the basic form of cortical representation that we appear to share with other mammals. Nevertheless, these capacities

would clearly pose a problem for an exclusively predictive model-based understanding of mental representation.

Finally, I have also not addressed many features of memory in this thesis. Of course, in one sense memory is at the core of theories such as predictive processing: generative models can be viewed as learned information about the structure of the world. Nevertheless, there are features of semantic and episodic memory that I have not touched on—for example, our capacity to remember highly discrete pieces of information such as someone's phone number and recall specific experiences. There is some fascinating work relating predictive processing to such phenomena—especially episodic memory (see, e.g. Henson & Gagnepain 2010)—but they clearly present a challenge for the simple view that mental representations are models of the causal-statistical structure of bodily and environmental domains that must therefore be addressed in more depth in future work.

As noted above, these three cases are by no means exhaustive, and I do not want to claim that they present insuperable problems for the broad framework that I have defended. Nevertheless, they help to showcase some concrete examples of psychological phenomena that I have not addressed in previous chapters.

(11.)**6. The Role of Public Representational Systems**

I have stressed throughout this thesis that much of philosophical orthodoxy concerning representation has emerged from a focus on language. I have argued that this focus is mistaken for the basic case of mental representation, which, if the account that I have developed is along the right lines, is not best understood in terms of the structural units, semantic properties, and forms of reasoning associated with language-like systems of representation.

Nevertheless, it could hardly be denied that natural language and other public representational systems play a central—and plausibly unique—role within human cognition, including the representational capacities of human cognition. How do such public symbol systems and the forms of combination and reasoning they facilitate relate to the account of mental representation defended here?

It is highly tempting to pursue something like the following view. Perhaps the account of mental representation in terms of multilevel probabilistic generative models outlined here captures the basic kind of mental representation in the neocortex (or, more broadly, thalamocortical system) that we share with other mammals. Nevertheless, through some

important set of neural and broader biological changes—for example, the relative enlargement of our neocortex (Rakic 2009), the capacity and motivation for large-scale dynamic social coordination (Heyes 2018), and the facility and inclination for cultural learning and thus cumulative cultural evolution (Henrich 2017)—humans have acquired the unique ability to flexibly exploit the arbitrary symbols of public language and the cumulative bodies of acquired knowledge any such language embodies. As Churchland (2012, p.26) puts it, in this way accounts of mental representation founded on a commitment to "Mentalese" might fail "to acknowledge the extraordinary cognitive *novelty* that the invention of language represents, and the degree to which it has launched humankind on an intellectual trajectory that is impossible for creatures denied the benefits of that innovation."

With such a view in mind, one might view the emergence of *public* symbol systems and the suite of cognitive capacities that facilitate their use as a kind of magic bullet in the current context, providing a way of answering many of the challenges raised above.

For example, perhaps the domain generality of "conceptual thought" in human beings is less a reflection of the basic form of representation inside our head than a reflection of the idiosyncratic contribution of natural language with its arbitrary symbols and accumulated body of cultural knowledge (Horst 2016). Likewise, perhaps the combinatorics of natural language provides a kind of representational format whose very dissimilarity to the generative model-based form of mental representation inside our heads is what confers its unique computational advantages, augmenting and transforming the basic representational capacities that we otherwise share with other mammals (Clark 1996). Finally, perhaps the specific kind of *thinking* identified by propositional attitude psychology is really just a matter of privately talking to ourselves in a natural language we understand—an idea that has enticed philosophers at least since Plato (see Cummins 2010, pp.174-92).

Perhaps. These suggestions are as attractive as they are speculative, and they risk generating more puzzles than they solve. Nevertheless, explaining how the predictive modelling engines that I have described in this thesis are augmented and transformed in the human context of discrete combinatorial symbol systems, richly structured social practices of coordination and evaluation, and cumulative bodies of cultural know-how is a core challenge for future work.[27]

(11.)**7. Conclusion**

[27] For some initial explorations in this area, see Clark (2016, ch.9) and Williams (2018b).

Chapter 3 began with the following quote from Kenneth Craik:

> "So we may try to formulate a hypothesis concerning the nature of thought and reality, see whether any evidence at present available seems to support it, and wait to find whether the evidence gained in future gives any support to it or to slightly modified forms of it" (Craik 1943, p.47).

Although many open questions and challenges remain, I hope that I have done enough in this thesis to show that the years since Craik died have been kind to his "hypothesis on the nature of thought."

References

Adams, R., Shipp, S., & Friston, K. (2012). Predictions not commands: active inference in the motor system. *Brain Structure And Function*, *218*(3), 611-643. doi: 10.1007/s00429-012-0475-5

Adams, R., Stephan, K., Brown, H., Frith, C., & Friston, K. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4.

Akins, K. (1996). Of sensory systems and the "Aboutness" of mental states. *The Journal of Philosophy*, 93(7), 337.

Allen, M., & Friston, K. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482.

Anderson, M. (2003). Embodied Cognition: A field guide. *Artificial Intelligence*, *149*(1), 91-130. http://dx.doi.org/10.1016/s0004-3702(03)00054-7

Anderson, M. (2014). *After phrenology*. Cambridge, MA. The MIT Press.

Anderson, M. (2015). Précis of After Phrenology: Neural Reuse and the Interactive Brain. *Behavioral And Brain Sciences*, *39*. doi: 10.1017/s0140525x15000631

Anderson, M., & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral And Brain Sciences*, *36*(03), 204-205.

Anderson, M., & Chemero, T. (2016). The brain evolved to guide action. In S. Shepherd (Ed.), *The Wiley handbook of evolutionary neuroscience* (pp. 1–20). London: Wiley.

Anderson, M. L., & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior*, *29*, 55–86.

Aristotle, (1984) De Anima and De Sensu, Barnes, S. (Ed.) vol. 1 of *The Complete Works of Aristotle: Revised Oxford translation*, (2 vols). Princeton, Princeton University Press

Ashby, W. R. (1952). *Design for a brain*. London: Chapman and Hall.

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman and Hall

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence, 48*, 57-86.

Barr, M. (2011). *Predictions in the brain*. Oxford: Oxford University Press.

Barrett, L. (2011). *Beyond the brain: How body and environment shape animal and human minds*. London: Princeton University Press.

Barrett, L. F. (2017a). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*.

Barrett, LF. (2017b). *How emotions are made*. New York: Macmillan.

Barrett, L.F., & Simmons, W. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419-429. doi: 10.1038/nrn3950

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., & Friston, K. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695-711.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Gulcehre, C. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *Proceedings Of The National Academy Of Sciences*, *110*(45), 18327-18332.

Bechtel, W. (1998). Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science. *Cognitive Science*, *22*(3), 295-318.

Bechtel, W. (2008). *Mental Mechanisms*. New York: Taylor & Francis Group.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421-441.

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72*(1-2), 173-215.

Bennett, M., & Hacker, P. (2003). *Philosophical foundations of neuroscience*. Malden, MA: Blackwell.

Bermúdez, J. (2010). *Cognitive science*. Cambridge: Cambridge University Press.

Bernard, C. (1878). *Lecons sur les phenomenes de la vie communs aux animaux et aux vegetaux*. Paris: Bailliere.

Bickhard, M. H. (2004). The dynamic emergence of representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representation* (pp. 71–90). Oxford: Elsevier.

Blakemore, S. J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself?. *Neuroreport*, *11*(11), R11-R16.

Block, Ned (1986) "Advertisement for a semantics for psychology," Midwest Studies in Philosophy 10: 615-678

Boden, M. (2006). *Mind as machine: A history of cognitive science, Volume 1 & 2*. Oxford: Oxford University Press.

Boyd, R. N. (1991). Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, *61*, 127-148.

Braddon-Mitchell, D. & Jackson, F. (1996) Philosophy of mind and cognition. London: Blackwell.

Brandom, R. (1994). *Making it explicit*. Cambridge: Harvard University Press.

Brooks, R. (1991). ‘‘Intelligence without representation,’’ Artificial Intelligence 47: 139–159.

Brooks, R. (1999). *Cambrian intelligence*. Cambridge, Mass.: MIT Press.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. Frontiers in Human Neuroscience, 4, 25.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Butz, M., & Kutter, E. (2017). *How the mind comes into being*. Oxford: Oxford University Press.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, *21*(1), 145-182.

Cannon, W.B. (1929). Organization of physiological homeostasis. *Physiological Reviews*, *9*, 399–431.

Chalasani, R., & Principe, J. C. (2013). Deep predictive coding networks. *arXiv preprint arXiv:1301.3541*.

Chalmers, D.J. (2003): 'Consciousness and Its Place in Nature', in S. Stich and F. Warfield (eds.), Blackwell Guide to Philosophy of Mind, Blackwell.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-823.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Chemero, A. (2013). Radical embodied cognitive science. *Review Of General Psychology*, *17*(2), 145-150. doi: 10.1037/a0032923

Chomsky, N. (1980). *Rules and representations*. New York: Columbia Univ. Press.

Chomsky, N. (1983). Mental representations. *Syracuse Scholar (1979-1991)*, *4*(2), 2.

Chomsky, N. (1995). Language and nature. *Mind*, 104(413), 1–61.

Churchland, P. (1986). *Neurophilosophy* (1st ed.). Cambridge: MIT Press.

Churchland, P. (1987). Epistemology in the Age of Neuroscience. *Journal Of Philosophy*, *84*(10), 544-553.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision. *Large-scale neuronal theories of the brain*, 23-60.

Churchland, P.M. (1995). *The engine of reason, the seat of the soul*. Cambridge, MA: MIT Press.

Churchland, P.M. (2012). *Plato's camera*. Cambridge, MA: MIT Press.

Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies*, *6*(11–12), 125–142.

Clark, A. (1996). *Being there*. Cambridge, Mass.: MIT Press.

Clark, A. (1999). An embodied cognitive science?. *Trends In Cognitive Sciences*, *3*(9), 345-351.

Clark, A. (2008). *Supersizing the mind*. Oxford: Oxford University Press.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind*, *121*(483), 753-771.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204.

Clark, A. (2015). Predicting peace: The end of the representation wars—A reply to Michael Madary. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*: 7(R). Frankfurt am Main: MIND Group.

Clark, A. (2016). Surfing uncertainty (1st ed.). Oxford: Oxford University Press.

Clark, A. (2018) "Strange Inversions," in Huebner, B. (ed.) *The Philosophy of Daniel Dennett.* Oxford: Oxford University Press.

Clark, A., & Toribio, J. (1994). Doing without representing?. *Synthese*, *101*(3), 401-431.

Colder, B. (2011). Emulation as an integrating principle for cognition. *Frontiers In Human Neuroscience*, *5*. doi: 10.3389/fnhum.2011.00054

Conant, R., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Corcoran, A., & Hohwy, J. (2018). Allostasis, interoception, and the free energy principle: Feeling our way forward. In M. Tsakiris & H. De Preester (Eds.), *The interoceptive basis of the mind*. Oxford: Oxford University Press.

Corlett, P., & Fletcher, P. (2015). Delusions and prediction error: clarifying the roles of behavioural and brain responses. *Cognitive Neuropsychiatry*, *20*(2), 95-105.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Craik, K. (1948) Theory of the human operator in control systems II. Man as an element in a control system. *British Journal of Psychology: General Section 38* (3): 142-148.

Craik, K. (1966) *The Nature of Psychology*, in Sherwood, S.L (ed.). Cambridge: Cambridge University Press.

Crane, T. (2003). *The mechanical mind*. London: Routledge.

Crane, T. (2009). Is Perception a Propositional Attitude?. *The Philosophical Quarterly*, *59*(236), 452-469.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge: MIT Press.

Cummins, R. (1994) Interpretational Semantics, in S. Stich and T. Warfield (eds.) *Mental Representation: A Reader*. Oxford: Blackwells

Cummins, R. (1996). *Representations, targets and attitudes*. Cambridge, Mass.: MIT Press.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.

Cummins, R. (2010). *The world in the head*. Oxford: Oxford Univ. Press.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge: The MIT Press

Davidson, D. (1967). Truth and meaning. *Synthese*, *17*(1), 304-323. doi: 10.1007/bf00485035

Davidson, D. (1987). Knowing One's Own Mind. *Proceedings And Addresses Of The American Philosophical Association*, *60*(3), 441.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Dawson, M. (2014). Embedded and situated cognition. In Shapiro, L. (Ed.) *The Routledge handbook of embodied cognition*. London: Routledge, pp.77-85

Dayan, P. (1997). Recognition in hierarchical models. In *Foundations of computational mathematics* (pp. 43-62). Springer, Berlin, Heidelberg.

Dayan, P., & Daw, N. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429-453.

Dayan, P., Hinton, G., Neal, R., & Zemel, R. (1995). The Helmholtz Machine. *Neural Computation*, *7*(5), 889-904.

de Lange, F., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception?. *Trends In Cognitive Sciences*.

Dennett, D. (1978). *Brainstorms*. Cambridge, Mass: MIT Press.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (1991). Real Patterns. *The Journal Of Philosophy*, *88*(1), 27.

Dennett, D. (1995). *Darwin's dangerous idea*. New York: Simon & Schuster.

Dennett, D. C. (2013). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences*, *36*(3), 209-210.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. London: WW Norton & Company.

Dewey, J. (1925). *Experience and nature*. London: Open Court Publishing Company.

Dewey, J. (1948). *Reconstruction in philosophy* (2nd ed.). Boston, Massachusetts: Beacon.

Dietrich, E. (2007). Representation. In P. Thagard (Ed.), Philosophy of psychology and cognitive science (1st ed., pp. 1–31). Oxford: North-Holland Publications.

Douglas, R., & Martin, K. (2007). Mapping the Matrix: The Ways of Neocortex. *Neuron*, *56*(2), 226-238.

Downey, A. (2017). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*. doi: 10.1007/s11229-017-1442-8

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge: MIT Press.

Dretske, F. (1988). *Explaining behavior*. Cambridge: MIT Press.

Dreyfus, H. L. (2002). Intelligence without representation–Merleau-Ponty's critique of mental representation The relevance of phenomenology to scientific explanation. *Phenomenology and the cognitive sciences*, *1*(4), 367-383.

Egan, F. (2013). How to think about mental content. *Philosophical Studies*, 170(1), 115–135.

Eliasmith, C. (2000) *How neurons mean: A neurocomputational theory of representational content*. Ph.D. dissertation, Dept. of Philosophy, Washington University in St. Louis.

Engel, A., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends In Cognitive Sciences*, *17*(5), 202-209.

Engel, A., Friston, K., & Kragic, D. (2015). *The pragmatic turn*. Cambridge: MIT Press.

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Feldman, H., & Friston, K. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers In Human Neuroscience*, *4*.

Feldman, J. (2013). Tuning Your Priors to the World. *Topics In Cognitive Science*, *5*(1), 13-34.

Feldman, J. (2015). Bayesian inference and "truth": a comment on Hoffman, Singh, and Prakash. *Psychonomic Bulletin & Review*, *22*(6), 1523-1525.

Felleman, D. J. & Van Essen, D. C. (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Fodor, J. (1975). *The language of thought*. New York: Thomas Y. Crowell.

Fodor, J. (1983). *The modularity of mind*. Cambridge: The MIT Press.

Fodor, J. (1985). Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. *Mind*, *XCIV*(373), 76-100. doi: 10.1093/mind/xciv.373.76

Fodor, J. (1987). Psychosemantics. Cambridge, MA: MIT Press.

Fodor, J. (2008). *LOT2: The language of thought revisited*. New York: Oxford University Press

Fodor, J. A. (1984). “Semantics, Wisconsin Style.” Synthese 59: 231-250

Franklin, D., & Wolpert, D. (2011). Computational Mechanisms of Sensorimotor Control. *Neuron*, *72*(3), 425-442.

Frigg, R., Hartmann, S. (2018). Models in Science. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/models-science/>.

Friston, K. (2002a). Beyond Phrenology: What Can Neuroimaging Tell Us About Distributed Circuitry?. *Annual Review Of Neuroscience*, *25*(1), 221-250.

Friston, K. (2002b). Functional integration and inference in the brain. *Progress in neurobiology*, *68*(2), 113-143.

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1456), 815–836.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *360*(1456), 815-836.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4(11), e1000211.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127-138.

Friston, K., & Buzsáki, G. (2016). The Functional Anatomy of Time: What and When in the Brain. *Trends In Cognitive Sciences*, *20*(7), 500-511.

Friston, K., & Price, C. (2001). Dynamic representations and generative models of brain function. *Brain Research Bulletin*, *54*(3), 275-285.

Friston, K., Mattout, J. & Kilner, J. (2011) Action understanding and active inference. *Biol. Cyber.*,*104*, 137-160.

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics*, *102*(3), 227-260.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*(1), 1-49.

Frith, C. (2007). *Making Up The Mind: How the Brain Creates our Mental World*. Chichester: Wiley-Blackwell.

Gadsby, S., & Williams, D. (2018). Action, affordances, and anorexia: body representation and basic cognition. *Synthese*.

Gallagher, S. (2017). *Enactivist interventions*. Oxford: Oxford University Press.

Gallistel, R. (2001). "Mental Representation, Psychology of" in P. Baltes and N. Smelser (eds.) The International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier, pp. 9691-9695

Geisler, W., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508-510.

George, D. (2008). *How the brain might work: A hierarchical and temporal model for learning and recognition*. Stanford: Stanford University.

George, D., & Hawkins, J. (2009). Towards a Mathematical Theory of Cortical Micro-circuits. *Plos Computational Biology*, *5*(10), e1000532.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452-459.

Gibson, J. J. (1966) The senses considered as perceptual systems. Boston: Houghton Mifflin.

Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Giere, R. (2004). How models are used to represent reality. Philosophy of Science, 71(5), 742–752.

Giere, R. (2006). *Scientific perspectivism*. Chicago: Univ. of Chicago Press.

Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, 193(2), 559–582.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy*,32(3), 337-355

Godfrey-Smith P (1996) *Complexity and the function of mind in nature*. Cambridge University Press, Cambridge

Godfrey-Smith, P. (2003) *Theory and Reality*. London: The University of Chicago Press.

Godfrey-Smith, P. (2006a) Mental Representation, Naturalism and Teleosemantics. In G. Macdonald and D. Papineau (eds.) *New Philosophical Essays in Teleosemantics*. Oxford, Oxford University Press.

Godfrey-Smith, P. (2006b). The strategy of model-based science. *Biology & Philosophy*, *21*(5), 725-740.

Godfrey-Smith, P. (2009). Representationalism reconsidered, in D. Murphy and M. Bishop (eds), *Stich and His Critics*, Wiley-Blackwell, pp. 30–45

Godfrey-Smith, P. (2018) "Towers and Trees in Cognitive Evolution," in B. Huebner (ed.), *The Philosophy of Daniel Dennett*. Oxford: Oxford University Press, pp. 225-249.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. London: MIT Press.

Goodman, N. (1969). *Languages of art*. London: Oxford University Press

Goodman, N. Tenenbaum, J.B. (2016). *Probabilistic Models of Cognition* (2nd ed.). Retrieved 2018-7-28 from https://probmods.org/

Goodman, N., Tenenbaum, J., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In Margolis & Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, *111*(1), 3.

Gordon, N., Koenig-Robert, R., Tsuchiya, N., van Boxtel, J., & Hohwy, J. (2017). Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *Elife, 6*.

Gregory, R. (1980). Perceptions as Hypotheses. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *290*(1038), 181-197.

Gregory, R. (1983). Forty Years On: Kenneth Craik's The Nature of Explanation (1943). *Perception*, *12*(3), 233-237.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends In Cognitive Sciences*, *14*(8), 357-364.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, *10*(1), 5-23.

Grush, R. (2003). In defense of some Cartesian assumptions concerning the brain and its operation. *Biology and Philosophy*, *18*(1), 53-93.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(03), 377–396.

Ha, D., & Schmidhuber, J. (2018). World Models. *arXiv preprint arXiv:1803.10122*.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology*, 5.

Hardin, C. (1988). *Color for philosophers*. Indianapolis, Ind: Hackett.

Haugeland, J. (1987). An overview of the frame problem. In Z. W. Pylyshyn (Ed.), The Robot's dilemma: The frame problem in artificial intelligence (pp. 77–93). Norwood, NJ: Ablex.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Hillsdale, NJ: Lawrence Erlbaum.

Hawkins, J., & Blakeslee, S. (2004). *On intelligence*. New York: Henry Holt and Company

Helmholtz, H. von. (1867). *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Henaff, M., Weston, J. Szlam, A. Bordes, A. LeCun. Y. (2016) Tracking the world state with recurrent entity networks. arXiv:1612.03969.

Henrich, J. (2017). *The secret of our success*. London: Princeton University Press.

Henson, R., Gagnepain, R. (2010. Predictive, interactive multiple memory systems. *Hippocampus*, 20(11), pp.1315-1326.

Heyes, C. (2018). *Cognitive gadgets*. London: Harvard University Press.

Hinton, G. (2010). Learning to represent visual input. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *365*(1537), 177-184.

Hinton, G. E. (2005). What kind of graphical model is the brain?. In *IJCAI* (Vol. 5, pp. 1765-1775).

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, *11*(10), 428-434.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in brain research*, *165*, 535-547.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The" wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*(5214), 1158-1161.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic bulletin & review*, *22*(6), 1480-1506.

Hohwy, J. (2013). The predictive mind. Oxford: Oxford University Press

Hohwy, J. (2014). The self-evidencing brain. Noûs, 50(2), 259–285.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness And Cognition*, *47*, 75-85.

Horst, S. (2016). *Cognitive pluralism*. Cambridge: MIT Press.

Huang, Y., & Rao, R. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580-593. doi: 10.1002/wcs.142

Hume, D. (1751/1998). *An enquiry concerning the principles of morals* (Tom L. Beauchamp, Ed.). Oxford, UK: Oxford University Press.

Hutto, D. (2017). Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese*.

Hutto, D., & Myin, E. (2012). *Radicalizing enactivism*. Cambridge, MA: MIT Press.

Hutto, D., & Satne, G. (2015). The Natural Origins of Content. *Philosophia*, *43*(3), 521-536. doi: 10.1007/s11406-015-9644-0

Isaac, A. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy*, *91*(4), 683–704.

Jacobs, R., & Kruschke, J. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8-21. http://dx.doi.org/10.1002/wcs.80

James, W. (1890). *The Principles of Psychology*.

James, W. (2000*). Pragmatism and other writings*. London: Penguin Books.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, 14(103), 109

Jebara, T. (2002). "Discriminative, Generative, and Imitative Learning."

Johnson-Laird, P. (1983). *Mental models*. Cambridge, MA: Harvard University Press

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *370*(1668), 20140169-20140169. doi: 10.1098/rstb.2014.0169

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., & Hudspeth, A. (2013). *Principles of Neural Science* (5th ed.). London: The McGraw-Hill Companies.

Kemp, G. (2012). *Quine versus Davidson*. Oxford: Oxford University Press.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, *13*(2), 150-158.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review Of Psychology*, *55*(1), 271-304. doi: 10.1146/annurev.psych.55.090902.142005

Kiebel, S. J., Garrido, M. I., & Friston, K. J. (2009). Perception and hierarchical dynamics. Frontiers in Neuroinformatics, 3(20), 1–9.

Kiefer, A., & Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*.

Kiefer, A., & Hohwy, J. (in press). Representation in the Prediction Error Minimization Framework. Routledge Handbook to the Philosophy of Psychology. (Eds.) J. Symons, P. Calvo, & S. Robins. Routledge.

Kornblith, H. (1993). *Inductive Inference and Its Natural Ground*. Cambridge, Mass.: MIT Press.

Kriegel, U. (2008). Real Narrow Content. *Mind & Language, 23*(3), 304-328. doi: 10.1111/j.1468-0017.2008.00345.x

Kripke, S. (1982). *Wittgenstein on rules and private language*. Oxford: Blackwell.

Kurzweil, R. (2012). *How to create a mind*. London: Viking.

Kveraga, K., Ghuman, A., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain And Cognition*, *65*(2), 145-168. doi: 10.1016/j.bandc.2007.06.007

Ladyman, J., Collier, J., Ross, D., & Spurrett, D. (2014). *Every thing must go*. Oxford: Oxford University Press.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2016). Building machines that learn and think like people. *Behavioral And Brain Sciences*, *40*. doi: 10.1017/s0140525x16001837

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal Of The Optical Society Of America A*, *20*(7), 1434. http://dx.doi.org/10.1364/josaa.20.001434

Llinás, R. R. (2001). *I of the vortex*. Cambridge, MA: MIT Press.

Lotter, W., Kreiman, G., & Cox, D. (2015). Unsupervised learning of visual structure using predictive generative networks. *arXiv preprint arXiv:1511.06380*.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Malafouris, L. (2013). *How things shape the mind*. London: MIT Press.

Marcus, G. (2004). *The Birth of the Mind*. New York: Basic Books.

Marcus, G., Marblestone, A., & Dean, T. (2014). The atoms of neural computation. *Science*, *346*(6209), 551-552. http://dx.doi.org/10.1126/science.1261661

Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman & Co.

Mathys, C., Lomakina, E., Daunizeau, J., Iglesias, S., Brodersen, K., Friston, K., & Stephan, K. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers In Human Neuroscience*, *8*. http://dx.doi.org/10.3389/fnhum.2014.00825

McDowell, J. (1996). London: Harvard University Press.

McGinn, C. (1991). *Mental content*. Oxford: Blackwell.

Michaels, C. F., & Palatinus, Z. (2014). A ten commandments for ecological psychology, in Shapiro, L. (ed.) *The Routledge Handbook of Embodied Cognition*. London: Routledge.

Michaels, C., & Carello, C. (1981). *Direct perception*. Englewood Cliffs, N.J: Prentice-Hall.

Miller, M., & Clark, A. (2017). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, *195*(6), 2559-2575. doi: 10.1007/s11229-017-1399-7

Millikan, R. (1984). Language, thought and other biological categories. Cambridge, MA: MIT Press.

Millikan, R. (1999). Historical kinds and the "special sciences". *Philosophical Studies*, *95(1/2)*, 45-65.

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B*, *364*, 1273–1280.

Mountcastle, V. (1978). "An organizing principle for cerebral function: the unit model and the distributed system," in *The Mindful Brain*, eds G. Edelman and V. Mountcastle (Cambridge, MA: MIT Press), 7–50.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, *66*(3), 241-251.

Mumford, D. (2002). Pattern theory: the mathematics of perception. *arXiv preprint math/0212400*.

Nair, V., Susskind, J., & Hinton, G. E. (2008) Analysis-by-synthesis by learning to invert generative black boxes. In *International Conference on Artificial Neural Networks* (pp. 971-981). Springer, Berlin, Heidelberg.

Nersessian N (1999) Model-based reasoning in conceptual change. In: Magani L, Nersessian N, Thagard P (eds) Model-based reasoning in scientific discovery. Kluwer/Plenum, New York, pp 5–22

Newell, A. (1980). Physical Symbol Systems*. *Cognitive Science*, *4*(2), 135-183. doi: 10.1207/s15516709cog0402_2

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

O'Brien, G. (2015). How does mind matter?—Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.), Open MIND: 28(T). Frankfurt am Main: MIND Group. doi:10.15502/ 9783958570146.

O'Brien, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind*. Amsterdam: Elsevier.

O'Brien, G., & Opie, J. (2008). The role of representation in computation. *Cognitive Processing*, *10*(1), 53-62. doi: 10.1007/s10339-008-0227-x

O'Brien, G., & Opie, J. (2010). Representation in analog computation. In A. Newen, A. Bartels, & E. Jung (Eds.), Knowledge and representation. Stanford, CA: CSLI.

O'Brien, G., & Opie, J. (2015). Intentionality lite or analog content? *Philosophia*, 43(3), 723–729.

O'Regan, K. J., & Degenaar, J. (2014). Predictive processing, perceptual presence, and sensorimotor theory. *Cognitive Neuroscience, 5*, 130–131.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a cognitive map*. Oxford: Clarendon Press.

O'Regan, J., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral And Brain Sciences*, *24*(05), 939-973. doi: 10.1017/s0140525x01000115

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, Mass.: MIT Press.

O'Reilly, Randall C., Dean R. Wyatte, and John Rohrlich. "Deep Predictive Learning: A Comprehensive Model of Three Visual Streams." *arXiv preprint arXiv:1709.04654* (2017).

Orlandi, N. (2014). *The innocent eye*. Oxford: Oxford University Press.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese, 195*(6), 2367-2386.

Palmer, S. (1978) ‘‘Fundamental aspects of cognitive representation,’’ in E. Rosch and E. Lloyd (eds.), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum, pp. 259–303.

Papineau, D. (1984). Representation and Explanation. *Philosophy Of Science, 51*(4), 550-572. doi: 10.1086/289205

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Elsevier Science.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Peirce, C. S. (1931–58). Collected papers of Charles Sanders Peirce (8 vols.), P.Hartshorne, P. Weiss, & A. Burks (Eds.). Cambridge, MA: Harvard University Press.

Penny, W. (2012). Bayesian Models of Brain and Behaviour. *ISRN Biomathematics, 2012*, 1-19. http://dx.doi.org/10.5402/2012/785791

Penny, W. (2012). Bayesian Models of Brain and Behaviour. *ISRN Biomathematics, 2012*, 1-19. doi: 10.5402/2012/785791

Perin, R., Berger, T., & Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proceedings Of The National Academy Of Sciences, 108*(13), 5419-5424.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 20*. Frankfurt am Main: MIND Group.

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends In Cognitive Sciences, 20*(6), 414-424. doi: 10.1016/j.tics.2016.03.013

Pfeifer, R., & Bongard, J. (2007). *How the body shapes the way we think*. Cambridge: Massachusetts.

Pfeifer, R., Iida, F., & Lungarella, M. (2014). Cognition from the bottom up: on biological inspiration, body morphology, and soft materials. *Trends In Cognitive Sciences, 18*(8), 404-413. doi: 10.1016/j.tics.2014.04.004

Pickering, A. (2010). *The cybernetic brain: Sketches of another future.* Chicago, IL: University of Chicago Press.

Pinker, S. (1994). *The language instinct*. London: William Morrow and Company.

Pinker, S. (1997). *How the mind works*. London: W.V.Norton.

Pinker, S. (2005). So How Does the Mind Work?. *Mind And Language, 20*(1), 1-24. http://dx.doi.org/10.1111/j.0268-1064.2005.00274.x

Pitt, D. (2017) "Mental Representation", *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/mental-representation/>.

Price, H. (2011). *Naturalism without mirrors*. Oxford: Oxford University Press.

Prinz, J. (2012). *Beyond human nature*. London: Penguin.

Quartz, S., & Sejnowski, T. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral And Brain Sciences*, *20*(04). http://dx.doi.org/10.1017/s0140525x97001581

Quine, W. (1960). *Word and object*. Cambridge: Technology Press of the Massachusetts Institute of Technology.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews Neuroscience*, *10*(10), 724-735. doi: 10.1038/nrn2719

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas In Psychology*, *40*, 3-12.

Rey, G. (1997). *Contemporary Philosophy of Mind*. Oxford: Blackwell Publishers.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field-effects. Nature Neuroscience, 2(1), 79.

Rescorla, M. (2013). Bayesian perceptual psychology. In M. Matthen (Ed.), *Oxford handbook of the philosophy of perception*. Oxford, NY: Oxford University Press.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind and Language*, 31, 3–36.

Roe, A., Pallas, S., Kwon, Y., & Sur, M. (1992). Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *Journal Of Neuroscience*, *12*(9), 3651-3664.

Rogers, B. (2017) Perception: A very short introduction. Oxford: Oxford University Press.

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press.

Rorty, R. (1989). *Contingency, irony, and solidarity*. Cambridge: Cambridge Univ. Press.

Rosenberg, A. (2015). The genealogy of content or the future of an illusion. Philosophia, 43(3), 537–547.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, Purpose and Teleology. *Philosophy Of Science*, *10*(1), 18-24. doi: 10.1086/286788

Ross, H. E. (1969). When is a weight not illusory? *Quarterly Journal of Experimental Psychology* , *21* , 346 – 355

Rowlands, M. (2015). Arguing about representation. *Synthese*, *194*(11), 4215-4232. doi: 10.1007/s11229-014-0646-4

Rumelhart, D. and McClelland, J. (1986) Parallel Distributed Processing, Vol. 1. Cambridge, MA: MIT Press.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). London: Pearson.

Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind and Language*, *19*(2), 211–240. doi:10.1111/j.1468-0017.2004.00255.x.

Ryder, D. (2009a), Problems of representation I: nature and role, in F. Garzon & J. Symons, eds, `The Routledge Companion to the Philosophy of Psychology', Routledge.

Ryder, D. (2009b), Problems of representation II: naturalising content, in F. Garzon & J. Symons, eds, `*The Routledge Companion to the Philosophy of Psychology*', Routledge

Ryder, D. (forthcoming) *Models in the brain.*

Ryle, G. (1949). *The concept of mind*. London: Routledge.

Schmidhuber, J. (2015) On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. ArXiv preprint, URL https://arxiv.org/ abs/1511.09249.

Schneider, S. (2009). The language of thought. In P. Calvo and J. Symons (eds), Routledge Companion to Philosophy of Psychology. New York: Routledge.

Schneider, S. (2011). *Language of thought*. Cambridge: Mit Press.

Seth, A. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97-118. doi: 10.1080/17588928.2013.877880

Seth, A. K. (2015). The cybernetic bayesian brain—From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.), Open MIND: 35(T). Frankfurt am Main: MIND Group. doi:10.15502/9783958570108.

Seth, A., & Friston, K. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *371*(1708), 20160007. doi: 10.1098/rstb.2016.0007

Shapiro, L. (2010). *Embodied cognition*. New York: Routledge.

Shea, N. (2014). Exploitable isomorphism and structural representation. In *Proceedings of the Aristotelian society* (Hardback) Vol. 114(2pt2), pp. 123–144.

Silvers, S. (1989). *Rerepresentation*. Dordrecht: Kluwer Academic Publishers.

Skinner, B. (1938). *The behavior of organisms*. Acton, Mass.: Copley Pub. Group.

Skinner, B. (1971). *Beyond Freedom and Dignity*. Del Mar (California): Communications Research Machines

Stephan, K. E., Petzschner, F. H., Kasper, L., Bayer, J., Wellstein, K. V., Stefanics, G., ... & Heinzle, J. (2017). Laminar fMRI and computational theories of brain function. *NeuroImage*.

Sterelny, K. (1990). *The representational theory of mind*. Oxford: Blackwell.

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. Cambridge: MIT Press.

Stich, S. (1983). From folk psychology to cognitive science (1st ed.). Cambridge, MA: MIT Press.

Stich, S. (1992). What Is a Theory of Mental Representation?1. *Mind*, *101*(402), 243-261. doi: 10.1093/mind/101.402.243

Stich, S., & Warfield, T. (1994). *Mental representation*. Cambridge, MA: Blackwell.

Suppes, P. (1960). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese*, *12*(2-3), 287-301. doi: 10.1007/bf00485107

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449-508. doi: 10.1007/bf00499820

Tarski, A. (1956). The concept of truth in formalized languages. In Logic, Semantics, and Metamathematics. Oxford: Oxford University Press.

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal Of Vision*, *11*(5), 5-5. doi: 10.1167/11.5.5

Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis* 55: 393–415

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279-1285. doi: 10.1126/science.1192788

Teufel, C., & Nanay, B. (2017). How to (and how not to) think about top-down influences on visual perception. *Consciousness and cognition*, *47*, 17-25.

Thagard, P. (1996). *Mind*. Cambridge: MIT Press.

Thagard, P. (2012). The cognitive science of science: Explanation, discovery, and conceptual change. Cambridge, MA: MIT Press.

Thompson, E. (2010). *Mind in life*. London: Harvard University Press.

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. 55, 189–208.

Tooby, J. & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. In Buss, D. M. (Ed.), *The Handbook of Evolutionary Psychology, Second edition. Volume 1: Foundations*. (pp. 3-87). Hoboken, NJ: John Wiley & Sons.

Truitt, T.D. & Rogers, A.E. (1960) *Basics of Analog Computers*. John F. Rider. Smith GW, Wood RC (1959) Principles of analog computation. McGraw-Hill, New York

Ullman, T., Spelke, E., Battaglia, P., & Tenenbaum, J. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends In Cognitive Sciences*, *21*(9), 649-665. doi: 10.1016/j.tics.2017.05.012

Usher, M., (2001) "A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation," *Mind and Language*, 16: 311–334.

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation?. *Journal Of Philosophy*, *92*(7), 345-381. doi: 10.2307/2941061

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and brain sciences*, *21*(5), 615-628.

Varela, F., Thompson, E., & Rosch, E. (1991). The embodied mind (1st ed.). Cambridge, MA: MIT Press.

Von Eckardt, B. (2012). The representational theory of mind. In K. Frankish, & W. Ramsey (Eds.), *The Cambridge handbook of cognitive science* (1st ed., pp. 29–50). Cambridge: Cambridge University Press.

Von Uexkull, J. (1957). A stroll through the worlds of animals and men. In: Instinctive Behaviour: The Development of a Modern Concept (eds. C. H. Schiller and K. S. Lashley), pp. 5–82. International Universities Press, Madison, Conn.(Original work published in 1934.)

Waskan, J. (2006). *Models and cognition*. Cambridge, Mass.: MIT Press.

Watson, J. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158-177. http://dx.doi.org/10.1037/h0074428

Weilnhammer, V., Sterzer, P., Hesselmann, G., & Schmack, K. (2017). A predictive-coding account of multistable perception. *Journal Of Vision*, *17*(10), 580. http://dx.doi.org/10.1167/17.10.580

Weisberg, M. (2013). *Simulation and similarity*. Oxford: Oxford University Press.

Wheeler, M. (2005). *Reconstructing the cognitive world*. Cambridge, Mass.: MIT.

Wiener, N. (1948) *Cybernetics*. Cambridge, MA. MIT Press.

Williams, D. (2017). Predictive Processing and the Representation Wars. *Minds And Machines*, *28*(1), 141-172.

Williams, D. (2018a). Hierarchical Bayesian models of delusion. *Consciousness And Cognition*, *61*, 129-147.

Williams, D. (2018b). Pragmatism and the predictive mind. *Phenomenology And The Cognitive Sciences*.

Williams, D. (2018c). Predictive coding and thought. *Synthese*.

Williams, D. (2018d). Predictive minds and small-scale models: Kenneth Craik's contribution to cognitive science. *Philosophical Explorations*, *21*(2), 245-263.

Williams, D. (forthcoming) Hierarchical Minds and the Perception-Cognition Distinction.

Williams, D., & Colling, L. (2017). From symbols to icons: the return of resemblance in the cognitive neuroscience revolution. *Synthese*, *195*(5), 1941-1967.

Wilson, A., & Golonka, S. (2013). Embodied Cognition is Not What you Think it is. *Frontiers In Psychology*, *4*. doi: 10.3389/fpsyg.2013.00058

Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin and Review, 9(4), 625–636. doi:10.3758/bf03196322.

Wittgenstein, L., & Anscombe, G. (1953). Philosophical investigations (1st ed.). Oxford: Blackwell.

Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, Oxford

Woodward J (2008) Mental causation and neural mechanisms. In: Hohwy J, Kallestrup J (eds) Being reduced: new essays on reduction, explanation, and causation. Oxford University Press, Oxford, pp. 218–262

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual conference of the cognitive science society* (Vol. 1, No. 2).

Zangwill, O. (1980). Kenneth Craik: The man and his work*. *British Journal Of Psychology*, *71*(1), 1-16. doi: 10.1111/j.2044-8295.1980.tb02723.x

Zhu,Q.,& Bingham, G.P.(2011). Human readiness to throw: The size-weight illusion is not an illusion when picking the best objects to throw. *Evolution and Human Behavior*, *32* (4),288–293