Predictive Processing and Mental Representation

Daniel Calder

PhD Philosophy
The University of Edinburgh
2017

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Signed Date

Daniel Calder

Abstract

According to some (e.g. Friston, 2010) predictive processing (PP) models of cognition have the potential to offer a grand unifying theory of cognition. The framework defines a flexible architecture governed by one simple principle – minimise error. The process of Bayesian inference used to achieve this goal results in an ongoing flow of prediction that both makes sense of perception and unifies it with action.

Such a provocative and appealing theory naturally has caused ripples in philosophical circles, prompting several commentaries (e.g. Hohwy, 2012; Clark, 2016). This thesis tackles one outstanding philosophical problem in relation to PP – the question of mental representation.

In attempting to understand the nature of mental representations in PP systems I touch on several contentious points in philosophy of cognitive science, including the explanatory power of mechanisms vs. dynamics, the internalism vs. externalism debate, and the knotty problem of proper biological function. Exploring these issues enables me to offer a speculative solution to the question of mental representation in PP systems, with further implications for understanding mental representation in a broader context.

The result is a conception of mind that is deeply continuous with life. With an explanation of how normativity emerges in certain classes of self-maintaining systems of which cognitive systems are a subset. We discover the possibility of a harmonious union between mechanics and dynamics necessary for making sense of PP systems, each playing an indispensable role in our understanding of their internal representations.

Lay Summary

Most scientists and philosophers treat the mind as a machine. In this machine ideas, thoughts, and desires are created, stored, and moved around by the brain. When new theories are developed, or new experimental discoveries are made, they are always interpreted in this mechanical way.

The subject of my thesis is an exciting new theory called *predictive processing*. Scientists working on predictive processing see the mind as a prediction machine, constantly generating and testing predictions about the outside world – perception is controlled hallucination, action is an experiment to test our mental picture of the world.

However, there are problems with the underlying mechanical view of the mind. In this thesis, I philosophically examine the nature of thoughts and ideas, what philosophers and scientists call *mental representations*. Mental representations are hard to make sense of using the mechanical picture of the mind. Our thoughts and ideas have meanings, and we don't normally think of the parts of a machine as the sorts of thing that carry meaning in this way. The question is, how can physical components mean anything?

I address this question by looking carefully at the mechanistic approach to predictive processing and its alternatives. By drawing on a broad range of literature, I develop a view that aims to synthesise the mechanical view of the mind with a dynamical view, and in so doing, provide a satisfying, naturalistic, way of understanding representations in a predictive mind.

Acknowledgements

The writing of this thesis has been a long and difficult journey for me, and I am indebted to many people without whom I most likely would not have been able to complete the project.

Though I did not begin the journey with her, my wife Jemma has been here since I met her in my first year, supporting me in too many ways to enumerate. I owe her many days of absence, both mental and physical, too preoccupied with this work to pay her the attention she deserves.

From a distance my family, especially my grandmother, M, and my uncle, Jason have provided many hours of conversation that have, directly and indirectly, helped me work through the many of the knotty issues I tackle here. Nor did they miss a chance to offer some wise words of encouragement, which got me back on track when I needed it most.

Of course, the quality of my work owes much to the diligent and perspicuous oversight of Andy Clark and Mark Sprevak. Andy’s unfathomable depth of knowledge and ability to connect my thoughts to other relevant work, all the while challenging me to hone my arguments in our many contentious conversations, has been the dominant influence in shaping my thoughts and words. Mark has not only lent me his sharp insight on issues of computation and philosophy of science, but also originally suggested this engaging topic to me, for which I am very grateful.

Finally, I must mention those who have indulged me in conversation over the years. Especially Joe Dewhurst and Jonny Lee, who patiently entertained even my most outlandish notions. Also, my pals from 2.16 Anna Ortìn, Nick Rebol, Giada Fratantonio, Rosa Hardt, and Di Yang, your constant support and philosophical input over lunch and coffee has been invaluable.

Contents

List of Figures

For Oliver

0. Introduction

The purpose of this thesis is to answer one simple question: If the brain is a ‘predictive processor’, does it use mental representations?

Predictive processing, also known as ‘active inference’ or ‘hierarchical predictive coding’, is a label for a loose set of computational systems that have been used successfully to model many cognitive phenomena. These models share certain core features in common – they generate predictions about the next input and process error from the last prediction; they learn by constructing a statistical model of their task domain and then use that knowledge to make better predictions; their core aim is to minimise the amount of errors they make in the long run.

Supporters make ambitious claims about the scope of the framework for explaining cognition, some even going so far as to claim that predictive processing is the first paradigm with the potential to yield a grand unifying theory (GUT) of cognition (Hohwy, 2014; Friston, 2010). A core feature of the emergent literature explaining and popularising the predictive processing paradigm is its use of representational language. Cognition is presented as involving inference, prediction and modelling at every level of processing, from simple perceptual procedures up to abstract planning and linguistic tasks.

Given the impressive potential of the predictive processing paradigm, and its novel claims, it is important for scientists and philosophers alike to have a firm handle on the best way to understand the representational language being used. This thesis aims for rigour through scepticism. I will be arguing, in the main, for eliminativist (anti-representationalist) conclusions. However, compromises will be made along the way, and among the end products is an extremely minimal, but robust notion of representation, with a firm naturalistic grounding that I hope will be useful for both scientists of life and mind, and philosophers of mind and cognition.

In this introduction, I will give a broad overview of the debate about mental representation within cognitive science. I will also set out precisely how I intend to tackle the question above, once the dialectical landscape has been established.

0.1 The State of the Art

There are several ways we can think of the mind as representing. The most vivid way is to just look around and see the world represented in your conscious experience. Your visual field is populated with all kinds of objects – desks, chairs, books, windows, trees, clouds and so on. These appear to have size, colour, texture, depth and other qualities. The simplest explanation for this experience is that the mind perceives by representing those objects as having qualities, that is, that there are mental objects that make up conscious experience that reliably relate to the real objects out in the world. This is known as *phenomenal consciousness*.

The next most vivid way the mind represents is in thought. If you conjure up a thought in your mind now, say, "the sun is the nearest star to earth", then you have some experience of that thought. That experience is a conscious experience, similar to the perceptual experiences mentioned above. It is not as vivid, but it is there – you might 'see' the sentence in your 'mind's eye', or 'hear' the sentence (in your mind's ear). This slight difference is captured by the term *access consciousness*.

Both phenomenal consciousness and access consciousness are elements of what Daniel Dennett has called the *personal level*. The personal level is "the explanatory level of people and their sensations and activities" (Dennett, 1969 p. 93). To be sure, there are many other goings on at the personal level – decisions, desires, beliefs – not all of which are necessarily conscious; but Dennett's work reveals also the *subpersonal level*, and there are representations here too.

The subpersonal level is the world of neurons and events in the brain. That sentence that we just called into our conscious experience is stored somewhere. You believe that the sun is the nearest star to earth even if you are not consciously thinking about it all the time. We might suppose then, that somewhere at the subpersonal level that belief has been stored as a memory, and that memory is also a kind of mental representation. Indeed – considering only memory, it seems as if the vast majority of our mental representations live at the subpersonal level. So, even if they are largely invisible to us, these subpersonal representations make up the majority of our mental lives.

But there is much more to the subpersonal level than just memory – all the brain's ongoing control of our bodily processes, crucial to our minute-to-minute survival, lies below the level of our awareness. All the fine control of our actions, once we have made the decision to perform them, occurs at the subpersonal level. There's a lot of thinking going on underneath the personal level, the level we are aware of, and it makes sense that at least some of that thinking might require using representations of the world, just as we use them at the personal level – to make good decisions, to guide our actions, to regulate our biological processes.

Predictive processing seeks to explain cognition, and cognition occurs mostly at the subpersonal level. Having explained these subpersonal processes, we may be able to construct good explanations of our personal level thoughts and perceptions, but for most of this thesis we will be concerned with this low-level guidance, regulation, and decision-making happening at the subpersonal level.

Cognitive scientists ultimately hope to explain cognition is by providing a *naturalistic* account of cognitive processes. That means that any theory of cognition seeks to describe our mental phenomena in purely physical terms. In this way our understanding of the mind can be unified with our understanding of the rest of the natural world. This constraint of naturalism is the source of the most difficult challenge for mental representation – mental representations possess meaning, and meaning is something that is difficult to explain in purely physical terms. So, if predictive processing requires mental representations, and also aims to provide a naturalistic account of cognition, then it requires a naturalistic account of the meaning possessed by mental representations.

The way that cognitive scientists typically work within the constraint of naturalism is by positing a *causal structure* that is capable of bringing about distinctly cognitive behaviours. What they propose is a kind of *mechanism* of cognition. Now mechanisms come in many different forms, but the current paradigm that dominates the science is that the mind is to be understood as a computing mechanism, or *computer* for short. A computer possesses certain components – for instance, one of the simplest kinds of computer is a Turing Machine, which can read and move a tape with a string of 1s and 0s written on it, a state memory, and a finite set of instructions, that tells the machine what to do when it is in a particular state and reading a particular symbol on the tape. Alan Turing proved that such a simple machine is capable of computing any computable function (Turing 1936–7, 135). Essentially, a computer is a mechanism that can process *vehicles* (the symbols on the tape, the input) according to rules that are sensitive only to the properties of those vehicles (Piccinini, 2017). These vehicles, shuffled about by computers are taken to be a kind of representation[1], which is an attractive starting point for naturalising the notion of mental representation – after all, the vehicles are just physical states of a physical mechanism.

Within the world of computational models of the mind, there are two main frameworks that are used by cognitive scientists: classical computing and connectionism. Classical computers work by following a finite list of instructions that specify an algorithm for solving a given problem. This is best visualised and understood as a flow chart. When we know the rules being followed we can work out exactly how the machine will behave given any input, and we can write this out in terms of the states of the computer and the intermediate instructions it executes to move between states. For example, consider the flow chart below for performing multiplication.

[1] By most – there are those who hold that computation is not necessarily representational (e.g. Stich, 1983; Piccinini, 2007, 2015; Milkowski, 2013; Fresco, 2014).

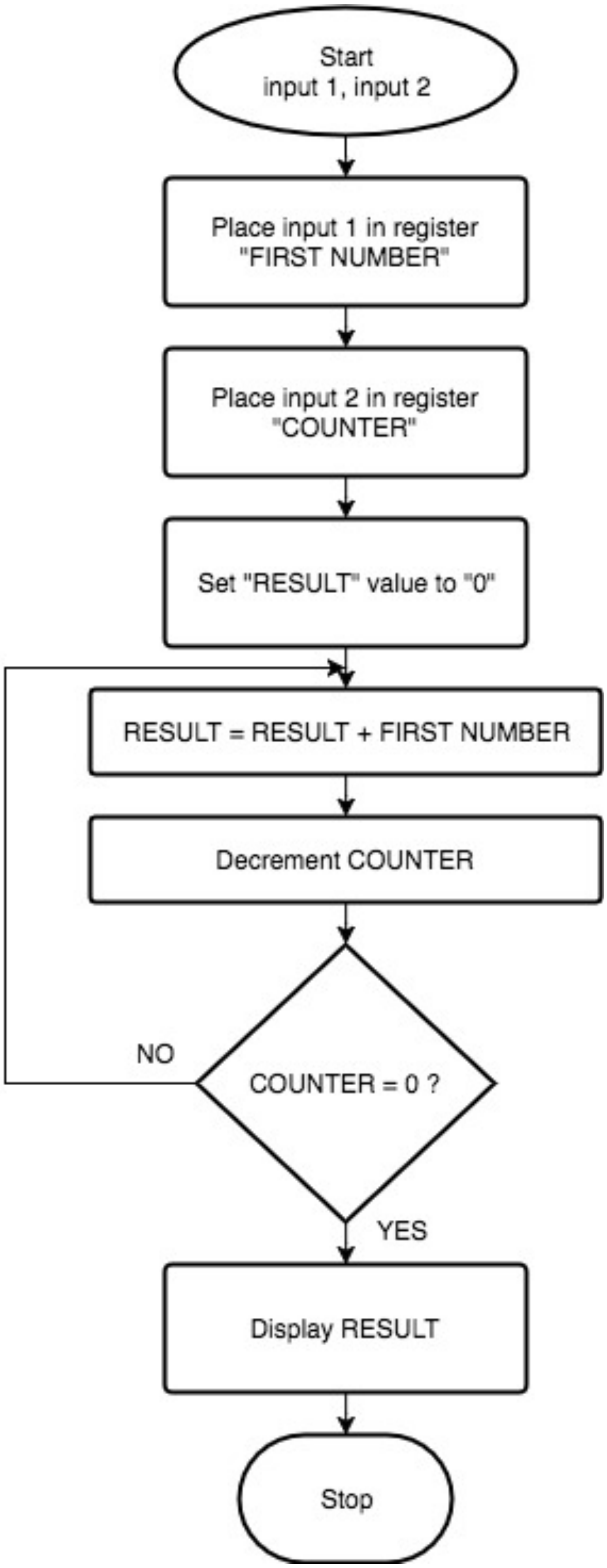
Start
input 1, input 2
Place input 1 in register "FIRST NUMBER"
Place input 2 in register "COUNTER"
Set "RESULT" value to "0"
RESULT = RESULT + FIRST NUMBER
Decrement COUNTER
NO
COUNTER = 0 ?
YES
Display RESULT
Stop

Fig 0.1: Algorithmic flow chart for a multiplying machine.

Two numbers are provided to the computer – the first number is held in a register called ‘first number’, and the second number is held in a register called ‘counter’. A third register is used, called ‘result’ and is set to zero. The algorithm then proceeds to add ‘first number’ to itself, storing the result in ‘result’ and after each iteration subtracts one (decrements) from ‘counter’. When ‘counter’ reaches zero, the computer offers the value of ‘result’ as output. Thus, multiplication is computed using much simpler functions – addition and decrement. The flow chart shows the instructions the machine will follow in any given situation, and by mentally simulating the algorithm we can check that it will indeed multiply any two numbers it is provided with. In a Turing machine, the numbers and instructions will be read by the machine as strings of binary code, made up of 1s and 0s, and these symbols represent the numbers being calculated by the computer.

Connectionist machines are very different from classical computers. They were inspired by the way the brain works. A connectionist computer is made up of many (usually identical) artificial neurons – which are excited by voltages, and when the total voltage received exceeds a certain threshold, the neuron fires – outputting some voltage to the neurons connected downstream. The neurons in connectionist machines are called ‘units’, and the voltages being passed between them are mediated by the ‘weight’ of the connection between the two units. When a neuron fires, the voltage it outputs to its neighbour can be amplified or inhibited by the connection weight, increasing or decreasing the relationship between the two units. Given that the units are homogeneous, the connection weights define the set of instructions – the software – being followed by the system. They define precisely how an input will be transformed into an output by the network. Below is a diagram of a simple connectionist network to help illustrate their structure.

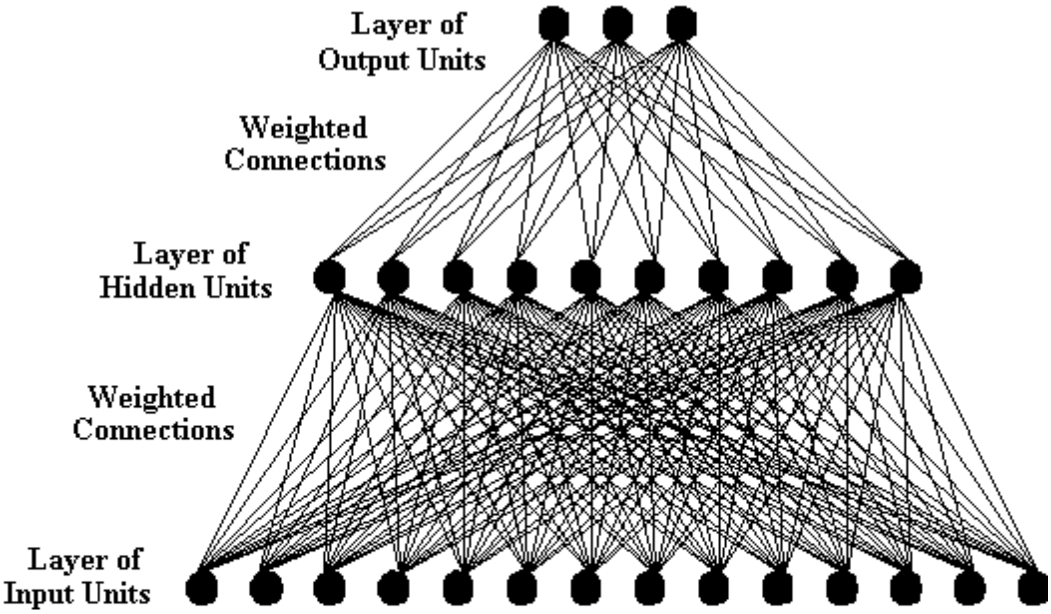

Fig 0.2: Schematic of a three-layer connectionist network.
(Source: http://www.ucs.louisiana.edu/~isb9112/dept/phil341/wisconn.html)

As is illustrated in the figure, connectionist networks are organised in a series of three layers – from bottom to top, the input layer, the hidden units, and the output layer. The input layer is a set of units whose activation is fixed by an input. We can imagine this as a binary string, much like a classical computer, as the units will either be on (1) or off (0). Similarly the output will be a binary string, with a length equal to the number of output units. There is one quick caveat to mention – the information output by each unit is actually more fine-grained than a binary digit; as the output of a connectionist unit is a voltage, the value is actually a continuous variable rather than a discrete value. When interpreting output, we would normally take anything above a threshold value as indicating a '1' and anything below as a '0', but this is not necessary, and it might be that we can identify infinite intermediate values which might be interpreted differently.

The processing done by a connectionist computer is clearly very different to that done in a classical computer. While each process and vehicle of a classical computation is simple, it is supposed to be *interpretable*, for instance as addition or as a decrement. This is not obviously true for connectionist networks which are understood as *subsymbolic* (Smolensky, 1988), as Robert Cummins notes 'connectionists do not assume that the objects of computation are the objects of semantic interpretation' (Cummins, 1989). If representation occurs in connectionist networks then it is at a higher level than that captured by individual units and weights (Hinton et al., 1986; Rumelhart et al., 1986; Smolensky, 1988, 1995). We will consider these ideas more closely in chapter two.

The mechanistic paradigm is not the only way of doing naturalistic cognitive science. Another approach is to instead find the *laws* that govern the behaviour of cognitive systems with respect to their environment. This is the method that has been adopted by *dynamical systems* theorists. Dynamical Systems Theory (DST) encompasses a set of mathematical tools for modelling complex systems. As the brain is the most complex system we have attempted to analyse, it makes sense to attempt to understand it using these tools. However, there is a fierce debate regarding the suitability of DST for the work of cognitive science and we will go into some of these issues in chapter nine. One controversial aspect of the theory is that it does not involve representations, which is precisely what makes it a useful foil for the purposes of this thesis. If predictive processing is a cousin of DST, then it is possible that we can make sense of it without invoking internal representational states.

Dynamical Systems theorists see the brain as a high-dimensional object – one might consider the voltage of each neuron and the weight of each connection between each neuron as one dimension along which the state of the brain might vary. Their aim is to 'provide a low dimensional model that provides a scientifically tractable description of the same qualitative dynamics as is exhibited by the high-dimensional system (the brain)' (van Gelder and Port, 1995 p. 28). In order to do this, dynamical systems theorists characterize systems in terms of their *phase space*. A phase space is a space with as many dimensions as the system possesses – if there are 100 billion neurons in the human brain, the phase space of the brain would possess at least 100 billion dimensions, because there are that many units which may change independently of one another. For instance, one way the dynamicist might reduce this dimensionality is by identifying large clusters of neurons in the brain and taking their average activation value as a single pertinent dimension, thus the dimensionality of the system might be usefully reduced to the number of neuron clusters we identify.

One of the crucial advantages the dynamicist has over the computationalist is sensitivity to the effects of time on the system. Everything in a computer progresses in discrete time-steps of processing which need not reflect the real passage of time in any way. On the other hand time is at the heart of the dynamic picture. DST characterizes the behaviour of a system in terms of its *trajectory* through its phase space. Once we know the laws governing the way the values of each variable of the space will change, we can map the trajectory of the system through that space. We can see this illustrated by a very simple example – a cannonball being fired by a cannon.

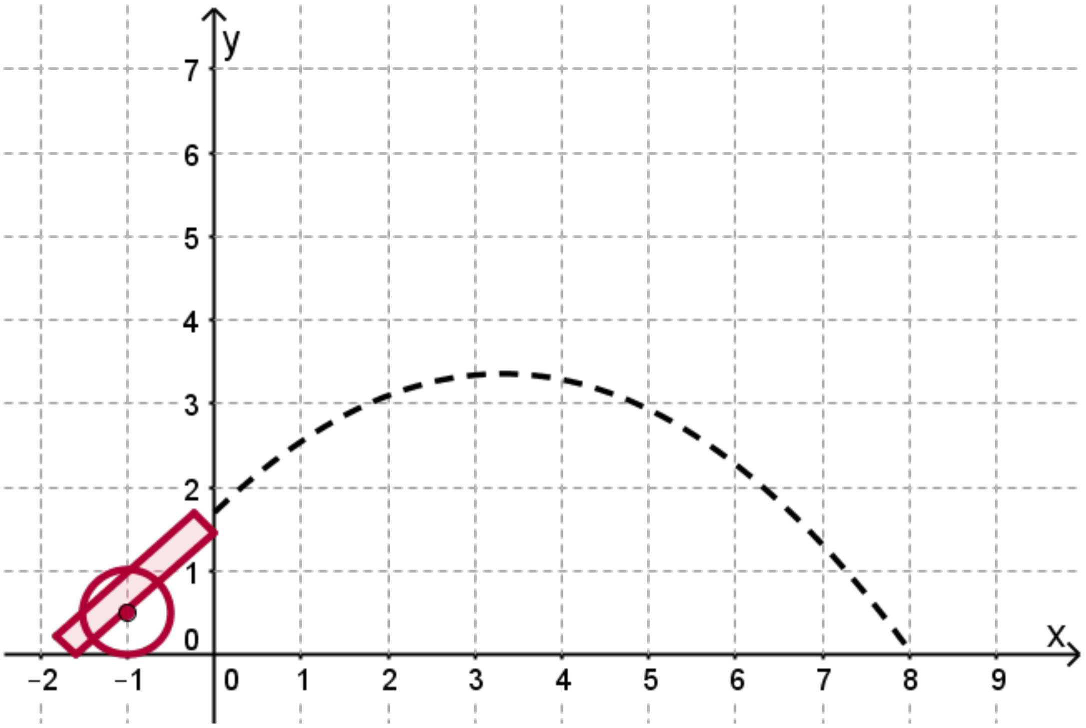

Fig 0.3: Simple state-space of a cannonball in flight, height is shown on the y-axis and distance from the cannon on the x-axis.
(Source: https://www.futurelearn.com/courses/maths-linear-quadratic-relations/0/steps/12125)

Here the cannonball is a system that can vary along two dimensions – its height relative to the ground (y), and its distance from the cannon (x). We then specify the *parameters* of the system (those factors that will influence the system, but are not themselves influenced by the system). Given that the cannon imparts some force, at a given angle, and the cannonball has a given weight, and starts at a given height, we can then, using Newton's dynamics, calculate the trajectory of the cannonball through this phase space and represent it graphically, as in the diagram above.

Along with this basic framework, DST also makes use of concepts such as *coupled systems* – two systems whose dynamics are deeply intertwined such that some change in system A will affect the dynamics of system B, and some change in system B will affect the dynamics of system A. The dynamics of two non-linear, coupled systems each described by one variable can be explained in the simplest form using the following equations:

System A: $\frac{dX}{dt} = f(X, Y)$

System B: $\frac{dY}{dt} = g(X, Y)$

As these demonstrate, the variable (X) that defines system A changes as a function of the variable that defines system B (Y), and vice-versa. This coupling relation does a lot of the work that is supposed to be done by representation in the computational paradigm by showing how subsystems of the brain might interact with each other to regulate and control behaviour (Eliasmith, 1996). We can also define types of phase space in terms of the *phase portrait* of the system, that is, the shape of the space defined by all the possible trajectories of the system within that space. For instance, some phase spaces possess *attractors* – an attractor is a point in phase space towards which the system will tend to return to – imagine a pendulum – however the pendulum might be perturbed, it will eventually come to rest at its lowest point.

These concepts and the mathematical tools behind them (i.e. the differential equations that define the dynamics of the system) have been successfully leveraged by many scientists to make sense of cognitive phenomena from motor coordination to language use and decision making. Excellent overviews of the literature are contained in Port and van Gelder (1995), Kelso (1995), Chemero (2009), Thompson and Varela (2001), Smith and Thelen (2003).

0.2 The Plan

That said, predictive processing is ostensibly a computational account of cognition, so that is where we will begin. In Part One, I will be investigating the foundations of representation within the mechanistic paradigm. In chapter one, I will outline the dilemma for using computational theories of cognition to ground a representational understanding of the mind. On one hand, computation implies representation, and computational theories are the most successful theories we have, so it follows that we should hold that the mind is representational. On the other hand, it's not clear that we can make sense of representational content in a naturalistic way, and if this is impossible, then it challenges the very possibility of a computational theory of cognition. I will sketch the main contenders attempting to solve this dilemma and the problems they face. Chapter two will examine the connectionist notion of representation. Connectionist models purport to make sense of representation in a very different way to classical systems, which I will again sketch. I will then raise the problems offered by William Ramsey (2007) which prompt doubt in whether connectionist networks truly represent. Chapter three will provide a summary of three further positions in the mental representation debate that will inform our discussion moving forward – action-oriented representation, the intentional stance, and the sceptical dynamical stance.

Having thus found our bearings in the mental representation debate, I will then use Part Two to look in detail at the predictive processing framework and some of the high profile theoretical issues which are at stake for the paradigm. Chapter four will provide a detailed summary of the mechanisms common to all PP models, prediction, error, precision, and generative models, in addition to discussion of the key contribution made by hierarchical neural organisation. The most salient philosophical disagreement between advocates of the paradigm is addressed in chapter five – internalism versus externalism – some (Hohwy, 2012) take PP to provide a powerful, modern model for understanding an encapsulated mind, but others (Clark, 2013, 2016) argue that PP provides a robust way of making sense of non-internalist (extended, embodied) research programmes. Chapter six provides a detailed survey of recent neuroscientific work that demonstrates the paradigm's plausibility.

Part Three will take lessons from Part One and Part Two, and attempt to resolve the problems that arise for PP from the mental representation debate. In chapter seven I will look at the usual mechanical-computational way in which PP models are discussed, and argue that the representational posits fail Ramsey's challenge. I then consider the deep problem of mechanical function that is responsible for this failure, how can we understand what it means for a posit to function as a representation if we do not fully understand what it means for a something to *function as* anything at all. Chapter eight moves away from mechanistic explanation, and discusses the promising possibility that the proper interpretation of PP is a dynamic interpretation, and thus suffers none of the problems of the mechanical-computational strategy. Optimism and positivity reach their maximum levels in chapter nine, in which I apply the exciting process ontological framework of Mark Bickhard and Richard Campbell to the problems of function and representation. As the dust settles, we discover a new and elegant way of reinterpreting these scientific projects: one that makes sense of the relationship between computation, dynamics, and representation in living systems, and that is fully consistent with the PP research programme.

Part One: Mental Representation and Cognitive Science

1. The Puzzle of Computation and Representation

The work of logicians in the early 20th century showed how reasoning processes could be formalised into a symbolic system. Simultaneously, the work of mathematicians such as Alan Turing showed how physical computing machines could be made to implement any formal symbolic system (Turing, 1936-7). The product of this marriage of ideas is the classical computational theory of mind (CCTM), the belief that the physical basis of the mind, the brain, is a computer that implements the laws of thought that govern our mental processes, that our concepts are symbolic representations in this machine, which are combined into sentences which can themselves then be processed by the system to make inferences or form attitudes that cause behaviour. Computation and representation go hand in hand – physical symbols possess an internal structure, called *syntax*, which governs how the machine will use them in computation – consider symbols in a pocket calculator – the internal symbol representing "5" will possess some property that defines it as a numeral, whereas the internal symbol representing "+" will have some other property, which picks it out as an operator. This syntax ensures that the machine will use the symbols correctly, so that when the user presses "5+5" and hits the enter key, the display should output the correct answer, "10". By using the right syntax, the system ensures that the link between the internal symbols and what they mean, viz. their *content*, is preserved. The calculator manages to calculate that 5+5=10 because the representations of "5", "+" and "10" stand in the right syntactic relations to one another within the system.

The link between computation and representation is so strong that Jerry Fodor claimed, not unreasonably, that there's "no computation without representation" (Fodor, 1975). Representations are the objects that computations are defined over – it is hard to imagine how to make sense of computation without representation[2]. We can formulate

[2] That said, Gualtiero Piccinini (2008) argues that computation is not defined over representations, but rather over functional states. The result of this debate will not be significant for the discussion in this thesis, which quickly moves on from discussion of classical computing in the following chapters.

this claim as a conditional in the following way: If a system is a computer, then that system possesses internal representations.

Taking this proposition as a given, we might argue two ways. On the one hand we might argue *modus ponens*, that the mind *is* a computer, so it follows that the mind possesses internal (mental) representations. On the other hand we can argue *modus tollens*: the mind does not possess internal representations, so it follows that the mind is not a computer.

In this chapter I will present what I call the puzzle of computation and representation. The puzzle is that there seem to very good reasons to suppose that the mind (well, the brain) is a computer, which would justify the modus ponens form of the argument. But, conversely, the philosophical community has tried and failed to make sense of how the mind could possess representations, which motivates the modus tollens. Having outlined both sides of this puzzle, I will indicate the most promising way forward in the literature, the teleological notion of intentional content, which will guide the discussion in Part Three.

1.1 The Mind is a Computer

In *The Language of Thought* (1975), Fodor argued that the marriage of formal logic and classical computing showed that the mind can be understood as computer with its own system of representation – the eponymous language of thought – which solves old philosophical problems such as how a purely physical system might exhibit rationality and how we as finite beings are capable of having a potentially infinite number of different thoughts (a feature known as *productivity*). Fodor called this the language of thought because the representations of concepts must be organised in a language-like way to take full advantage of the logical formalism that underpins reasoning. For instance, to reason that 'all cats have four legs, so Garfield has four legs' the predicates 'is a cat' and 'has four legs' must stand in appropriate relations to each other, as must the name 'Garfield' pick out some object that lies within the class of cats. The syntax of the representations implementing this process will ground those relations – ‘is a cat’ will have some representation picking it out as a predicate, as will ‘has four legs’, and those will be related in some way that allows the inference of universal instantiation to be carried out on them. ‘Garfield’ will have some internal representation, but that will pick it out not as a predicate, but as an object which related to ‘is a cat’, thus permitting the inference. This is an extremely simplistic example, but Fodor made the point that cognitive scientists had already been assuming something like the Language of Thought as the basis for their computational research, building far more complex and informative models than ones about Garfield’s legs.

A now classic example of the power of classical computationalism is *Vision* (1983) by David Marr. Marr showed how rich visual percepts of the three-dimensional world around us could be built out of imperfect, two-dimensional, noisy, retinal stimulations. Another famous example is the SOAR architecture, a computer that has been used to model a wide variety of cognitive tasks from playing games like tic-tac-toe to diagnosing medical conditions (see Rosenbloom et al., 1991; Laird, 2008, 2012). SOAR uses a working memory to access relevant data from a much larger memory store of knowledge, and then searches the possible space, aiming to satisfy the preferences of the user – e.g. SOAR might search its medical knowledge to find the most preferable treatment for a cancer patient with some specific age, sex, and medical history. Going back into history we might pick out the hugely influential work on cognitive development done by Jean Piaget (1932, 1936, 1945, 1957, 1958) which shows how, beginning with a simple set of concepts (schemas) and operations, a child can show some kinds of intelligence but not others, and that through a process of development can acquire new abilities such as operational (logical) thought, and abstract thought, which have a distinctly computational structure that operates over conceptual schemas.

This diversity and success of the computational paradigm in cognitive science shows that there has been a broad consensus that computation is a good way of explaining cognition. I will not follow Fodor all the way to the conclusion that cognition is thus governed by a Language of Thought which comes bundled with some uncomfortable ideas (e.g. that we must have innate symbols for all concepts from birth, aka *nativism*).

For now, we are just seeking to show that the mind is a computer, so we can activate the modus ponens argument to the conclusion that there are mental representations.

To do this we want to show not just that computation is a good way of explaining cognition, but our *best* way. If this is established we can run an *inference to the best explanation* to the effect that cognitive systems are computers. Minds are a kind of cognitive system so minds must be computers too. Looking back on the 20th century, there doesn't seem to be much work to do here – the computational paradigm has been totally dominant, and although questions have been raised in the latter quarter of the century with the advent of connectionism and by enactivism (more on these later), computation is still going strong as a way of understanding and explaining the mind. Popular today are *Bayesian* theories of cognition, which understand cognitive systems to be performing a specific type of computation – namely Bayesian reasoning (see Griffiths, Kemp, and Tenenbaum, 2008 for an excellent overview). Old sentential logic (and with it the Language of Thought) is replaced by probabilistic inference governed by Bayes' Theorem that explains how we are able to reason effectively in a world full of partial information and uncertainty.

On the other hand, alternative approaches such as DST have been gaining popularity, and offer a genuine alternative (unlike connectionism) to computation. I will be going into more depth on the knotty issue of relative explanatory value of dynamical explanation versus computational explanation in chapter seven, but for now let us simply grant for the sake of argument that computation is the best explanation we currently have for cognition. Just in terms of pure numbers of models, scholars, and scientists working on the assumption that the mind is a computer, this judgement is justifiable for now.

Once it is conceded that computation is the best explanation for the workings of a cognitive system, that is, that treating a cognitive system as a computer will best explain how it is able to do the things it can do, then we are able to reason that we better believe that cognitive systems are computers. After all, following Putnam (1975) if it were the case that cognition were not a kind of computation, then wouldn't the predictive and explanatory success of our computational models be a kind of miracle? That is, it seems very unlikely that our models could be so good, and yet also be radically incorrect. Thus we can infer that the computational theory of cognition is approximately true – to do otherwise would be to seriously entertain the idea that a false theory could get things right so consistently for so long, which is absurd.

So, that cognition is a kind of computation is established by inference to the best explanation, and as the mind is a *cogniser*, it follows that the mind is a kind of computer. From this we invoke our conditional that if a system is a computer then it possesses internal representations, and come to the representational conclusion that the mind possesses internal representations.

A similar argument of this form that may be of interest to us comes from biology. William Bechtel argues that neuroscience also works on the assumption that brain areas represent, and that furthermore, this assumption is used to inform and guide research. If this assumption is false, it is a miracle that neuroscientists have been successful in their predictions. In the language of Imre Lakatos (1978), a commitment to representation is part of the hard core of a progressive research paradigm and supports a strong positive heuristic that continues to generate new predictions which are often vindicated empirically. As there are no comparable alternatives that are not committed to representation, it is fully rational to commit to representation.

Bechtel's story of place cells goes like this. In 1971, O'Keefe and Dostrovsky found eight neurons that "responded solely or maximally when the rat was situated in a particular part of the testing platform facing in a particular direction" (p. 172). In 1976, O'Keefe produced the results of follow-up experiments and introduced the language of place cells, finding 26 cells whose firing rate correlated only with location (and not with orientation). Borrowing theory from Tolman (1948), O'Keefe and Nadel (1978) argued that the hippocampus instantiates an allocentric map of the creature's environment, facilitating the use of novel routes and navigation. In a telling passage from a paper with Conway (O'Keefe and Conway, 1978), O'Keefe clearly presents questions that constitute the positive heuristic of his research programme, and we find them to be concerned primarily with the identification of representational contents:

> “Is [representing a place field] due to something the rat does in the place field or to some environmental factor? If the latter, is the cell responding to a stimulus, or is it signaling more abstract information such as the place itself, as we have previously suggested? How does the cell identify the place? Does it do so on the basis of a special set of cues or will any cue do?” (O’Keefe and Conway, 1978)

Answers to these questions were forthcoming. For instance, in the same paper, O’Keefe and Conway showed that normally two cues were sufficient to prompt activity in place cells (from a set of cues including, “lights, sounds, and feels… and [place cell activity is] not necessarily dependent on distal cues” ibid. p. 589). O’Keefe published further support in 1987 (O’Keefe and Speakman, 1987), showing that place cells remain active during working memory tasks, suggesting that they not only covary with location but encode a persistent spatial map. Developments in the 1990s (e.g. Quirk et al, 1990; Bostock et al., 1991; Markus et al., 1995) explained how place cells could instantiate maps of different locations under different circumstances, and further suggested a mechanism for how that ‘remapping’ occurs. Bechtel continues this history, following developments right through to contemporary studies (e.g. Colgin and Moser, 2010). Thus furnishing advocates of neural representation with a huge body of evidence in support of their inference to the best explanation.

This detour into neuroscience shows that not only is there some good reasons to directly hold that representations are a good explanation of certain cognitive tasks, but also for the computational paradigm which has occupied us thus far. O’Keefe’s study of hippocampal maps fits neatly into a computational understanding of neural processing, further bolstering the computationalist’s claim to possess the best explanation of cognitive activity, not just in the abstract realm of models, but also in concrete biological systems.

So we have a good case for mental representation in virtue of the explanatory power of computational theories and neuroscience. Now let us consider the possible modus tollens, the aspect that makes the relationship between computation and representation a puzzle.

1.2 The Symbol-Grounding Problem

We have heard that computation involves systematic transformations of representations. In this sense computation takes representation for granted – there just must be representations in order for computation to occur. This means that understanding computation tells us precious little about what representation consists in. Just saying that there is a system of symbols at work when cognition occurs doesn't tell us what those symbols *mean*. This is the *symbol-grounding problem*. How are we to make sense of the claim that these physical vehicles in a computing mechanism have meaning? What grounds the meaning in natural terms?

One way of understanding this problem is to consider the distinction between representational *vehicles* and representational *contents*. The vehicle is the physical object that carries the representation – the physical properties of symbols in the computer. Content is the meaning of the representation, what it is standing for. The symbol-grounding problem demands an answer to the question: "how do representations get their contents?"

Those of a pragmatic mind-set might not worry too much about the symbol-grounding problem. After all, it is evident that computation occurs, so therefore there are representations, why do we need to know more than that to make sense of the world? But, what if it were to turn out that there is *no* good way of answering the symbol-grounding problem? Well then we would have to conclude that there is no such thing as internal representation, and computation would be impossible.

From Symbol-Grounding to Intentionality

The absurdity of this conclusion is clear – it is patently obvious that computation is possible – we are surrounded by computing machines: laptops, tablets, mobile phones, even many simple tools have on-board computers to facilitate ‘smart’ functions. However Robert Cummins (1996) provides a more satisfying, philosophical response to make this point. Cummins argues that if a system of symbols works successfully as part of a computer program it forms a structure that represents in virtue of its formal properties which it shares with its content. Sharing these formal properties, this *isomorphism*, is what grounds content in computing systems. The multiplication computer program represents multiplication because it shares a formal structure with the abstract function of addition – that is, it relates its parts (symbols) in the same way that numbers are related in the addition function. In Cummins’ terms, the *content* of a representation is the internal, formal structure of the representation[3]. This allows us to interpret that representation in particular ways to perform useful functions such as addition.

However, content is just one part of Cummins’ theory, because, on its own, content as internal structure doesn’t make sense of the problem of error – an entity can never be in error about its internal structure. What Cummins adds is *targets*. A representation has a particular target – for example, addition. When the content of the representation is not isomorphic with the target then misrepresentation, error, occurs. In the case of ordinary computers, we, the designers and users of software, define the targets of computation through our intended use of the program. For example, when I use a calculator to add numbers together, the target of computation is the addition of the particular numbers I input. But what of systems that have no intended use? What of biological systems?

The problem of meaning for biological systems thus comes down to *target fixation*, how are we to understand the proper targets of biological computations (cognition)? The symbol-grounding problem thus morphs, when we consider natural systems, into

[3] This is what Cummins (1996) terms the*picture theory of representation*. He attempts to avoid the old problems of resemblance theories of representation such as panrepresentationalism by biting the bullet – the content of a representation just is a formal structure that can be shared in by many things. However, Cummins rejects the vehicle/content dichotomy in favour of his understanding in terms of vehicles, content, and target, which allows him to avoid the problems entailed by non-unique contents.

Brentano's problem of intentionality (Brentano, 1874)[4]. How are we to understand the way a mental representation is *about*, or *directed towards* (targeted at), something out in the world? This is the real question that troubles us about mental representation, and that may cause problems for the thesis that minds are computers – that is, if intentional relations are impossible, and representation requires intentional contents, and computation requires representation, then the mind is not a computer. To get started on this problem, let's briefly consider how everyday representation works.

Semiotics provides us with a solid starting point for understanding ordinary representations; Charles Peirce (1958, 1977) developed an account which has since been refined and clarified (von Eckardt, 1993). Peirce distinguishes three kinds of representation: icons, indices, and symbols. These three kinds all share the quality of being signs that are linked to some object(s). They differ in the way that link is realised. Icons are taken to be *isomorphic*, or similar, to their object in some way, a representation such as a portrait or a map would thus be an icon on Peirce's taxonomy. Indices are the most interesting for the cognitive scientist as they connect to their object through *causal or nomic relations*; thermometers being a standard example of indices. They allow for "natural signs" (Grice, 1957; Dretske, 1986; Ramsey, 2007, p. 21), e.g. smoke means fire, rings in the trunk indicating the age of a tree. These natural relationships seem to be a promising avenue for the philosopher committed to naturalism to find a basis for mental representations. Conversely, symbols are linked to their object solely through *convention*, a thoroughly non-natural means of connection.

Aside from taxonomy, Peirce's analysis provides two further philosophically interesting morsels. The first of these is that ordinary representations are constituted by a *three-way relation*, with the sign itself standing in proper relationships with its object and some *interpreter*. The second being his analysis that representations have the *functional role* of *standing in* for an object, for the interpreter. Both of these suggestions are attractive and intuitive. For example, a road sign which indicates the

[4] Cummins himself notes that what he calls targets are what others call intentional contents (1996, p.129).

road is subject to the national speed limit would no longer represent if there were no drivers who understood how to interpret the sign. Furthermore, the sign is a representation in virtue of it functioning as a means to inform drivers about the legal status of their speed on the road.

While our human ability to identify a representation is extremely reliable for the vast majority of cases encountered in day-to-day life, this talent is almost exclusively reliant on our capacity to interpret the symbol being presented as standing for something else, or at the very least to imagine that the symbol might possibly be usefully interpreted in such a way. However, mental representation does require interpretation – if something is being interpreted, then its meaning is being *derived* from the meaning already possessed by something else. What we are looking for is *underived meaning*, that is, underived *content* for our representations, what Daniel Dennett calls 'unmeant meaners' (Dennett, 1987, p. 288). The trick is to squeeze Peirce's three-way relation into a two-way relation that does away with an interpreter.

I will now survey the two main contenders proposed to perform this feat, and thus solve the symbol-grounding problem. The first is the *causal theory of representation*, and the second is the *teleological theory of representation*. We will discover that both have serious deficiencies, and that there is still a genuine problem at hand, and thus an important unsolved puzzle for advocates of the computational theory of mind. However, we will also see that the teleological theory has important advantages over the causal theory, and in Part Three I will spend some time deconstructing the core notions at the heart of the teleological view, and reconstructing them with stronger foundations based in the conceptual framework provided by predictive processing's close cousin, *the free-energy formulation*. For now though, we will make do with the following sketches that will provide us with a handle on the main problems facing theories of representation.

The Causal Theory of Representation

The causal theory of representation is a misnomer – in fact there are a number of theories that seek to solve the problem of intentionality by appeal to causal relations. However, these theories are all similar insofar as they have been developed in response to the problems suffered by the naïve formulation of a causal account of representation. This naïve view is that X represents Y if and only if X *reliably indicates* Y (cf. Crane, 2016). This idea of reliable indication has its roots in the idea of *natural meaning*.

The philosopher of language Paul Grice provided a tempting morsel for theorists attempting to naturalise content. He referred to 'natural meaning', contrasting it with the kind of meaning that linguistic utterances have, 'non-natural meaning' (Grice, 1957). Natural meaning is a kind of covariance or indication relation, briefly mentioned in chapter one. One state of affairs (A) can possess natural meaning about another (B), if A reliably obtains when B obtains, so that having access to A allows one to draw inferences about B. Classic examples of this are wisps of smoke in the wilderness which allow lookouts to respond to forest fires, and the relation that holds between the number of rings in a tree trunk cross-section and the age of that tree[5]. Fred Dretske used this notion to suggest that the content of neural representations can be determined by examining the causal covariance relations between states in the external world and brain states, combined with an analysis of the function of those brain states (Dretske, 1986). For example, if a brain state not only responds to visual stimuli provided by fluffy white sheep, but also to those provided by sheared black sheep, and to auditory "Baa" stimuli, we are justified in supposing that that brain state has the content "sheep". Bechtel (2014) goes into a great deal of depth to demonstrate that an important task in modern cognitive science is to search for representational contents of neural areas by identifying these Dretskean causal covariance relations.

[5] The breadth of tree-trunk rings also reliably indicates growth rates during the year that ring was formed, allowing researchers to use these examples of natural meaning to inform them of climate conditions in the corresponding time-period.

The reason Dretske (1986) invokes function is to avoid the first major problem with the naïve causal theory which grounds representation in terms of reliable indication. This is the *problem of error* and it runs like this: some representations are not accurate, for example, a child's drawing of a dog might show only two legs, or one might believe that the Queen of the United Kingdom is called Fiona; despite being inaccurate, these are nevertheless representations (in the first case, of a dog, and in the second, of the Queen), so *representations must be able to be in error*; but a truly reliable indicator will never be in error, so a reliable indicator is not a representation because it does not have all the properties of a representation. Thus Dretske hypothesizes that the content of a representation is picked out functionally in terms of complex, multi-modal covariance, so that a representation might be triggered by one property of the content (perhaps I smell cheese, so I expect to find a nice slice of cheese on toast in the kitchen) yet nevertheless be in error (instead I find a smelly pair of socks on the floor).

Jerry Fodor also constructs his response to the problem of intentionality in a way that avoids the problem of error, but is also sensitive to the second major objection to the naïve view – *the disjunction problem*. Many natural signs appear to have more than one cause. Crane (2016) highlights an instructive example from cognitive ethology. Work by D. L. Cheney and R. M. Seyfarth into alarm calls made by vervet monkeys hypothesized that particular kinds of alarm have particular meanings, determined by the cause of the alarm call. The disjunction problem is made clear in their discussion of the 'leopard alarm'.

> "The meaning of leopard alarm is, from the monkey's point of view, only as precise as it needs to be. In Amoseli, where leopards hunt vervets but lions and cheetahs do not, leopard alarm could mean, 'big spotted cat that isn't a cheetah' or 'big spotted cat with the shorter legs' … In other areas of Africa, where cheetahs do hunt vervets, leopard alarm could mean 'leopard or cheetah'. (Cheney and Seyfarth, 1990)

The leopard alarm call is evidently a kind of representation – but the (naïve) causal theory of representation allows that the content of the representation is indeterminate. What at first appears to be a representation of leopards is, in different circumstances, actually a representation of 'leopard *or* cheetah'. Applied to our own lives, we might think we have a simple representation ‘face’ that reliably indicates faces, but experiences in the dark shared by us all attest to the fact that ‘face’, on the naïve causal theory, in fact means ‘face *or* cushion in the dark *or* cupboard door in the dark *or* clock in the dark *or…*’ because ‘face’ seems to be set off by all kinds of everyday objects in darkness.

This is a problem for the naïve causal theory because it entails that we have no principled basis for saying that ‘face’ or ‘leopard alarm’ mean what they are supposed to mean, in fact there is a whole mishmash of worldly states that might trigger a given representation, with the supposed content being just one case in a vast disjunction of possibilities. To overcome this requires some addition to the naïve causal theory that shows why the supposed content is special, and the other disjuncts are accidental, and thus examples of error.

Jerry Fodor (1990) has put forward one of the best known responses to the disjunction problem, and the problem of error, in the form of *asymmetric dependency*. Fodor argues that the causes of a given representation are *asymmetric*, in the sense that one is intentional and the others are accidents, because the accidental causes *depend* on the representation’s causal relationship to the intentional cause, which is the real meaning. For example, a clock face in the dark may trigger our ‘face’ representation, but that is only because of the way in which ‘face’ is causally related to real faces – presumably through a sensitivity to roughly round shapes with a certain kind of internal structure. In the dark of course, a face and a clock could look very similar, which explains why ‘face’ might misfire in the presence of clocks. The crucial point is that ‘face’ only represents clocks in the dark *because it represents faces*. The accidental representation of clocks *asymmetrically depends* on its intentional representation of faces.

However, in dealing with the disjunction problem, Fodor’s asymmetric dependency theory drops the ball on the problem of error, indeed Fodor himself acknowledges this (Fodor, 1990, ch. 4). Asymmetric dependence is a purely ontological relationship between representation and intentional content. That is, all the relations appealed to in the theory are real relations – the relation between the representation and intentional content, and the dependence relation between the intentional content and accidental content are simple informational relations grounded in covariance. But famously, ontology can't inform us about normativity – what *is* can't help us with what *ought to be*, and the problem of error requires an account of how a representation is normative – why it is *true*, *correct* or *accurate* in certain circumstances and *false,* or *incorrect* in others.

To solve this problem we return again to Dretske (1986), and his invocation of *function*. Dretske clearly places importance on the causal relationship between representation and content, hence his inclusion in this section, but it is his insight that function is required to really make sense of representational content by overcoming the problem of error that marks him out as a *teleological* theorist.

The Teleological Theory of Representation

Telos, the Greek for 'end', indicates that *teleology* points a way to achieving some end, some goal, some purpose. Nowadays we talk about function. The teleological theory of representation, or *teleosemantics*, is an account of the way representations help achieve some goal – that is, they are involved in some function. Whilst this does mark a transition from the realm of ontology to the realm of normativity, this should not be a source of worry. After all, the computational theory of mind starts from the position of defining cognitive functions, then follows up by proposing some algorithm (the flow chart) that shows how a physical system could perform that function, providing a bridge from the normative to the ontological, thus helping us naturalise our understanding of the function in question. Teleosemantics is the theory that codifies the recognition that just like specifying computations as, say, cognitive, in order to specify the content of representations, we have to understand their function.

Grounding intentional content in function rather than causation has other benefits. Causal theories struggle to account for representations of past and future events, of fictional events and objects, and of counter-factuals. After all, none of these things can possibly stand in a causal relation to the system, only real, present, objects and present events can have causal influence on a system. On the other hand, we can quite happily conceive that a state might function to represent the past, future, fiction, or counterfactual states of affairs – the tricky part is to make sense of how something might function in this way in an underived, naturalistic way.

While we noted that Dretske (1986) made this important insight, it is Ruth Millikan (1984, 1989) who is best known for formulating and defending teleosemantics. Millikan refuses to squeeze Peirce's three-way relation into a two-way relation – maintaining instead that we can have our cake and eat it with a naturalistic three-way relation, between two systems and a representation. One system is the *producer* of the representation, the other is the *consumer* of the representation. Put very simply, Millikan thinks that the meaning of a representation is grounded in role it plays for the consumer system. Her now classic example (Millikan, 2004) is the dance of honeybees discovered by von Frisch (1967). When a honeybee finds a rich source of nectar, they return to the hive and perform a 'waggle dance'. The other bees in the hive watch the dance, which consists in the bee moving in a straight line waggling its abdomen before circling around and repeating the movement. The direction and length of the movement represent the direction and distance of the source of the nectar, thus specify its location relative to the hive. Millikan's theory is that the dance represents the location because that is how it is used by the consumers of the representation – the audience of bees in the hive – who then take off towards the source of the nectar. It is because of this emphasis on the function for the consumer that Millikan's theory is sometimes referred to as *consumer semantics*.

However, it is not obvious what computers have in common with a beehive – Millikan uses talk of producers and consumers in two different senses when discussing internal representations. On one hand, we might divide our computing machinery into subsystems, and have one subsystem produce representations, while another consumes them. But some computing mechanisms do not decompose neatly in this way. In these cases, Millikan talks of the same system acting as producer *and* consumer – at time A, the system produces the representation, and some time later, at time B, the system might then consume the representation. Time A and time B need not be very far apart at all, on this understanding, and in this way we can make sense of the way a computer stores a representation in a register for later use.

There are two extant problems for teleological theories of representation. The first is the disjunction problem, again. The second is the notion of function – it is not clear that it is easier to naturalise this notion than it is to naturalise the notion of intentional content, in which case teleological theories might be overcomplicated answers to the problem of intentionality.

The disjunction problem arises for the teleological theorist in a very similar way as it does for the causal theorist. Let us return to the example of the vervets' leopard alarm. Rather than say that the alarm means 'leopard' because the cry is caused by the presence of a leopard, the teleosemanticist says that the alarm means 'leopard' because its function is to represent leopards, and thus warn the vervet troop. Presumably the alarm has this function because running away from leopards has survival value for the vervets, as organisms are more likely to survive when they avoid predators. But then what is the function of the representation? Does it function to represent 'leopard', or does it function to represent 'predator'? It seems that contents picked out by function are indeterminate, so unless we concede that representational contents are indeterminate, we are left with the disjunction problem, and both options are unattractive (Fodor, 1990).

To avoid the problem of error in natural, biological systems, any old notion of function won't do – just like representation, function is normally imposed on a system by us – we press computing systems to use in specific contexts, designing them to perform functions useful to us. What we are looking for is a naturalised account of *proper function* – the function a biological system is *supposed to have*, whether or not it manages to carry out that function. By far and away the most popular notion is based in evolution. This is *teleofunctionalism* and is endorsed by Millikan herself (1989); it is in virtue of evolution that we can ascribe contents to mental representations.

Teleofunctionalism is an *etiological* theory of function. That is, the teleofunctionalist believes, roughly, that functions are *selected effects* – the effect of a mechanical component that has been selected for by evolution through natural selection. It is the evolutionary history of an organism that defines a component or processes' proper function (Neander, 1991a). For example, the proper function of a heartbeat is to pump blood around the body. It is possible to think of other functions of a heartbeat – nowadays athletes monitor their heart-rate in order to regulate their training. However, these other functions, such as its role in the regulation of fitness training, have not been relevant for its evolutionary success. The selected effect theory of function thus provides us with plenty of scope for normativity – proper function is fixed by a component's history, not by any of the component's properties. The component's properties may be incapable of performing its proper function, which leads to malfunction. This translates extremely naturally into an understanding of intentional content – a neural process may have the proper function to represent, say, rivers, but if this process occurs at the wrong time, or fails to occur at the right time, then we might rightly say that it is misrepresenting. This may manifest as a case of representing a river when no river is present, or of failing to represent a river when there is a river present – thus, error, and misrepresentation.

The selected effect theory of function has recently been extended and improved by Justin Garson (2011, 2012). Garson argues that selection is the important part of teleofunctionalism, and that natural selection is just one, albeit important, way selection can occur. However, brain processes and components undergo selection on a much faster timescale than that required for natural selection, namely, through learning. Garson contends that what he dubs *neural selection* is an alternative route through which proper function can be acquired. Thus we can avoid some of the classic objections to teleofunctional theories – viz. that they have difficulty accounting for the brain's clear capability for dealing with modern artefacts. Smartphones and catch-up television, even things like bread and other products of agriculture have appeared in our milieu too recently for us to have evolved proper responses to (and representations for) them, yet we manifestly have. The most popular responses appeal to a clunky and unconvincing process through which representations of the modern are constructed out of more basic representations generated through natural selection (Neander, 1999). Garson's theory (2012) circumvents these issues – allowing teleofunctionalists to account for proper function and neural misrepresentation by appeal to the brain's *learning process*, which imposes selective forces on its mechanism. In this way we can say that a certain process has the proper function to represent Coca-Cola cans because it is the product of a learning process that selected it for its capacity to represent Coca-Cola cans. If in the future it fails in some way to function as a representation of Coca-Cola cans, then we can legitimately call this a case of misrepresentation.

Despite these advances, teleofunctionalism, and the broader selected effects theory of function have a serious vulnerability. Although Garson (2012) offers us good reasons for a more general notion of selection, broader than natural selection, theories of function based on selection are still, and always will be, etiological. Etiological theories face serious conceptual difficulties due to old and well-known problem examples such as 'swamp man' (Davidson, 1987). Let us imagine that due to a freak quantum accident in a modern swamp, organic matter forms itself into precisely the right organization of a 35-year-old modern man. We can push further and stipulate that this man, 'swamp man', has all the right neural circuitry to recognize and talk about the wonders of the modern world. What makes swamp man a problem for etiological theories of function is that he does not have an evolutionary history, nor does he have a neural history, swamp man has no history at all. Etiological theorists are thus compelled to conclude that swamp man's organs have no proper function, nor does his brain contain functional components such as representations. What makes this denial absurd however, is that swamp man is absolutely identical to a modern man in all respects except his development. He can navigate his way to an airport, use an internet café to book himself a ticket to the nearest metropolitan centre; from there he might visit the local university and engage in some discourse on the conceptual foundations of biological function. Swamp man may be able to do this, but the etiological theorist must maintain that his biological and neural processes are completely lacking in function, and thus his thoughts have no content whilst his identical peers have rich mental lives enabling them to engage in identical activities.

Apart from being horribly unfair to swamp man, the etiological approach leads to absurdity when we try to explain his biology and behaviour. As swamp man's heart doesn't have a proper function, we can't explain that his cells remain oxygenated in part because he has a heart connected to a network of blood vessels. In order to explain this we would have to provide detail down to an atomic level specifying the precise causal chain that results in the phenomenon we want to explain. That is, with swamp man, unlike any other identical human, we are not permitted to appeal to function in our explanations. So, absurdly, we might have two identical systems (a man, and swamp man), performing exactly the same tasks (talking about philosophy, breathing, staying alive), and be unable to provide the same explanation in each case, simply because of an accident of history.

There is, of course, a substantial literature on this kind of argument and its implications for etiological theories of function (e.g. Neander, 1991b; Garson, 2011). For the purposes of this thesis however, I will leave the discussion here, having noted this serious problem for the teleofunctionalist and selected effect theorists, who currently carry the baton for the notion of proper function. My aim is only to show that there are no simple answers to the problem of intentionality, which is the problem of grounding symbols in natural systems. Cummins (1996) who I used to introduce this problem, only re-names it – calling it the task of target fixation, but it amounts to the same thing – grounding of intentional contents – and he discusses the very same theories mentioned here.

1.3 Moving forward

The puzzle of computation and representation in natural, biological systems such as us, is this:

> Premise 1: If a system is a computer then it possesses internal representations.
>
> Premise 2: Our best explanations of cognition assume that cognitive systems are computers.

Conclusion 1. So, cognitive systems are computers, via inference to the best explanation applied to P2.

Premise 3: We have no way to ground the intentional content of internal representations.

Conclusion 2. So, it is doubtful that natural states possess intentional content.

Conclusion 3. From P1 and C1: We should believe that cognitive systems possess representations.

Conclusion 4. From C2: It is doubtful that cognitive systems possess representations.

C3 and C4 are in clear tension, and the source of our puzzle. In a radical mood we might even draw the further conclusion:

Conclusion 5. From 1 and C4: It is doubtful that cognitive systems are computers.

Which is in tension with C1. The failure to ground intentional content thus bites on both our understanding of mental representation, *and* our conviction that minds are computers.

Whilst all three premises are potential targets for attack, as I have noted in the main discussion above, it is by denying Premise 3 that we might solve the puzzle, rather than deflate it. The puzzle really is a simple consequence of the problem of intentionality, and the dominant computationalist paradigm in cognitive science which provides predictive processing with its core assumptions.

However, predictive processing is not a *classical* computational theory. Models of predictive processing are usually implemented in biologically inspired connectionists networks, which we sketched in the introduction. In the following chapter I will now relate the discussion here, which has focused on issues for classical computation, to computation in connectionist models, so that we can get a firm grip on the salient points of discussion when we sharpen our focus to predictive processing in Parts Two and Three.

2. Connectionism: Symbols and Systematicity

We are most familiar with *digital* computation. All the wonders of modern 'smart' technology are digital computers with von Neumann architectures. When we imagine the language of computers we think of strings of binary code somehow being processed incredibly fast on tiny pieces of silicon. When one first learns about connectionism, it comes as some surprise that these models are supposed to be a kind of computer – a computer that isn't digital. This surprise is somewhat justified, given the significant differences between digital and connectionist machines. The flowchart style process that exemplifies classical computation shows how algorithms are to proceed in a stepwise, *serial*, fashion. Connectionist units do not process information through serial logic functions, but operate in parallel as described in the introduction. The activity of units in a connectionist (or neural) net is captured with the use of dynamics – equations that govern when and how they fire (Macdonald, 1995). This is an important point of commonality that connectionist models have with DST. Furthermore, the activation of units isn't to be understood semantically – that is, they aren't to be interpreted as classical symbols. Rather, neural network processing is said to be *subsymbolic*. Processing occurs below the level of semantic evaluability.

Some connectionists do deny that connectionist processing is necessarily subsymbolic, these are the implementationists, who believe that connectionist networks can straightforwardly implement classical computations (e.g. Ballard and Hayes, 1984; Ballard, 1986). If the mind is a classical computer, then something like this must be the case – after all, the brain is a kind of neural network, and it must be implementing classical computation somehow. However, for the most part, connectionists believe that while the processing is subsymbolic, at a higher level of description, semantically evaluable states (viz. familiar, chunky symbols) emerge. The activation of clusters of units, or a vector that describes the activation of the whole system, or at least large parts of it, are taken to be states with intentional content. In this way *radical* connectionists (those that are not implementationists) take the processing of connectionist networks to have a *split-level* description (Macdonald, 1995). On the top level one can analyse aggregates and find symbolic processing, and thus make sense of intentional cognitive phenomena, but on the bottom level, the algorithmic level of units and connection weights, there are no semantically evaluable states. This is the main difference between connectionism and classicism. In classical architectures, the algorithm operates on semantically evaluable, 'chunky' symbols; whilst in connectionist architectures, it does not.

Classical computation provides an elegant and intuitive account of the *systematicity* of thought. Systematicity is the ability to decompose and recompose thoughts using the same parts but to mean different things. For instance, we can think the thought 'Smolensky wrote about Fodor', and we can also think the thought 'Fodor wrote about Smolensky'. Despite being composed of exactly the same words[6], the two thoughts clearly have very different meanings – one might be true while the other is false. In fact, the two sentences are only related in virtue of the *semantic* properties of the constituent words. Systematicity is thus the ability to take the parts 'wrote about', 'Smolensky', and 'Fodor' and recombine them in ways that mean different things, but all the while those constituent pieces retain their meaning. In the last chapter I mentioned how the classical theory produces productivity, highlighting the way our finite minds can, in principle, think an infinite number of thoughts. Similarly, systematicity is a property of our mental processes that cognitive science must explain – if a paradigm is unable to explain the systematicity of thought, then that paradigm is a non-starter, not worth pursuing within cognitive science.

The classical theorists Fodor and Pylyshyn (1988) argued that while systematicity comes naturally to classical systems, that connectionist systems cannot account for it, so connectionism will never be able to provide adequate models of cognition. The first task of this chapter will be to provide a brief overview of this debate. In his thoughtful response to this systematicity challenge, Paul Smolensky developed an attractive and robust notion of representation for connectionism, which will in turn help us get a grip on the ways in which predictive processors might represent. The second task of this chapter is to challenge that notion of representation. We will find that unlike classical

[6]… and those same words possess the same meaning in both sentences, a property known as *compositionality*, which will be important in short order.

systems, whose non-intentional content (recall Cummins, 1989) is easy to identify, connectionist representations are slippery and not obviously useful explanatory posits (Ramsey, 2007). So defenders of connectionist representation face the dual challenge of first proving their explanatory worth, and then finding an adequate theory of intentional content to answer the problem of intentionality.

2.1 The Systematicity Challenge

Systematicity requires that thoughts possess structure. The ability for the two thoughts ‘Fodor wrote about Smolensky’ and ‘Smolensky wrote about Fodor’ to be both related and different requires that the representation of those thoughts possess some structure internal to both that can be related in some way – that is, they are made up of the same parts (Fodor and Pylyshyn, 1988). Classical computation with its symbols standing in robust syntactic and semantic relations entails that systematicity holds. On the other hand, connectionist representations are unstructured. The aggregate activation of the units in a connectionist network possesses no internal structure. Representations are *distributed* wholes that are not made up of semantic parts, and so are unable to possess the kind of semantic structure necessary for systematicity. That is, while the connectionist network might *token* both ‘Fodor wrote about Smolensky’ and ‘Smolensky wrote about Fodor’, the network will not be tokening ‘Fodor’ in both those sentences.

This complaint regards *compositionality*, a feature closely related to systematicity. We have discussed it obliquely already, but to be explicit – compositionality describes the property of thought and language, that representations can be composed out of constituent representations, and those constituents can participate in many different representations, but *retain the same semantic content across contexts.* In other words, the semantic building blocks of thought are *context insensitive*.

Now, one of the attractions of connectionist models are their context sensitivity. They can learn to become sensitive to fine-grained properties of their domain and take them into account in their tasks. Small differences in input can result in large differences in the processing and output of a connectionist network due to the massively interconnected architecture and resulting distributed processing. However, this does make it difficult, intuitively, to square the compositional nature of thought with a connectionist vision of cognition. This difficulty then makes systematic thought impossible, hence the challenge.

Paul Smolensky (1987, 1988, 1995a) resists this conclusion. He shows how the aggregate activity of a neural network can exhibit compositionality and systematicity, yet not be a mere implementation of classical processing. The theory Smolensky presents is a technical method of instantiating representations in connectionist networks, which he calls *tensor product representations*. The essential idea is that in some networks, analysis of connectionist activation vectors is possible in a way that they decompose into meaningful parts that enable compositionality, and a part that defines their structure in a way that enables systematicity. In networks where this is possible, systematicity is possible and thus the challenge is answered. The goal is not then to show that all connectionist networks exhibit systematicity, but just to deny Fodor and Pylyshyn's (1988) claim that no connectionist network can exhibit systematicity without being a mere implementation of classical processing, by describing a subclass of models with distributed, but structured representations. I will not go deep into the technicalities here, but a sketch is necessary to have some grasp of the shape of Smolensky's response.

To produce a tensor product representation, we must identify *filler vectors* and *role vectors*. A distributed, but systematic representation is simply the product of two such vectors. These representations can then themselves be combined through addition to produce more complex representations. The filler vector is a context-independent vector with some content that is invariant across contexts, thus facilitating compositionality. The role vector specifies the role of the filler vector. So, considering our representation of 'Fodor wrote about Smolensky', in a classical system the representation may be structured like this [W,[F, S]]. For a tensor product representation of this sentence, we need two role vectors **r1** and **r2** (one for position 1 in the sentence and one for position 2), and three filler vectors, **W**, **F**, and **S** (one for each meaningful element). We can thus construct the representation by taking the product (x) of each role vector and filler vector, and adding (+) them sequentially to produce **r0**x**W** + **r1**x[**r0**x**F** + **r1**x**S**] (adapted from Smolensky, 1995a). The resulting aggregate is a distributed representation defined over the activity of connectionist units in a distinctly non-classical system, but one that can be broken down into structured, context invariant parts.

What is important to note about the tensor product theory is that, as an account of systematic representation in a properly connectionist network, there remains (at least according to Smolensky) a disconnect between the symbolic level of description and the causal processes underpinning the computation. As mentioned above, unlike classical computation, representations in connectionist networks are not syntactically related. What Smolensky has shown is that there is an alternative way of understanding internal structure that at once makes sense of systematicity, but does not permit syntactic individuation of symbols. Connectionist processing is still to be described in terms of the dynamics of the network whose evolution is governed by differential equations. Representations exist at a separate level of description, as 'vectors partially specifying the state of a dynamical system' (Smolensky, 1995a). This level of description allows us to understand the way connectionist networks perform computations, but the nature of this computation is radically different to the classical conception.

2.2. The Job-Description Challenge

The debate about systematicity in connectionist models was by no means settled by Smolensky's tensor product analysis (1987) (See for example, Fodor and McLaughlin, 1990; Chalmers, 1993; Aizawa, 1997; Jansen and Watter, 2012). However, for our purposes it is useful just to see the shape of the connectionist response, which helps us get a firmer grip on the kind of representation that must be invoked to handle the demands of cognitive science. The split-level understanding of connectionist computation, with a gap between the symbolic and algorithmic level poses a problem when we consider again the symbol-grounding problem. Let us once more recall the classicist's solution to the problem.

Above, I briefly described Cummins' dissolution of the symbol-grounding problem for classical systems. There is no symbol-grounding problem according to Cummins (1996), because the syntactic relations between symbols constitute a formal structure. The syntax itself thus ground the meaning of the symbols. In the words of John Haugeland (1981), "if you take care of the syntax the semantics will take care of itself". Clear from Smolensky's response to the systematicity challenge, is that connectionists receive no such free ride from syntax to semantics. The structures that represent are not those that are causally involved in running the program (Smolensky, 1995a).

This multiplies the connectionist's workload when it comes to understanding representation and intentionality. It opens the door to objections against the claim that what Smolensky is describing as a representation, really functions as a representation. This is the point that William Ramsey applies pressure to in *Representation Reconsidered* (2007).

Ramsey (2007) poses the question along these lines, "are the states and structures that connectionists call 'representations' functioning as representations? If so, how?" For example, there is a molecule, mRNA, that cellular biologists describe as functioning as a 'messenger'. What makes this a good description is that the molecule acts as intermediary between DNA and the developmental process of protein-building. Describing mRNA as a messenger aids explanation and biologists seem to be justified in their functional ascription, in a way they wouldn't have been if they described it as a 'passenger' or a 'learner', for example. Thus we can say that mRNA is *really* a kind of messenger given that it fulfils the function typically fulfilled by messengers and that it performing this function is vital for certain explanations in biology. Thus we are convinced that mRNA are *messengers* because they perform a messenger function, *and* that messenger function is important for biological explanations.

Ramsey argues that if we are going to be convinced that certain states and structures are representations, we require a similar argument to the one made above. That is, to believe that a certain state *is a representation* requires that they perform a representational function, *and* that that function is important for explanations in the relevant domain. These requirements constitute the *Job-Description Challenge* (JDC), so called because any theoretical posit being called a representation must show that it fulfils the representational 'job-description'. There is still at least one important question to be resolved before proceeding: what is a 'representational function'?

In this case, Ramsey hesitates to commit fully to a single characterisation of representation. He states that 'in the case of both non-mental and mental representation, the relevant roles include things like *informing*, *denoting* or *standing for something else*', adding that 'it is not at all clear how these sorts of roles are supposed to be cashed out in the naturalistic, mechanistic framework of a cognitive theory' (2007, p. 25). This short aside highlights the slippery nature of the symbol grounding problem that remains for the connectionist. Ramsey's insight is that what are important in solving this problem are identifying *functional* requirements that satisfy the constraints of the problem. This goes for any everyday term that scientists invoke – mRNA is only correctly termed a 'messenger' because it really performs a messenger's function.

Perceptual and categorisation tasks are a particularly popular application for connectionist models. Sejnowski and Rosenberg's (1986) grapheme-to-phoneme converter, NETtalk, is an exemplar of this kind of research. After its training, NETtalk would read one letter at time, in the context of three letters either side and activate an output unit associated with a certain phoneme. For example, in the diagram below NETtalk is reading the letter 'c', and taking into account its context, the network outputs the phoneme /k/. For NETtalk and other models like it, some subset of units in the network is responsible for detecting a certain kind of input by entering a particular activation pattern, for instance there might be some set of units in NETtalk that activates only (or primarily) when the character 'c' is the input. Recall the kind of content grounded in natural meaning such as smoke means fire, it seems plausible that the components of NETtalk possess content in this kind of way. Such components are referred to as 'detectors' or 'indicators' and representational language is usually invoked to make sense of their function. However, Ramsey (2007) labels this kind of representation the ‘*receptor* notion’, arguing that in fact these components are not playing a special representational role at all.

While it makes sense to label these units as being sensitive to a particular class of stimuli, this doesn’t entail that they are functioning as representations in the sense demanded by the JDC. On the representational interpretation, these units are indicating the presence of a certain object to the system, which seems to be an uncontrovertially representational function. But Ramsey has a different and more parsimonious interpretation of their function, which is as a ‘reliable causal mediator’, which is a role filled by many components in other systems which we do not interpret as representations at all. For instance, a tap that switches on using some kind of heat or motion sensor uses these components to activate the flow of water when a hand is present, but these components do not ‘represent’ hands, though they seem to be playing a very similar function to receptor-notion representations.

Here it is important to stress the strategy Ramsey (2007) is employing here. Ramsey concedes that the receptor notion of representation may help us identify intentional content – that is, we can see what distal states the activation patterns covary with (see our discussion of causal accounts of intentionality above). However, simply being a receptor is not *sufficient* to perform a representational function. If it were, we would suffer an explosion in the sorts of things that would count as representations, rendering the notion too cheap to perform the explanatory work it is used for. The moral of this argument is that having contents is not enough to be a representation, that is, if the ‘contentful’ state is not performing a representational function, it should only be considered to have ‘potential content’, in that the state could potentially function as a representation of that content, if it were employed in the right way by the system. For example, our motion sensor in the tap could potentially be used to represent the presence of a hand, if for instance a researcher wanted to know how many hands were washed by such a tap in a train station, they might collect data about how many times the sensor was activated (n) and use that to infer that 2n hands had been washed by that tap during the period of the study. Not a particularly exciting example, but functional. To the researcher, each activation of the motion sensor represents a pair of hands being washed, and she can perform inferences using those representations. However, for the tap, the sensor functions only as a switch, no semantic content is necessary to make sense of how the system works. This curious feature of receptor-style representations highlights how Ramsey has brought the debate forward, identifying the two-part demand on representations, one, they must have content, and two, they must perform the right kind of function. You might have one without the other, but that would not be sufficient for RL to properly apply.

The second notion of representation that Ramsey attributes to connectionism is 'tacit representation'. Like the receptor notion, connectionist researchers frequently invoke this kind of representation. Rather than perceptual faculties, tacit representations are used to make sense of memory networks, another important class of model. A contemporary example of this is Facebook's 'Question-Answering (QA)' network, MemNN (Weston et al., 2014). The success of these networks at answering a broad variety of questions of many different forms and requiring a diverse array of reasoning processes is explained by adopting RL, particularly by postulating that the connection weights of the network encode knowledge of the training data.

Ramsey labels this stored, learned knowledge *tacit* representation because the states are not explicit in the system. That is, unlike receptor representations, or classical symbolic representations, you cannot localise a component in the network and say, "that part of the system represents the determinate contents *x*." Tacit representations are *distributed*, that is, representations of many different contents are stored in the same network of connections. The evidence for this really being knowledge, and thus representation of a sort, is that connectionist networks employ learning algorithms in order to alter their connection weights in such a way that they make fewer mistakes in future on the task they are being trained to perform. When they are tested, these networks often perform very well, thus demonstrating that they 'remember' the information they have been 'taught'.

However, Ramsey's criticism of cases like this is convincing. Success at question-answering tasks and similar tests is a result of the *dispositional* properties of network architecture. If we attempt to deconstruct the network mechanistically, there need be no component or process that is performing a representational function. As Smolensky admits – structures at the representational level of analysis are playing no causal role. All the work is being done by a set of connections which, when activated in certain ways, result in a predictable set of output pattern that a learning algorithm has gradually moulded the system into producing. However, the learning algorithms in place do not help form representations or inferential processes, they merely reinforce desired associations and inhibit undesirable associations. While it is amazing and useful that inference-like and knowledge-like behaviour can result from the recurrent application of these 'dumb' algorithms to 'dumb' neural networks, representations nevertheless play no causal role for the system. In other words, nothing in the system is performing a representational function, what we have is a complex component (the network) which has a set of dispositions (to produce the right answers), and that component can be broken down into subcomponents (individual units) with their own causal capacities. But at neither of these levels is a representational function being performed by a physical element in the network. To some, eliminating representation from explanations of these systems would cripple our understanding of them: it just makes sense to talk about them representationally. But not only does Ramsey's argument demonstrate that nothing of explanatory value is gained by describing these systems as representing, it also misconstrues their mechanical structure. Further, the non-representational understanding demonstrates precisely that the *appearance* of representational thought and behaviour can emerge naturally from the dynamics of a system involving no representations at all. This argument motivates an ambitious eliminativist position.

2.3 An Answer

Ramsey (2007) has suffered criticism for failing to consider a third way that representation is understood to function in connectionist models. That is, as a kind of *model*[7]. Recall once more the way Cummins (1996) deflates the symbol-grounding problem for classical computation. He appeals to formal isomorphism, grounded in the physical structure instantiated by the syntax of the system. A model is a kind of formal isomorphism, indeed, Cummins (1996) uses the example of maps and models numerous times to illustrate his arguments. If we could show that connectionist networks instantiate a kind of model, then we can answer the JDC, and ground representational function in just the same way as classicists do.

Oron Shagrir (2012) makes the argument that neural networks instantiate structural isomorphisms that can ground representational function in the way Ramsey demands. Drawing on work by Amit (1989), Seung (1996), Eliasmith and Anderson (2003) and others, Shagrir aims to show that non-classical networks utilise isomorphism by proving the case in a single instance, the biological oculomotor working memory, and generalizing to the whole class. I will briefly summarise the pertinent details of these oculomotor memory networks and Shagrir's analysis to demonstrate one way we can ground the semantics of neural networks.

The function of the oculomotor memory is to keep the eyes steady between saccades. A semantic story is easy to tell. Following each saccade, the network encodes the position of the eyes, which is then used by motor neurons to ensure the stability of the eye until the next saccade. The state of the network at any given time is thus taken to represent the current eye position. The network receives information only about the saccade, that is, its velocity and direction. It is thus often referred to as a *neural integrator* (Robinson, 1989; Seung, 1998; Eliasmith and Anderson, 2003), as it is taken to perform an integration function, taking its previous state and the velocity of the saccade as arguments to transform into a new state.

On this, bare-bones understanding, the network is only taken to utilise receptor representations. 'Pulse inputs are correlated with eye velocity, and certain persistent

[7] This notion is of especial interest to us, as predictive processors involve the use of *generative models* which I will discuss in more depth and with proper technical context in Part Two.

collective activity in the neural network is correlated with a certain eye position' (Shagrir, 2012, p. 531), these causal correlations amount to the reliable activation patterns discussed in relation to the flawed receptor notion, this much is conceded by Shagrir. Recall that reliable correlation of this sort is not sufficient for passing the job-description challenge. Despite a certain pattern of activation in the oculomotor integrator being a good indicator of a certain eye-position, this does not entail that the pattern of activation is playing a sufficiently robust representational role to qualify as a mental representation[8].

However, this isn't the whole story. Shagrir aims to show that the oculomotor memory network utilises these receptor-style inputs as part of a model of angular eye positioning. By showing that these receptor representations not only correlate with some content, but also function in virtue of having that content, which permits them to graduate from being what Ramsey calls potential representations (a contentful state that isn't playing a representational role) (Ramsey, 2007, p. 139) to fully-fledged, properly grounded, mental representations.

The model Shagrir appeals to begins to emerge when we employ the tools and methods of dynamics. First we construct the state-space of the network. This is the mathematical space of the system's potential states. Each point in the space corresponds to an activation vector, with each element of the vector corresponding to the activation value of a single cell in the network. Imagine measuring the activity of each neuron in the oculomotor integrator at a certain point in time, then writing down the measurements in a list. This is one activation vector, which can be plotted on an imaginary graph with the state of each of the *n* neurons having its own axis, producing an *n*-dimensional space. These state-spaces are useful as they allow us to map all possible states of the network. Once we have some mathematical laws governing the dynamics of activation patterns, we can draw trajectories in our space, i.e. given some

[8] Ramsey (2007) provides an excellent example of this. He asks us to think of a pistol – striking the firing pin invariably correlates with a bullet being fired; however, we would not go so far as to call the firing pin a representation of a shot.

initial state, we can see how the state of the network will develop over time, plotting each discrete activation vector on the graph.

One further tool of dynamical systems theory is necessary to grasp how Shagrir intends to cash out the representational nature of the network, *attractors*.

Some activation states of the network are stable, and others are unstable. As these labels suggest, the network tends to remain in stable states and, over time, drifts away from unstable states. Using the state-space, we can map the relative stability of each activation. Stable states appear as troughs, and unstable states appear as peaks. The resulting gradient of the space will reveal attractors as troughs which are surrounded by peaks. The state of the network will tend to drift from peaks to troughs via the most direct route, and will drift faster along steep gradients, and more slowly down shallow gradients (Seung, 1998; Robinson, 1989; Eliasmith and Anderson, 2003).

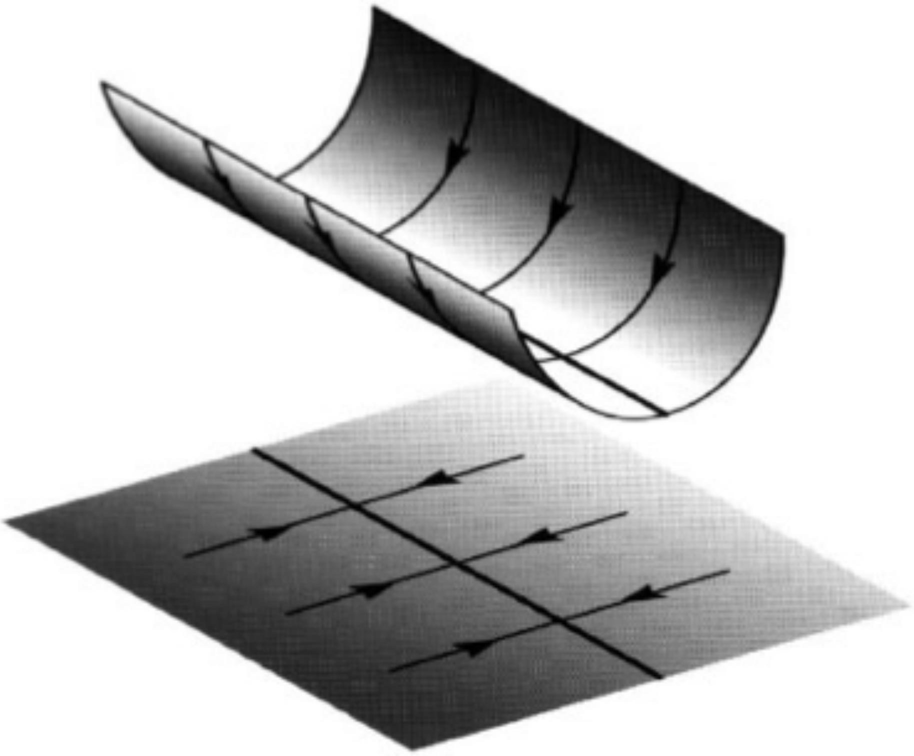

Fig 2.1: Line attractor (From Seung (1996) p. 13340; Copyright (1996) National Academy of Sciences, USA)

The state-space of the oculomotor integrator network takes the form of a *line attractor*, pictured above. Just a narrow range of states are stable, and the cliffs either side denote which states of the network will tend to certain points of the attractor. The network is perturbed by each saccade, and settles to a new point in the attractor. Shagrir argues that the attractor is a model, with a formal structure, and the states of the network have semantic content in just the same way as classical symbols have content in virtue of their participation in a formal structure:

> 'Each state, Si, along the line attractor encodes a different eye position, and the distance between two states, Si and Sj, corresponds to the distance between two eye positions, Ei and Ej. We thus have a sort of isomorphism between the representing network and the eye: the function that maps the stable states, Si's, to the corresponding eye-position states, Ei's, is type preserving, in that the distances between two states mirror the distances between eye-positions' (Shagrir, 2012, p. 532).

The isomorphism present in this specific case is an integration relation; the pulse inputs to the memory network stimulate the cells in such a way that they move to a new point in the state-space describable as integration, and the actual eyes move to their new position which is described as an integration over the velocity of the saccade.

By functioning as an integrator, the network preserves salient relations between the velocity of the saccade and eye-position, using the pulse input and the activation state of the network as representations. These count as grounded representations because, in the context of the model, the distance between the points in state-space are isomorphic to the horizontal distance between two eye-positions. In simple terms, the motor control system for the eye uses the oculomotor memory as a map of where the eyes ought to be at any given time. Just like our previous example, the map only works because it preserves the right relations between its elements.

In short, with his analysis of the oculomotor integrator, Shagrir claims to have firmly established the existence of isomorphism in neural networks. Supplemented by the observation that the use of attractor networks is a popular paradigm in neuroscience (for instance its use in head-direction networks is noted by Eliasmith and Anderson, 2003), Shagrir claims that isomorphism can be used to ground the representational ascriptions in vast swathes of neuroscientific research. We can simply extend this claim to connectionist networks, whose behaviour can be modelled using the tools of DST in exactly the same way as the oculomotor integrator.

2.4 Connecting Connectionism

Connectionist models of cognition face a tougher challenge than classical models. Not only do classical computers have certain distinctly cognitive qualities built in, such as productivity, systematicity, and compositionality, but they also have a neat solution to the symbol-grounding problem, if not the broader problem of intentionality. Advocates of neural networks also must answer the problem of intentionality, but before that they have to make sense of what it could mean for high level descriptions of a system to be representations – grounding the basic building blocks of thought in a way that accounts for our rational cognitive capacities.

This is all hard work, and Smolensky's answer is just one possibility. However, he introduces the constraint that a notion of neural representation must preserve the distinction between classical computing and connectionist computing. That is, at the algorithmic level, connectionist computing is subsymbolic. This hard work is rewarding however, after all classical systems simply *assume* systematicity as part of the architecture – the connectionist demonstrates how systematicity of thought can be *explained* by neural networks (Smolensky, 1995b). The problem this split-level conception of connectionist computing generates is that representations are no longer causally, and thus functionally, involved in cognitive processes. This invites the kind of criticism we saw from Ramsey (2007), whose charge that representations in Smolensky's parallel distributed networks are not functioning as representations at all, but as mere dispositions.

The answer we discussed in response to this problem from Shagrir (2012) attempted a solution through careful analysis of neural network dynamics using the tools of DST. The shape of a system's phase-space can exhibit isomorphisms that can be exploited in just the same way that the syntactic structure of a classical computer is exploited to ground representational content. Applying DST to connectionist systems in this way is a theme I will return to in chapter seven, when we consider the distinctive framework developed by Karl Friston – *free energy minimisation* – that we can use to get a firmer grip on the kind of representation being exploited by predictive processors.

The problem of intentionality for connectionists is identical to the problem for classicists – once the connectionist has identified the states supposed to be playing a representational role (at whatever level of description they emerge), we can ask how they gain their intentional content. The causal and teleological theories discussed in chapter one are as applicable to these models as they are to classical models, so the problem of intentionality remains, given that we currently have no convincing account available.

There is just one final task to perform before we take a closer look at predictive processing architectures. The discussion so far has kept a close focus on the main thread of debate surrounding the related problems of symbol-grounding and intentionality. Now however, I will consider some alternative theories which have had an impact on these debates, and will be useful to bear in mind as we move forward with our discussion of predictive processing.

3. Stance and Orientation: Dennett, Chemero, and Action

The complexity of the brain, with its 100 trillion connections between relatively simple neuronal processors, along with the connectionist suggestion that the processing of such a machine occurs at the subsymbolic level invites doubt that the human brain is a representational system at all. At least, we might doubt that the computations it performs can be adequately captured in terms of maps and models of the environment, or in terms of the neat beliefs and desires appealed to in everyday, folk psychology. When we make new discoveries about the brain, the rift between our manifest image of the mind and the scientific image seems to grow. So, what justification do our intuitions provide for the claim that our mind computes classical-style representations, of the sort proposed by Fodor and Cummins we considered in chapter one? Indeed, if there is a compelling alternative, that can undermine our computational vision of the mind, what justification do we have that there is any such thing as mental representation?

In this chapter I will provide sketches of three sceptical approaches to the notion of representation. We begin with the least sceptical, which is in fact compatible with the classical picture, but prompts us to re-imagine the kind of content we should expect representations to possess – *action-oriented representation*. Next I will consider the *intentional stance* of Daniel Dennett (1987) which defies neat categorisation, but occupies a position somewhere between outright belief in representation and outright scepticism – rather, Dennett focuses on the explanatory utility of invoking representations to explain cognitive systems, a strategy which will prove useful for us as we consider the utility of invoking representation in explanations of predictive processing in Part Three. The final theory for consideration is Anthony Chemero's (2009) eliminativist position – *the dynamical stance*. Employing a conscious parallel with Dennett, Chemero argues that the notion of representation has no explanatory utility when we adopt the tools of DST.

However, to provide some context for these alternatives, I will briefly survey what has become known as *4E cognitive science*. A collective term for *embodied*, *embedded*, *extended*, and *enactive* approaches to cognition, 4E is now an established sector of cognitive science and the broad lessons it teaches have some implications that are widely recognized, and others which are perhaps underappreciated. Widely recognized now is that action is an important part of cognition, precisely how important is debated – but action planning and execution play a role in an ever-growing number of cognitive domains (Matheson and Barsalou, 2017). However, whether or not cognition is *necessarily* action-involving is a contentious issue.

Advocates of *embodied* cognitive science focus on the important role that a system's body plays in cognitive tasks – usually those properties of the body that affect the action output of the neural system. Embodiment theorists recognise that the neural system is just an important organ in the wider cognitive system – when planning to throw a ball, for example, the nervous system need not compute things such as muscle-elasticity into its calculation. Those things are 'calculated' by the muscles themselves, all the neural system needs to have done is adapted to the actual properties of the body it is attached to in order to make effective use of that body in planning and performing actions. Thus, the embodied cognitive scientist argues, cognition is not just in the head, the body itself plays a constitutive role too.

Researchers studying *embedded* cognition go a little further. They may accept the lessons of embodiment, but also argue that we should recognise the way that our particular way of cognising is affected by our immediate environment. A common way of talking about these effects is to refer to the way the environment *scaffolds* cognition. Scaffolding can make tasks easier or more difficult. Think about playing scrabble, boggle, or watching countdown. One way of finding more words is to physically move the letters around. It seems obvious that, if we were presented with the same set of letters on different occasions, but they were ordered differently (and we weren't allowed to move them about) we could end up with very different lists of words constructed out of those letters, and end up with very different final scores as a result. The position of the letters scaffolds the cognitive task of finding words. Success at the task is not a simple case of brain power, but also depends on the way our cognitive system is embedded in the environment.

Pushing the limits of cognition even further the *extended* cognition advocate argues that cognition is not just scaffolded by the environment, but sometimes the environment is *part* of the cognitive system. Humans are particularly good at recruiting external objects to help with cognitive tasks. For instance, when you solve a problem using a pocket calculator, that calculator was part the cognitive system that solved the problem, it was part of *your* cognitive system. The insight that advocates of extended cognition make is that components of cognitive systems are identified by the function they play in the system. Sometimes, external objects play distinctly cognitive functions, viz. functions that would count as cognitive if they were being played by some structure in the head (a famous example is the practice of writing important bits of information in a notebook – that notebook is functioning as part of your memory store, as much as any part of your hippocampus functions as a memory store), so those external objects are used as part of the cognitive system just as much as any neural component might be. Cognition is not just influenced by the environment, but, like the body, sometimes the environment really does cognitive work.

The fourth E is for *enactive* cognition. Enactive theorists emphasise the role of action in cognition. For example, Alva Noë (2004) argues that action is a constitutive element of perception. Perception is a skilful activity that involves active exploration of sensory states, and our percepts are thus action-involving: we know that the cat that is occluded by a fence is, in fact, a whole cat and not slices of cat divided by lengths of wood because we know that if we were to hop over the fence, that we would find a whole cat behind it, and not the gory remains of a cat cut into strips. This seems obvious to us, but that is because we are all extremely skilled perceivers, who have got very good at incorporating this action-involving knowledge into our cognitive routines. The broader conclusion of the enactive theorist is that all cognition is action-involving. Varela, Thompson, and Rosch (1991) make the case that it is through action that a cognitive system and environment co-evolve and help to create each other. Digesting the lessons of the other three Es, the enactive programme stresses that mentality is not a passive capacity of a system, but a product of the rich, skilful, two-way interaction between system and environment.

Enactive cognitive science is a fairly radical position, but all of the four Es are, in principle, compatible with the computational methods of classical, connectionist, and hybrid cognitive science. However, there are those that eschew computation altogether, these are those who endorse *radical embodied cognitive science* (Chemero, 2009). These theorists argue that what 4E theories teach us is that in fact, there is no principled way to decouple the brain-body-environment systems. To explain behaviour, we have to take into account variables in all three supposedly separate systems. Treating any of the three separately will result in a loss of explanatory power. Being thus unable to decompose these systems into their constituent, mechanical-computational, parts, the way we have to make sense of this complex system is to understand its dynamics. That is, not to separate these closely coupled systems, but to identify the variables and parameters that are salient for predicting and explaining their behaviour.

3.1 Action-Oriented Representation

As discussion of the role of brain-body-environment interactions has grown, this has led to the development of a corresponding notion of representation – *action-oriented* representation. Action-oriented representations (AORs henceforth) are representational states that do not involve action-neutral knowledge of a world reconstructed by perception. Rather, AORs occupy a middle ground. Adopting this approach bypasses the use of heavy-duty world models for many cognitive tasks, suggesting that cognitive systems need to encode only properties of the world that are relevant for the agent's ongoing engagement with its immediate environment. Rather than my brain having to represent a coffee cup as "25 centimeters away from my right hand", it can get by more effectively just by representing it as "reachable". Engel et al. (2013) thus instruct us that 'system states acquire meaning in virtue of their role in the context of action'. AORs are often hailed as a very minimal form of representation, or even 'maximally minimal' (Gallagher, 2008) and given that a large chuck of cognitive behaviour is now quite generally recognised to involve sensorimotor modulation or be otherwise action-involving, it is hoped that classical representations can take a back seat, only getting involved in more specialized tasks.

Classical computational research into vision (e.g. Marr, 1982) drove the 'reconstructive' approach to perception, and this has set a precedent for other areas of cognitive science that investigate faculties which lean on perception (such as tracking, grasping and locomotion). The precedent is that perceptual systems generate a model of the world that represents the size, shape, weight, colour, texture etc. of objects in the environment, encoding these properties symbolically for use by other systems. This programme has been largely successful, and accords fairly well with our phenomenology. However, there have been puzzles to solve, and Anderson (2014) notes the persistence of certain illusions that 'remain even in the face of experience and knowledge' (p.173). These cases, he contends, suggest that there are flaws in the conceptual framework being brought to bear by the classical approach. For instance, consider their proposed explanation of the *size-weight illusion*.

The size-weight illusion is a well-known phenomenon in which experimental participants will consistently report that a large object is lighter than a smaller object of the same weight (Ross, 1969). The effect is typically accounted for by a discrepancy 'between the expected and received sensory input' (Ross, 1969). However, this hypothesis is called into question, given that the illusion persists despite having explicit knowledge of its existence (the researchers themselves are consistently fooled). Anderson (2014) suggests an alternative explanation that makes use of properties of the object that have greater 'ecological relevance', offered by Zhu and Bingham (2011), is superior. The property in question is 'throwability', which subjects are asked to judge, given a selection of objects. In their study, participants were offered objects of six different sizes and eight different weights (48 in total). In each size class, the subject was asked to order the top three objects for throwing. When asked about the weight of objects, the participants suffered from the size-weight illusion, yet interestingly objects judged to be of equal weight were also judged to be equally good for throwing. Anderson's deduction is that subjects extrapolate the property of weight from the action-oriented property of throwability, which is a relational property we are very good at tracking (the preferred objects were reliably thrown the furthest distances). The classical way of looking at things is reversed in this explanation, rather than using an action-neutral model to determine action plans, Zhu and Bingham reveal that estimation of action-oriented properties are used to make a second- order estimation of weight, which is consequently inaccurate. Of course, this theory is not beyond criticism, but I will continue, for now, within the framework that this explanatory inversion offers us, and briefly explain what sort of representation is being proposed.

We find proto-AORs in the work of Ruth Millikan and her discussion of *pushmi-pullyu representations*. These refer to some state or process that specifies an action in response to certain environmental features (Millikan, 1995). For now, this interpretation might be usefully described as specifying neural correlates for Gibsonian *affordances*, that is, brain states that latch on to possibilities for action specified by distal stimuli. A more developed notion is presented as a target for Gallagher's antirepresentational attack. Drawing on work by both Wheeler (2005) and Clark (1997), Gallagher states that "AORs are temporary egocentric motor maps of the environment that are fully determined by the situation-specific action required of the agent (organism or robot)." Representation of this sort is extremely agent-involved, shunning disembodied 'knowledge-that' about a world beyond, and replacing it with just what the agent needs to get along, relative to its own capacities and in its own, current situation. It should be noted that Gallagher's characterization of AORs is not implied by the working definition established in my introductory paragraph. This strong claim opens the door to classical-style representations that would be necessary to facilitate higher-order offline reasoning (cf. Clark and Toribio, 1994). It is thus interesting to note the continuum along which AORs vary – they may be lean, as in Gallagher's example, but need to be propped up by strong classical action-neutral systems in order to account for a full variety of cognitive capacities. On the other hand, our original definition which follows Clark, allows for slightly beefier AORs which can be interpreted as 'both map and controller' (Chemero, 2000) but require far less support from classical systems.

Parsimony

It’s important to consider why AORs might be considered more parsimonious than traditional representations. The reason itself is fairly simple – they permit a system to solve problems with fewer computational resources than if it were using a more classical ‘calculate-then-act’ mode of operation. By continually performing small, easy calculations that result in immediate action and repeating the process until the problem is solved, a system employing AORs can reach the solution more quickly and accurately than one forced to plan and co-ordinate a course of action at the beginning. The most popular example of this is the outfielder problem.

Imagine you are a spectator at a cricket match. The bowler delivers a poor ball and the batsman has plenty of time to strike the ball sweetly and you watch it fly off his bat at great speed. Your friend asks you to predict where the ball will land and you indicate a spot somewhere near the boundary. In fact the ball lands well short of the boundary, to the disappointment of the crowd. As a species, we are typically very bad at making accurate guesses like this, because the calculation has to be done very fast, and there are lots of complicated variables to consider such as the mass of the ball, the drag caused by air resistance and also problems related to the perspective you have of the shot which may not give you ideal information. However, as a species, we are nevertheless rather good at *catching* cricket balls, even under these tricky conditions. Now imagine you are a fielder on the bowler’s team and the ball is flying into the area you cover. If you performed the same calculation as the spectator, and then ran to the spot where you predicted the ball would land, you wouldn’t be a cricketer playing in front of a big crowd, because you’d nearly always miss. However, if you keep your eye on the ball, and run at just the right speed to keep the ball in the middle of your visual field, then as it approaches the ground you’ll usually be in a good position to catch it. Indeed, this is how psychologists now believe we solve this sort of very tricky problem. The moment-to-moment co-ordination of one’s body and visual field requires just the sort of perception-for-action that falls very easily out of a framework involving AORs. Furthermore, by maintaining representations that themselves map percept to action hugely simplifies the computational work required to solve these kinds of problem.

As the necessary computational resources required for many kinds of cognition are greatly reduced by the use of AORs, ontological bloat, compared to classical systems, is proportionally mitigated. It is important to note that AORs may make certain kinds of reasoning more difficult – that is, reasoning about properties of the world that are not action-oriented (recall the extra step necessary to associate weight with the property of throwability). Nevertheless, given that much subpersonal processing will be action-involving, rather than requiring a rich reconstruction of the world, we have a fairly powerful case that AORs offer an attractive alternative to classical reconstructive models.

3.2 The Intentional Stance

AORs mark an attractive point in the space prepared by the symbol-grounding problem for classical computation – they agree that representation is grounded in a kind of model, they just disagree with the traditional way that model has been conceived. In contrast, Daniel Dennett's target is the problem of intentionality. We might call those persuaded by causal or teleological theories of intentionality *intentional realists*. That is, they believe that there are real states that possess intentional content in cognitive systems. Dennett is not an intentional realist in this sense, he has been described as an *instrumentalist*, which captures an important slant of his position, but he believes that this ascription is perhaps too radical. Suffice to say, Dennett occupies a position that is not wholly realist, but he nevertheless believes that we can make sense of intentionality in cognitive systems, but doing so requires observing patterns in the high-level behaviours of the system. This facet of Dennett's work leads still others to describe him as a *behaviourist* (Dahlbom, 1993). Here I will sketch his theory of intentionality, characterised by *the intentional stance*, and leave labels aside.

Characteristically, Dennett argues his point using a powerful thought-experiment (Dennett, 2009). He asks us to consider that we come across two black boxes on a strange planet, box A and box B, connected by a single lead. Box A has two buttons =a= and =b=, Box B has three LED lights, a green light, a red light, and an amber light. We observe the behaviour of the two boxes. Whenever we press =a=, a few seconds later box B flashes its green light; and whenever we press =b=, a few seconds later, box B flashes its red light. This is almost universally true, but in very rare circumstances we do observe that pressing =a= is followed by a red light on box B, or pressing =b= is followed by a green light on box B. The amber light never seems to be used.

Puzzled, we decide to open up the boxes to see how they work. Both boxes are extremely complex computers. We monitor their activity to see what is causing the lights to flash. After a little analysis we find that when we press =a= on box A it is processing a long string of bits, and it then sends the same string to box B, which then processes it and then flashes green. Similarly for button =b=, a string of thousands of bits is processed and subsequently causes a red light to flash on box B. So we look for the differences between the strings – we analyse thousands of examples of '=a=' strings and '=b=' strings to determine what is causing the green lights to light up, and what is causing the red lights to light up, but even laborious statistical analysis reveals no detectable pattern.

Not to be deterred, we begin to provide our own strings to box B, to see whether, by experimentation, we might come to discover the source of the regularity between =a= and green, and =b= and red. However, to our annoyance whenever we provide a string to box B, neither the red nor the green light flashes, but instead the *amber* light. Only very occasionally, once every several million attempts, do we get a red or green light to flash. It seems that despite having all the information about the causal structure of the system, that is, knowing everything there is to know about the syntax of the boxes, we still are unable to explain the regularities in the flashing lights.

Years later we become familiar with the natives of the strange planet who engineered the black boxes. Desperate, we ask them the secret. Box A and box B, they explain, were trained to store propositional knowledge from the (alien version of the) internet.

Box A was then simply programmed to access a random true proposition when =a= is pressed, and a random false proposition when =b= is pressed, then to transmit that proposition to box B for verification, which then flashes green for true propositions and red for false propositions. The rare cases in which =a= is followed by a red light or =b= is followed by a green light are simply instances of disagreement resulting from idiosyncrasies in the learning processes of the two boxes. If the string used is meaningless to box B, not formatted correctly, or with poor grammar, then the amber light flashes. The regularity between the boxes thus only makes sense once we understand the *semantics* they are processing. The semantic interpretation is the only good explanation available for the regularity.

This elegant thought experiment clearly demonstrates a case in which not just the best explanation of the phenomena invokes intentional contents, but the *only good* explanation *must* invoke intentional contents. When we adopt this strategy – the explanatory strategy of ascribing intentional states to systems – we are using *the intentional stance*. We are looking at the systems *as intentional systems*.

The justification for adopting the intentional stance is also contained in the thought experiment – there is a *real pattern* (a causal regularity in the case of box A and box B) in behaviour that is extraordinarily difficult, or impossible, to explain without invoking intentional contents (using what Dennett calls *the physical stance*), but is incredibly easy to explain when we do invoke intentional contents. Dennett's contention is thus that cognitive systems are just like the black boxes – the complex patterns of behaviour exhibited by a cognitive system can only be adequately explained by ascribing intentional contents.

However, the states that possess intentionality just are the real patterns of behaviour, and these may be extremely abstract. This doesn't bother Dennett. For instance, it doesn't make sense to ask what a center of gravity *is*, as it is a well-defined element of our best physical theories, even though they aren't made up of physical stuff. Representations bearing intentional contents are as real as centers of gravity and other *abstracta*, we might not be able to say what stuff they are made of, but we can talk about them in the way our best theories treat them, e.g. as having causal powers, being brought about by certain processes etc.

The problem of intentionality is thus not solved, but dissolved. Whether or not a system should be understood as possessing intentional contents is not a matter of metaphysics – there need be no causal theory or proper function grounding intentionality – rather, it is a matter of epistemology, to be settled by our best theories of cognition. Dennett (1987) clearly believes that our best theories of cognition are computational, and thus that the intentional stance is required to make sense of cognitive systems. But now let's consider the more skeptical alternative preferred by Anthony Chemero (2009), who takes Dennett's lesson about explanation, but rejects his confidence in the computational paradigm.

3.3 The Dynamical Stance

Whilst one way of arguing for eliminativism is using Ramsey's arguments against the legitimacy of certain kinds of implementations of representation (2007), another more radical path, is to deny that cognition is best explained computationally. It is this radical position that Anthony Chemero (2009) sets out in detail. In this section I will briefly summarise his arguments that motivate what he terms 'the dynamical stance' in opposition to mechanism, and then go on to sketch the dynamical image of cognitive science.

The dynamical stance is an explanatory strategy. Consciously mirroring Dennett's 'intentional stance', we find that sometimes we find it fruitful to explain the behaviour of a system by using the tools of dynamical systems theory. The claim that Chemero defends is that cognitive systems, contra Dennett, are best explained using the dynamical stance, not the intentional stance. As a toy example, Chemero cites the famous Watt governor example of van Gelder. Whilst there is a computational description of the Watt governor available by taking the intentional stance, we are much better off taking the dynamical stance, which provides an explanation that is 'precise, perfectly general, [and] counterfactual-supporting' (2009, p. 73).

The primary argument Chemero makes in favour of using the dynamical stance is that, when a dynamical explanation is available, there is nothing left to be explained. That is, the abstract characterisation facilitated by the dynamical stance makes detailed and general predictions, which, if correct, capture all salient aspects of the system's behaviour. All that an intentional, computational explanation may seek to do is provide an explanation with *as much* predictive power as the dynamical stance. A good dynamical explanation places a limit on the explanatory power of any explanation – "the representational gloss does not predict anything about the system's behaviour that could not be predicted by the dynamical explanation alone" (2009, p. 77). All that RL does is impose a confusing, clunky explanatory framework over the dynamics, which are already doing all the explanatory and predictive work.

What dynamicists must do then, is meet the challenge posed by those in the representational corner – provide good explanations for even those phenomena that appear to require representations. Whilst Chemero is optimistic about the ability of the dynamical stance to handle these cases, the proof must be provided.

What would this mature, anti-representational cognitive science look like? On this score, Chemero points to a model developed by Randy Beer (2003). Beer modelled an agent whose task was to collect circular objects while avoiding diamond-shaped objects. The agent's behaviour was controlled by a neural network that was refined using a genetic algorithm, they then analysed the behaviour, not just with reference to the neural controller but in two further ways. First they modeled the behaviour of the whole brain-body-environment system, and second the coupled dynamics of the agent-environment interaction. For example, analysis of the fully integrated system enables Beer (2003) to make accurate predictions about the number of collisions the agent suffers during its search. Modelling the coupled agent-environment dynamics facilitates precise predictions about when the agent will 'decide' to catch or avoid an object.

The tools of dynamical systems theory permit us to find these elegant mathematical models that indicate the key variables and parameters relevant to behaviours of interest. Using these we can construct a general model of the behaviour which is going to be a function of the nonlinearly coupled dynamics of the whole brain-body-environment system. However, we can also perform a decomposition of sorts by then subsequently zooming in and focusing on the way in which just the agent and the environment interact – that is, provide an explanation of the moment-to-moment behaviour of the system given the occurent states of the agent and the environment. Zooming further in, we can also isolate the dynamics of the agent's nervous system. By examining these, we can explain and predict the way in which the sensory states of the agent help modulate its behaviours. However, importantly, Chemero notes that these models are useful "only when combined with the models of the whole coupled system and the agent-environment dynamics. These three dynamical systems compose a single tripartite model" (2009, p. 38). The reason for this is that neural dynamics depend on the previous states of the environment, which can only be fully captured when the three models are taken together as a full explanation of the behaviour.

Thus, the dynamical stance provides a scientifically progressive way of explaining cognitive behaviour without recourse to representation or intentionality. Rather, by providing formal mathematical models at various levels of description we obtain far more precise predictions than we could with just a mechanical decomposition.

It goes without saying that the vision of the dynamical stance is a mere sketch. At this point my aim is to give a sense of the flavor of the theory. Recall the way that connectionist networks are described in terms of their dynamics, indeed the way Shagrir (2012) utilized DST to argue in favour of neural representations grounded in isomorphism. The interplay between connectionist and dynamicist cognitive science will be a theme I return to in greater depth in chapter eight when I will consider the strengths and weaknesses of the paradigm with respect to predictive processing.

3.4 Conclusions

Here I have outlined three important points of reference in the mental representation debate. Combined with the detailed discussion of classical and connectionist representation in previous chapters we are left with a fractured space of possibility. We might accept or deny that the mind is a kind of computer – denial leading to instead understanding the mind using DST and Chemero's dynamical stance (2009) that eliminates mental representation from our explanations. If we accept that the mind is a kind of computer, then we can ask whether it is classical or distinctly connectionist (that is, not merely a connectionist implementation of a classical system). Classical systems are on firm ground with regards to the symbol-grounding problem, thanks to Cummins (1996) and we might interpret these map-like structures as reconstructive models, or more action and body involving AORs; on the other hand, connectionists must work hard to justify that their symbols are properly grounded, i.e. that they are playing a standing-in role, per the demands of Ramsey (2007).

Both advocates of classicism and connectionism must tackle the trickiest problem of all, that of intentionality. We saw in chapter one that the two main contenders, the causal and teleological theories face serious issues such as the problem of error, the disjunction problem, and in the case of teleology – with grounding the notion of proper function. While I introduced Dennett's intentional stance as a third way of responding to the problem of intentionality, if it is possible to ground the notion naturalistically, that will be the most philosophically satisfying route to take. Thus in Part Three, with the details of predictive processing firmly established, I will attempt first to ground the notion of function, and then build a respectable notion of intentional content on those foundations. I will show that this notion is not exclusive to predictive processing models, but is especially amenable to it. Further my notion will help us understand how computational and dynamic explanations might be synthesized within cognitive science in order to construct full explanations of cognitive phenomena at various levels of description.

Part Two: Predictive Processing

Introduction to PP

Why do we, as a species, tend to have such great confidence in the knowledge generated by scientists? While this question has a venerable philosophical literature behind it, the main driver of our intuition seems to be grounded in the thought that the methods of scientists are somehow *maximally rational*, or at least *more* rational than other methods for generating knowledge. By contrast, consider knowledge arrived at by intuition. Beliefs generated in this way are notoriously unreliable, often leading us to distrust perfectly trustworthy people, and other times resulting in us falling victim to simple con-tricks.

The problem with intuitions is that they are based on very little evidence, and even when more evidence is offered in support of an intuition, it has often been gathered and interpreted in a biased way. One way to demarcate scientific methods from unscientific methods is by appeal to the standards required of the evidence, and the standards required of the collection and analysis of that evidence. In both collection and interpretation of results, objectivity is demanded. It is important to consider many possible hypotheses – different explanations of the evidence – and evaluate the plausibility of each upon receipt of a new body of evidence.

In order to avoid as much bias as possible, and thus ensure our evidence-based beliefs are as rational as they can be, we might use Bayes' Theorem to aid us with these evaluations. Bayes' Theorem is a mathematical formula which helps us to update the confidence we have in each competing hypothesis (measured as a probability value) based on the likelihood of the new evidence *given* each hypothesis and our prior confidence in each hypothesis. Once many pieces of evidence are presented, given that they represent a fair sample, and that we can correctly judge the importance of each piece, we will have a set of beliefs that has high confidence in the hypotheses that are correct and low confidence in those that are incorrect[9]. This is exactly what we want

[9] Indeed, even if our prior probabilities are wildly inaccurate to begin with, given enough evidence, Bayes' Theorem will correct them. This is known as *washing out the priors*.

from a rational method of evaluation, and indeed, Bayesian reasoning (the use of Bayes' Theorem to update beliefs) is the ideal to which scientific methods aspire.

Apart from bringing us closer to the truth about our domain of enquiry, it would also be very *useful* to be able to reason perfectly like this. Indeed, if the success of science is to be measured pragmatically, by the usefulness of the technology we can produce, then even if we don't consider our hypotheses to be bringing us closer to the truth, they can nevertheless suggest avenues for technological innovation. The usefulness of updating our beliefs using Bayes theorem or an approximation has inspired a research programme (see Knill and Pouget, 2004 for an overview) based on the possibility that evolution has furnished complex organisms such as humans with cognitive mechanisms approximating Bayesian reasoning abilities (Griffiths and Tenenbaum, 2006). Animals optimising their beliefs in this way gain an evolutionary advantage as they are better able to recognise regularities available in their ecological niche and exploit them. The principle that Thomas Bayes formalized mathematically, as a way of rationally pursuing knowledge, may be instantiated by our brains naturally, simply as a way of coping with the world.

Independently, a separate research programme led by the work of Karl Friston posits a single unifying principle to guide explanations of life and mind. This is the Free-Energy Principle or FEP. According to this principle all complex self-organising systems can be characterised as aiming to minimise their free-energy (more on what this is shortly). The simple idea is that there is a small subset of states a system must remain within in order to resist entropy and persist through time. Friston recommends framing the domain in terms of 'anticipation' and 'surprise' (or more correctly, surprisal): in other words, self-organising systems can persist in their ecological niche by anticipating the state of their surroundings and attuning themselves accordingly. Minimising the amount they are 'surprised' by their environment in the long term is a good strategy to maintain their viability. Free energy is the mathematical upper bound on surprise (to be explained further in chapter four) and so systems that minimise this variable in the long term will tend to survive longer than those that don't. The interesting aspect of this framework is that although we can analyse how a whole organism minimises its free energy, we can also consider individual elements of the organism to get a more detailed picture. When we analyse the central nervous system through the lens of the FEP we find a Bayesian machine emerges – that is, the best kind of system to minimise surprise, and thus free-energy the most effectively, is a system whose processes approximate Bayesian inference, with expectations generated by approximations to Bayes-optimal beliefs. FEP thus adds further theoretical weight to the possibility that brains are rational, scientific machines that make theories, and use evidence to evaluate those theories in a Bayesian way.

As we can see, two relatively independent research programmes indicate that there are excellent theoretical reasons to suspect that biological brains instantiate some form of probabilistic knowledge, regularly testing and probing the environment in order to update those beliefs in order to have a stronger chance of survival in their particular ecological niche. In this chapter I will elucidate the framework behind these ideas of a rational, scientific, experimental cognitive system which has become known as "predictive processing" (PP) (Friston, 2005, 2010; Clark, 2013, 2016; Hohwy, 2012).

I will begin with some technical details regarding the mechanism which will be germane for the following philosophical discussion of representation. There are many important aspects which will be dealt with in turn: the overall structure of these systems, which utilise a hierarchy of discrete layers to learn more effectively; the way that prediction and error signals are generated; how error is systematically minimised by the system, which is the key difference between this approach and others; the learning algorithms employed and the implementation and analysis of what is learned. Several case studies will be examined (e.g. Rao and Ballard 1999, Friston 2010) to illustrate how everything works together and the kind of success the PP programme has achieved.

The second part of this section will examine competing interpretations of the framework. Whilst excitement about the potential of PP is widespread in the philosophical community, the implications are not agreed upon. Thus far, debate is focused on what a successful PP paradigm would mean for extended, embodied and embedded theories in philosophy of mind. Given that PP is usually understood to be a modern incarnation of Helmholzian inferentialism (1867/1910), whereby a system infers the nature of the external world using only proximal stimuli, proponents such as Jakob Hohwy argue that PP is a strongly internalist framework. This would be a serious blow to the externalist movement that has been gaining momentum in recent decades. On the other hand, there are elements of PP that tie in very neatly with important theses of the embodied approach. Indeed, these elements are central to the theoretical structure of PP and are thus taken as indicators that PP is an exciting new framework to flesh out an extended mind hypothesis for action and perception. The outcome of this debate is linked, we shall later see, with our questions concerning the nature and status of PP's apparent use of representations.

The final part of the section will consider some of the evidence supporting PP. Though the programme is rich in theoretical development and the corpus of supporting empirical work is growing rapidly, there are elements still requiring confirmation and plausible neurobiological implementation which motivates scepticism and debate regarding the scientific viability of the framework. Further, it will become apparent in this section that PP is an extremely ambitious programme, with advocates such as Karl Friston suggesting that it is our best candidate for a Grand Unifying Theory (GUT) of cognition (Friston, 2010). It is yet to be decided whether PP is powerful enough to make good on this promise in any non-trivial way. Keeping concerns such as these in mind is vital to understanding the scope of what is being discussed in this thesis.

4. Mechanisms

It is certainly possible to go into great depth of detail when explicating how the models characterised as predictive processors actually work. In this section I hope to strike a balance that provides a level of detail slightly above what is necessary for the following discussion in order to facilitate a full understanding of the topic, but also to retain some generality in order that the relevance of these details to the argument is transparent and that one is able to see the defining features of the category 'predictive processor' plainly. The focus is to provide a working knowledge of the framework in order to inform the exclusively philosophical discussion in later chapters. To this end I will structure this chapter by dealing with more generally applicable details – the hierarchical structure and deep learning algorithms, before moving on to the prediction and error flows and error minimisation methods that define a PP model.

4.1 Hierarchical Models

To continue the analogy begun in the introduction, domains of scientific enquiry can be organised into a hierarchy of abstraction, or scope. Quantum mechanics provides us with predictions and theories about phenomena that happen very quickly and on very small scales. Chemists study phenomena at a slightly larger scale, both spatially and temporally. Biologists and geologists abstract further, concerning themselves with how chemicals combine to create larger, more complex objects and how they change over longer lengths of time. Cosmologists abstract the furthest, looking at the behaviour of extremely large objects over extremely long periods of time.

Our best artificial learning models (e.g. Hinton, 2009) mirror this hierarchy of abstraction by stacking layers of learning processes on top of each other. Each layer abstracting details from the layer below as the learning algorithms extract patterns in the data as presented at that layer or level, much like each set of scientists present in the analogy. The bottom processing layer, closest to the raw input, is a little like physics – it learns about the very quick dynamics of the atomic elements of the input, for example the patterns present in the way each rod and cone cell on the retina will activate over short timescales. Layers a little further along will abstract some patterns from that data, patterns that occur over slightly longer timescales, and between occurrent stimuli over a slightly extended field – e.g. finding patterns in how groups of rod and cone cells activate collectively, and how that pattern of activation will change over the course of a few seconds. Much higher up the hierarchy, processing layers will abstract out broad categories that might match on to common linguistic type-ascriptions and also be sensitive to the common ways in which stimuli evolves over long periods of time – e.g. top level learning layers may identify a certain visual scene as a picture of an elephant and may be able to categorise some sections of an audio recording as conversations.

Structuring a learning machine hierarchically helps to prepare that machine to learn in a specific way, constraining its dynamics so that it is quicker to extract and store high level regularities and abstractions that are useful to categorise its input. Rather than building a sprawling, one-layer network with many interconnections that effectively applies a learning algorithm once (for each item of training data), using a hierarchically structured machine, whose connections are constrained into a system of layered, virtual learners, the learning algorithm is applied many times per training stimulus, once for each layer in the hierarchy, with each layer learning about the layer below itself. It is this kind of iterative learning that results in the extremely accurate results that has fuelled the growing popularity of these systems in artificial intelligence research groups (Silver et al., 2016).

4.2 Prediction and error

In the previous section we saw the importance of layered processing as a means of facilitating learning in structured domains. This kind of machine is very popular, and provides a way for artificial cognitive systems to extract interesting regularities from the vast amounts of data beginning to be produced in the internet age. However, what sets predictive processors apart from their cousins is how they process information by making predictions and accounting for error.

In a predictive processor, the first processing layer above the input layer produces a prediction of what it expects the next stimulus to be. Before training of course, this prediction will just be random noise, but as the learning algorithm is applied over many thousands of training examples, these predictions will improve until the layer can anticipate the activity below with decent accuracy. All the layers above the input layer make predictions about the activity of the layer directly below. These predictions propagate downwards so they can be compared with the actual activity and also sideways so that neural populations on the same level can communicate and adjust their own predictions. Lower layers use the predictions of higher levels to inform and contextualize their own predictions, which are then in turn sent down to the next layer and sideways to their peers. This kind of process may account for phenomena such as bi-stable stimuli (e.g. Hohwy et al., 2008). The classical setup to induce visual bi-stability is for a subject to wear a pair of special glasses that present one image to the left eye, and a different image in the right eye. For example, the image in the left eye might be of a house, and the image in the right eye might be of a face. In this case, typical observers begin by perceiving the drawing as either a house or a face. This perception is fairly robust, for a short time, but then the percept will switch to the other image, and in turn that percept will remain robust. This flipping between percepts is surprising – we might expect to experience a hybrid 'house-face' percept, but instead the two percepts seem to be mutually exclusive. The PP explanation of this would cite the existence of high-level priors that there is no such thing as a house-face, so the probability of their being a house-face in front of the subject is much lower than the probability that one of the two images being received by each eye is wrong. However, the brain has no further information to consult about which eye it is that has the more reliable image, so hedges its bets by switching between the two, on one hand putting more stock in the reliability of the left eye, and on the other taking the right eye as the more reliable source.

Prediction signals are usually characterised as a downward flow of information, as they are sent from higher layers to lower layers and also thought to constitute a kind of 'top-down' process that can guide perception and action. There is also a reverse pathway, that corresponds to a 'bottom-up' flow of information. In predictive processing this information is characterised as ‘error’. High level predictions are compared to reality, and the mistakes are sent upstairs to be processed by the learning algorithm. Again, error messages are also propagated sideways to other populations on the same level which enhances the learning process. Mechanistically, error is a fairly simple concept – layer A sends down its predictions of the activity of layer B using a matrix that contains a value representing the predicted activation of each node in layer B. Then the actual activation values of layer B are subtracted from the matrix, and the new matrix is passed upwards and sideways as an error signal.

The error signal is aptly named because it returns just the *mistakes* in the prediction. These mistakes are the residual differences between the predicted activation of the input layer, and the actual activity of the layer. If the prediction of a particular node’s activation is correct (up to estimated noise levels), then no value is sent back for that node in the error matrix. This mechanism also yields the useful benefit that bigger mistakes in the prediction generate bigger error signals – for instance if the activation level 0.2V is predicted, and the actual activation is 0.98V, then the error signal will return the relatively large 0.78V value for that node. Layers receiving these signals may thus pay greater attention to bigger mistakes and work to correct them with more urgency than small mistakes. Although when elucidated mechanically, this is a job built into the learning algorithm of the system, it also offers an explanation for our phenomenological experience of novel stimuli, in particular the way they grab our attention – according to PP, our cognitive system is *surprised* by a new, unexpected stimulus and is struck by a flood of strong error signals that need to be corrected.

Error also provides a clue to the context of the subject’s situation. An animal’s cognitive system needs to have a certain flexibility in order to make accurate predictions in all the possible scenarios that occur day-to-day. For example, in order to predict visual activity properly, there needs to be some high-level expectation about the location of light sources. Outside, there is generally just one, fixed source, the sun, but indoors there may be several light sources such as table-lamps or ceiling spot-lights. Indeed, if you’re out in the evening, the lighting may be dynamic – moving and changing colour. In order to be effective then, the visual system needs to keep some prior expectations on stand-by, ready to be activated when the situation changes. When, for example, we step from a dark, lamp-lit side-street into a bright home with a fire roaring in the sitting room, our brain's visual predictions will suddenly suffer a tsunami of upward-flowing error signals. Potter et al. (2014) suggest that this influx may be used to rapidly determine context.

Together, the distinct downward flow of precision-weighted predictions and the upward flow of precision-modulated error signals are the central theoretical posits of PP. In this section I have given a brief explanation of how they function to deal with perceptual tasks, but many scientists remain hesitant to commit to the framework as a viable theory of neural computation before these flows of information have been empirically found to perform error and prediction functions in the brain.

4.3 Case study: Recognising natural scenes

The highly influential 1999 paper by Rajesh Rao and Dana Ballard, *Predictive coding in the visual cortex*, neatly illustrates the main ideas covered so far in this section. Rao and Ballard suspected that predictive processors modelling perception would display similarities with the observed behaviour of neurons in visual cortical areas V1, V2, V4 and MT. The feature that interested them was the phenomenon named 'end-stopping'. This refers to a certain pattern of neural activity whereby a neuron or population of neurons is typically excited by a stimulus, but that activity is suppressed when the stimulus extends beyond their classical receptive field (RF). For example, a population of neurons may become excited when a bar-like object appears, but if the image of the bar extends beyond the RF of that population, neuronal activity is suppressed. In order to test their theory, Rao and Ballard constructed a predictive processor and trained it to recognise a series of images, five photographs of natural scenes.

The network was organised into a simple two-level hierarchy, plus an input layer. Each node in the first processing layer had a narrow receptive field of 16 pixels by 16 pixels (px). These receptive fields were overlapping and each level two node shared connections with three level one nodes, resulting in each level two node having a receptive field of 16px by 26px. The image was processed by three level one 'modules', each responsible for processing a third of the image, divided vertically. Each module had 32 neurons to store a model of the input, 32 error neurons that sent error up to level two, and 256 prediction neurons to send predictions input to the input layer. Level two was just one module, and integrated information from all three level one modules. This top level had 128 feedforward input neurons from the level one modules, a further 128 neurons to store its own model and another 96 to transmit predictions of level one activity.

Figure 1 below visually demonstrates this structure. I, I_1 and I_2 represent the three sections of each image processed by the level one modules. PE stands for the predictive estimators constituted by the model-encoding neurons at each level. Arrows pointing from left to right represent the upward flow of error and arrows from right to left represent downward flowing predictions. Between level one and the input layer (level 0), the diagram represents the inhibition process, which is the way predictions are cashed out at the bottom level – predicted activation is mechanically inhibited and thus prevents error flowing back when the prediction is correct.

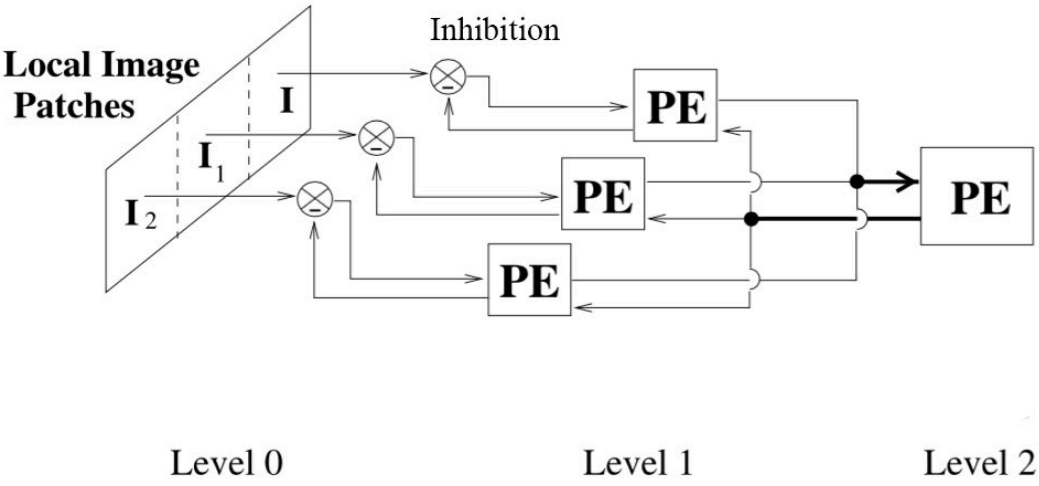

Fig 4.1: PP Schematic from Rao and Ballard (1999).

The end-stopping effect was measured by recording the error responses of level 1 neurons when presented with highly artificial, novel stimuli. The network was tested by comparing the level one error signal when presented with a short black bar image, with the level one error signal when presented with a longer black bar, which extended beyond the receptive field of a single level one node. The results showed that short bars resulted in significant error being transmitted, but negligible error being sent when the bar was longer. This is explained by Rao and Ballard as a result of the training the network on natural scenes, in which short bars are very uncommon, but long bars are not (e.g. trees, grasses, striped animal coats). Thus, level two would not predict the short bar stimulus, resulting in a strong error signal. This finding matches up with the empirical data mentioned earlier, in which real neural responses were suppressed in certain populations of neurons when objects extended beyond the receptive field of that population. Rao and Ballard's hypothesis is that those areas are responsible for propagating error to cortical layers further up the hierarchy which have a larger receptive field and do not expect very small visual elements.

One mechanical detail that will be important for our discussion of representations in part three is the precise way in which predictions from higher levels influence the activity of lower levels and thus the predictions that they pass down in turn. In this case study, predictions from level two play no role in the learning algorithm that determines the connection weights in level one, so there is no direct long-term effect filtering down. Rather, the top-down signals influence the occurrent activation levels of neurons in level one. This way, once the network has been trained, predictions from on high can have an instant but transient effect on the behaviour of lower levels, which prevents them disrupting or contaminating the models below, preserving their ability to perform their important functional role: picking out low level regularities in the stimulus.

Error on the other hand is a key variable in the system's learning algorithm. In the following section I will examine the mechanics of error minimisation in PP networks, and touch on lessons we can learn from Rao and Ballard's implementation before illustrating with further examples.

4.4 Error minimisation

One of the most alluring features of PP is the way it subsumes whole categories of cognitive function into a single, system-wide goal: minimisation of prediction error. So far I have described the structure and mechanisms of PP systems, as being ordered hierarchically and featuring a downward flow of predictions and an upward flow of error. I have also mentioned a few examples of the sorts of tasks these kinds of systems are capable of mastering. However, whilst on a computational level we might speak of these systems as learning to recognise handwriting or categorise images of natural scenes, algorithmically they all aim to minimise the long-term upward flow of error. That is, the learning and prediction generating algorithms are designed to tend towards accuracy in the long run.

The long-term viewpoint is specified in order to account for phenomena such as the Muller-Lyer illusion – a cognitively impenetrable visual illusion, the explanation of which invokes the existence of deep prior beliefs regarding perspective. Whilst our visual system is never able to properly account for the true nature of the stimulus in this case and others like it (see the checkerboard illusion), we recognise that these small, inconsequential mistakes are a reasonable price to pay for the huge gains in processing efficiency in other situations where the same priors are at work.

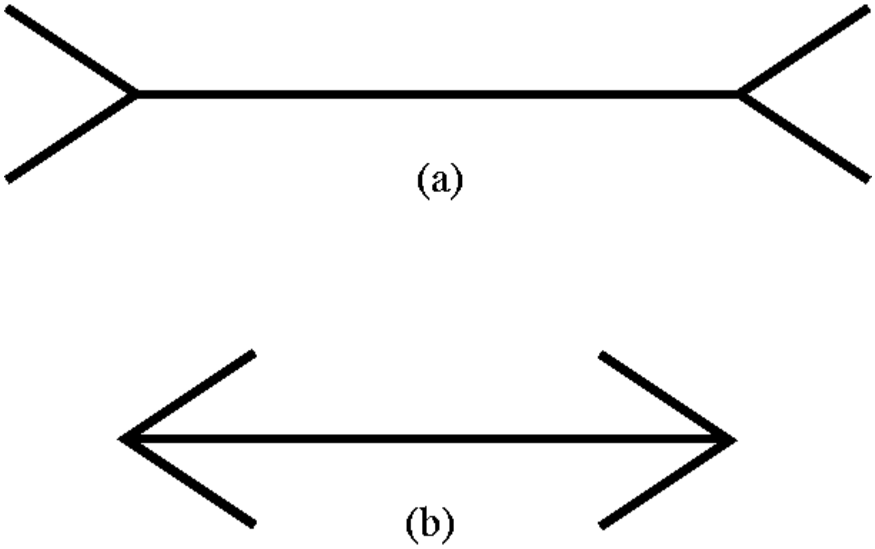

Fig 4.2: Müller-Lyer illusion. The horizontal lines in (a) and (b) are of identical length.

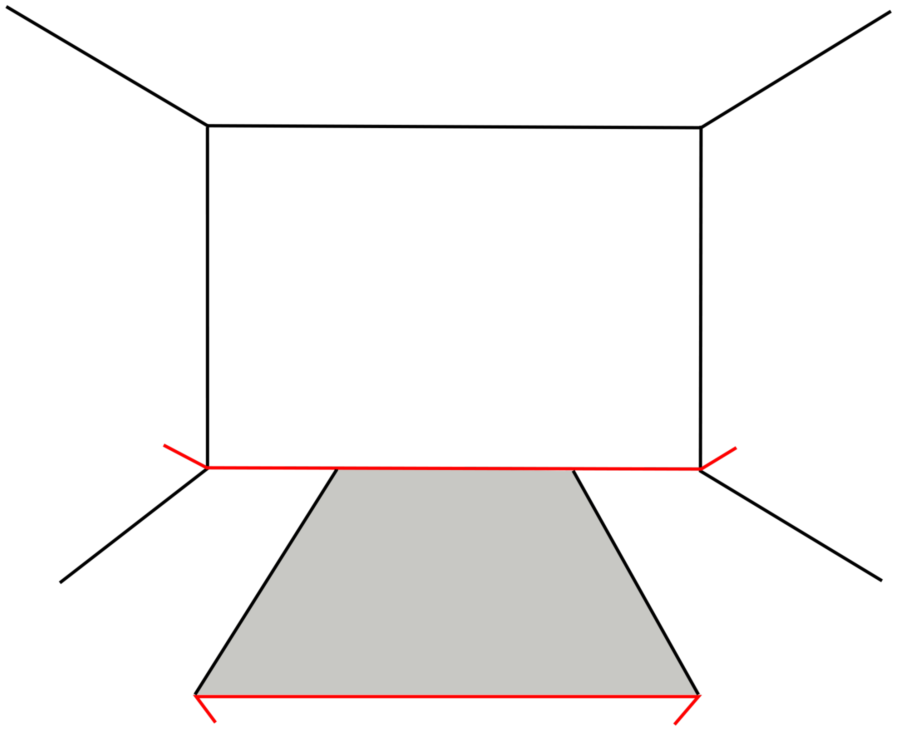

Fig 4.3: This diagram offers a visual explanation of the perspectival priors at work in the Müller-Lyer illusion. Though the horizontal portions of both red lines are of equal length, the one which appears to be further away, also seems longer due to the angle of adjacent lines.

Gary Lupyan (2015) makes the detailed case that failures in these niche circumstances are a very small price to pay for an otherwise well-functioning visual system. While one might argue that a better visual system would see the two lines as the same length, even in the optical illusion, and especially once it has been proved to the observer that they are in fact equal, this improvement would come at the cost of more drastic perceptual failings in more ordinary circumstances. One can imagine any number of tragic scenarios that could result from precepts related to the processing of horizontal lines with perspectival context, e.g. a driver failing to estimate the width of an oncoming lorry's chassis on a dual-carriageway. It is evident that there are sound evolutionary (and developmental) reasons for our brains, in some cases, to rely on extremely firm, all but immutable priors.

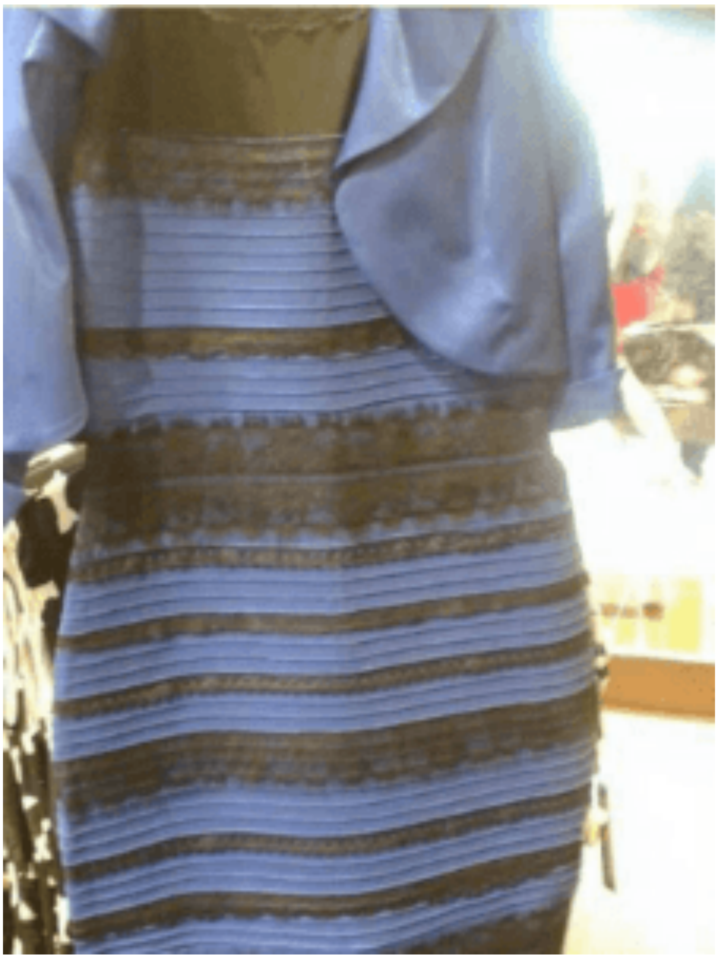

Fig 4.4: “The Dress” viral phenomenon. Image from Wikipedia.org.

A further enlightening example originated as a viral image online, and became an instant sensation (reaching UK national news) due to the striking differences between subjects’ percepts. By some people, the dress is seen as white and gold, while others perceive it as blue and black[10]. It has been hypothesised that the lighting context is suitably ambiguous so as to fail to trigger the correct priors in some people’s visual systems (Winkler et al., 2015). The suggestion is that there are a set of priors for outdoor lighting, and another set for indoor lighting. With the ‘outdoor priors’ the dress seems white and gold, but with the correct priors it seems blue and black (the photo

[10] If you have not seen this illusion before, it may seem inconceivable that such radical differences in perception are possible. If so, show the image to a few friends and you’ll soon discover that in certain circumstances, people see the world very differently. It was later revealed by the source of the original photo that the dress is, indeed blue and black.

was taken in a department store changing room). What is remarkable about this case is that it is cognitively impenetrable. Those who see the dress one way cannot see it the other way even once the alternative is articulated to them, and the phenomenon well-understood.

The way error is minimised in the Rao and Ballard (1999) paper discussed earlier indicates another benefit of the PP approach. Minimising overall system error, Rao and Ballard write, 'is equivalent to using the *minimum description length principle*, which requires solutions to be not only accurate but also cheap in terms of coding length.' So the optimization algorithms that work to minimize error in their case study, though ostensibly designed to minimize error, will also have the effect of ensuring that the system's models do this in the most parsimonious way possible – i.e. using a model with the fewest parameters capable of predicting the data accurately.

In the previous sections I have hinted that the system works to minimise error by learning how to predict the incoming sensory manifold with greater and greater accuracy. However, the goal of error minimisation presents a second means to fulfilment, at least for systems with any control over their own movements. For these systems, with a capacity for action, incoming signals can be brought into line with expectations. For example, my visual system might expect everything in my visual field to move to the right, and one way the accuracy of this prediction can be ensured is to simply move my head to the left and keep my eye position fixed. This kind of error minimisation has become known as *active inference*, mainly through the work of Karl Friston.

Active inference can also be exploited by the motor system, thus pressing PP into service for yet another cognitive faculty – action. Predictions about proprioceptive data can signal muscle movements that bring about that data. For example, my motor system could issue a prediction about the proprioceptive signal related to muscular tension in my shoulder, bicep and forearm to initiate a typical lifting action. The neural structure of the motor system has been demonstrated to mirror the perceptual systems (Clark, 2013; Friston, 2003) and the principled basis for predictions playing this kind of role has also been established and extensively illustrated by Adams, Shipp and Friston (2013) who state that "… under active inference, descending signals do not enact motor commands directly, but specify the desired consequences of a movement. These descending signals are either predictions of proprioceptive input or predictions of precision or gain." (Adams et al., 2013). Though I will be going into more detail on the nature of precision in PP systems in the following section, the salient point here is the way predictions are pressed into duty as motor commands. Though they do not specify the necessary movement, they specify the expected proprioceptive state, and the error signals resulting from these predictions effects the predicted movement. This brings the proprioceptive signal into line with the expectation, thus eliminating any further upward propagation of error. These systems are compatible with certain dynamical systems approaches to motor control, for example Feldman and Levin's (2009) equilibrium point hypothesis.

4.5 Precision

Whilst a great deal is possible using just the resources outlined so far, the addition of precision weightings increases the power of PP systems by several orders of magnitude. PP architectures that include a precision calculation not only process the error and prediction, but also assign a numerical value to the *confidence* or value the system assigns to those signals. An error signal that the system assigns high precision to will have a greater impact on the system's hypotheses than a less salient signal. Similarly, a prediction that the system is more confident in will be more resistant to sensory counter-evidence than a low-certainty prediction. For example, in a dark room our visual stimuli will be processed with a low precision weighting, and the precision of auditory and tactile stimuli may be heightened to compensate.

Mechanistically, the processing of precision has a well-established basis in practice. Precision is the estimated inverse variance of the error, and is calculated in parallel with prediction, and is assigned in accordance with the system's generative model. This mechanism allows the system to pay greater attention to some parts of its input, whilst suppressing others that it judges to have lower importance. For example, if there is a rustling in the bushes, we may not know what is in there, with several hypotheses with near-equal plausibility, thus the prediction about what will emerge from the bushes will have low precision, causing any error signals emerging from that part of the visual field to have very high salience, whereas areas of the visual field expecting grass-like appearances will have high confidence, so errors in those areas may go ignored – they are treated as ‘noise’ rather than ‘signal’.

When error messages are treated as noise, the system is overriding data from the external world with its own model. It is suggested that this mechanism may offer a robust explanation for the phenomena of hallucination and dreaming. It is hypothesised that the chemical imbalance in the brain is resulting in higher precision weightings being assigned to downward flowing predictions, which in turn leads to perception being overtaken by imagination (Fletcher and Frith, 2009; Corlett et al., 2010; Hobson and Friston, 2012).

The brain’s ability to attend more or less closely to certain stimuli is also explained neatly by the inclusion of precision in PP, and this is the focus of several studies, e.g. Feldman and Friston (2010) and Friston (2010). These case studies note that modulation of precision is governed by the top-down processes of the system’s generative model in tandem with other downward-flowing predictions. But over longer time-scales, parameters influenced by the system’s learning algorithm also affect precision weighting, allowing the system to adjust its precision expectations based on past experiences. Precision estimations are thus learned in the same way as other predictions, and function as priors over subsequent processing.

4.6 Summary

The aim of this chapter was to provide the reader with a moderately technical introduction to predictive processing. I have described a class of systems that are capable of autonomous learning; that are inherently active, and use action to guide their inferential process; that are flexible and context-sensitive; that employ parsimonious models involved in a constant cycle of expectation and error processing. The driving force behind these features is prediction error minimisation.

Some areas of philosophical interest will be noted in chapter five, and chapter six will summarise the current neurobiological evidence that gives us reason to believe that PP may be an accurate model of human and animal cognition.

5. Internalist and Externalist Approaches

Despite the extensive technical progress that has been made in order to bring PP to the forefront of cognitive research and modelling, and despite agreement among a cadre of philosophers as to the theoretical importance of the framework, the philosophical implications are yet to be settled. Whilst the main aim of this thesis is to consider one of these implications, the question of mental representations, it is beneficial to consider some of the other debates, as they will inform important aspects of our discussion. Indeed, if we decide one way for mental representation, that conclusion might entail certain positions on other questions, and if we take a stance on some other problem, that might influence our conclusions on mental representation. In this chapter I will focus on the implications PP has on the extended mind debate - Jakob Hohwy strongly suggests that PP is incompatible with the embodied cognition (Hohwy, 2013, 2014), instead offering a compelling account that explains how our self-contained nervous system is able to understand the world around it, despite there being a clear barrier between the two. On the other hand, Andy Clark (2013, 2015, 2016) sees PP as furthering his early work on embodied and extended cognition in the way it presents a thorough mechanism for understanding cognitive processes that are necessarily action-involving and take advantage of the body and environment to facilitate cognitive short-cuts. I will also discuss the work of Karl Friston, arguing that although he advocates a broadly internalist picture, that this is a mistake, and that in fact, his work supports a radically extended paradigm, and blurs the distinction between cognition and other kinds of biological activity.

5.1 Hohwy's Internalist Predictive Processing

> 'PEM [prediction error minimisation] reveals the mind to be inferentially secluded from the world, it seems to be neurocentrically skull-bound than embodied or extended, and action itself is more an inferential process on sensory input than an enactive coupling with the body and environment.'
> (Hohwy, 2014)

The first philosophical gloss I will examine is produced by Jakob Hohwy. As the quote above demonstrates, Hohwy believes that PP takes cognitive science back to a ‘neurocentric’ perspective that he presents in direct opposition to embodied and extended approaches to the mind. He argues that PP offers a complete and coherent account of mentality, appealing only to states of an internal generative model and the inferential processes therein. Even processes of active inference are properly ‘mental’ insofar as they result from the internal, neural states. In this section I will provide a little background on the philosophical tradition and examples that motivate this view, and then go on to elaborate on Hohwy’s position that the predictive brain is a ‘fragile mirror of nature’, enclosed by a well-defined ‘evidentiary boundary’ which gives us firm, principled reasons for rejecting embodied and enactive cognition.

Helmholtz’ Inference Machines

Predictive processing reverses the way we understand perception. The intuition that our sense-organs are bombarded with signals from the world and that our brain pieces together this puzzle to form perceptual representations has guided cognitive science for decades (e.g. Marr, 1983). If PP is a correct understanding of the way perception works, the puzzle begins completed, and we check the pieces to make sure they’re in the right place, and if not, shuffle them about or replace them, making better and better guesses about where each piece should be. Over time, through repeated cycles of hypothesis, prediction, error and inference, the picture (the generative model) becomes more and more accurate. However, this idea is not new, with Hermann Helmholtz making very similar claims as long ago as the 1860s. Through processes of probabilistic estimation and revision he called unconscious inference, Helmholz believed that the mind constructs an internal model in much the same way PP is suggesting today (Hatfield, 2002).

Helmholtz drew parallels with science, suggesting that the mind probed the world scientifically, and reasoned about it scientifically, making inductive inferences and testing its predictions through action. The PP models, built by Rao and Ballard and Friston, discussed above, demonstrate how machines and brains can perform the unconscious inferences resulting in a flow of predictions about the world that can be taken as constitutive of perception.

Hohwy takes the inferential talk inspired by Helmholtz extremely seriously. Neural correlates of PP networks deal in representations that track multi-layered features in the world, generating inferences about the probability that any given object or scene is 'out there' moment-by-moment. Whichever represented hypothesis is assigned the highest probability gets predicted, and so, perceived. These predictive processes, the learning processes and active inferences that result in error minimisation are all that constitute the mind for Hohwy, as they are all that is necessary for PP to explain cognitive phenomena. This results in his strong commitment to a mind bounded by the sensory peripheries, limited to the physical structures making up the nervous system.

The Leaky Dam

To help us get a better grip on Hohwy's view, I will now summarise an analogy he makes himself to aid his readers' understanding of the framework. While the example is illuminating in the way it captures the basics of PP, it is doubly illuminating in the evident way it reflects Hohwy's internalism.

> 'Imagine being charged with plugging holes in a large, old and leaking dam… the occurrence, frequency, and nature of the leaks all depend on the water pressure on the other side, the water levels, the consumption on this side, the state of repair of the dam, and so on'.

Hohwy (2013) presents a dam as our sensory periphery, with water leaking through playing the part of incoming stimuli. Our only job is to minimise the amount of water leaking through (i.e. minimise upward flowing error), and though we have no knowledge of the nexus of causes behind the dam wall, we are given a limited set of tools to patch up the wall as best we can.

Over time, Hohwy suggests, we would begin to notice simple patterns in the leaks. Perhaps a big leak in location A is often followed by a series of smaller leaks in location B, we could anticipate the leaks at B and patch up the wall before they sprung, whenever we noticed the tell-tale gusher at A. As we built up knowledge of several patterns, we might build a machine to respond to these patterns automatically, lightening our load a little. With our hard-earned free time, we might begin to notice larger patterns in the way the machine responds – whenever pattern X is halfway through execution, pattern Y initiates – we can then fine tune the machine to make these patterns of response more efficient. Furthermore, the longer we do the dam-plugging job, the more patterns we might notice in the leaks themselves. Perhaps there is an annual cycle affecting the number, position and severity of leaks, maybe a year of light leakage is typically followed by a heavier year. Recognising all these patterns allows us to further refine our anticipation of leaks and allow us to improve our machine.

The key contention is that our machine begins to represent the underlying causes of the leaks. Short term response patterns might represent typical undercurrents in the local area, day-to-day variations come to account for tidal effects, and at a higher level seasonal and macro-meteorological factors come to be represented. But this rich model of the world on the other side of the dam emerged only from our guiding influence, to minimise the amount of water leaking through. Howhy (2013) notes 'the crucial bit, however, is that in achieving this successful representation of the causal structure of the world beyond the dam, you didn't have to *try* to represent it' (original emphasis).

The internalist intuitions are well and truly pumped by this analogy. Through an ongoing process of leak minimisation, the machine begins to mirror the hidden causes of those leaks in order to better anticipate them. This mirror is hierarchical, allowing it to make certain associations at several levels of remove from the cold face of the dam wall, and model patterns taking place over extended periods of time. However, Hohwy acknowledges that the example provides no way for the dam-plugger to act on the world behind the dam, but eludes that this wouldn't present too much of a change to our understanding, with the machinery behind the dam remaining largely the same, the only addition being the ability to '[change] the location of the dam and the direction of the flow of rivers and so on' (Hohwy, 2013).

The analogy also fails to explicitly account for the role of precision in PP, as decisions about where and how to plug the dam do not involve processing confidence in each hypothesis or potential error'. This will become especially relevant when considering Andy Clark's opposing view which stresses the highly context-sensitive, 'quick and dirty' strategies modulation of precision affords.

Fragile Mirror of Nature

So far I have presented a few ways in which to understand the angle Hohwy wants to maintain on PP. In this section and the next, I will expand in more detail on the content, implications, and arguments for this angle. First I will summarise the arguments put forward in *The Predictive Mind* (Howhy, 2013), that PP networks are 'fragile mirrors of nature', and attempt to build representations of the world in a truth-seeking way. The following section expands further on this theme, drawing from Hohwy (2014), in which Hohwy seeks to establish that the PP paradigm entails a definite split between mind and world, an evidentiary boundary. He claims that there is no way to weaken this boundary and still maintain the core elements of PP, viz. a strongly representational model of the world and the process of Bayesian inference by which it is updated. This brings with it the implication that minds are vulnerable to radical Cartesian scepticism, which Hohwy claims is a direct consequence of PP.

The story so far has done much to establish the way a generative model mirrors the hidden causal nexus whose elements combine non-linearly to impinge on the system's sensory periphery. Through repeated and ongoing Bayesian inference, the generative model represents a rich structure of nested hypotheses about the next input for each sense modality. Each hypothesis will be weighted according to precision, and both the content of those hypotheses and the confidence assigned to them shifts over time, according to the feedforward flow of error. In a system operating well, that is, minimising error effectively in the short and long run, we can expect that the set of hypotheses assigned high confidence by generative model will be more or less accurate, that is, they will possess good representational schemata.

This mirroring, will of course be species-relative, Hohwy concedes. He notes that 'a given organism tends to sample the sensory evidence that defines its phenotype'

(Hohwy, 2013, p. 53). Here Howhy acknowledges that the set of states that minimise prediction error in an organism will vary from species to species, and following Friston (2010), that those states actually serve to define that phenotype. Nevertheless, the organism's model will mirror the environment in a way that tracks truth:

> 'the mind is a truth-tracker in the sense that it is a mechanism that is optimized such that it reliably recapitulates the structure of the world. However, the nature of the mechanism is such that when its normal conditions are tampered with, for example in the shape of input that is surprising relative to the states the organism is expected to be found in, then it stops reliably recapitulating the world' (Howhy, 2013, p. 230).

It is in this sense that Howhy understands the system's generative model to mirror the world, that is, only when the system is in a (narrow) set of states in which it is comfortable, which it already *expects* to be in.

The mirror is 'fragile' because it tends to get things drastically wrong when the organism strays into an unexpected situation. One of Hohwy's favourite examples is the 'rubber hand illusion'. This strange effect occurs when a subject's arm is placed *under* the surface of a table, and instead a rubber arm is placed on the table. An experimenter then touches the subject's hand and the rubber hand simultaneously. Subjects typically feel the sensation 'in' the rubber proxy, with the strength of the illusion varying depending on certain factors. For instance, the illusion still occurs, but to a weaker degree, when the rubber arm is replaced by completely different object, indeed some subjects experience the illusion when there is no proxy, and the experimenter just taps on the table surface (Armel and Ramachandran, 2003).

The rubber hand illusion is an especially strong example of the fragile nature of our world model as it concerns our own body. We often consider knowledge about our bodies to have a privileged epistemic status, insofar as we have high confidence in it. Furthermore, our body-model is central to our ability to perceive and act in the world. The rubber hand illusion 'flies in the face of what must be a very strong, deeply embedded prior belief about your body, namely that when you can feel a touch it is delivered to your own body and not to what ought to be an inanimate object' (Hohwy, 2013, p. 230).

Of course, according to PP, the illusion is the result of the brain doing its best to minimise prediction error by assigning high precision to the incoming visual stimuli that contradicts the body model. This demonstrates, Hohwy argues, that though the brain is *primarily* concerned with minimising prediction error, the *way* it does this (viz. by typically assigning high precision to error), is fundamentally 'truth-seeking' (Hohwy, 2013, p. 226). This is important to note for the forthcoming discussion about representational content, and the degree to which we can say that mental representations are capable of being true or false, correct or incorrect. It is clear that Howhy believes that states of a generative model aim at truth, and can get the world right or wrong. If this turns out to be an untenably bold position, we might balk at the bullet we have to bite in order to swallow Hohwy's 'fragile mirror of nature' interpretation.

Evidentiary Boundaries

A mirror requires a surface in order to reflect incoming light and, not to stretch the analogy too far, Hohwy believes that adopting PP requires the acknowledgement of a similar boundary which keeps the mind, and cognition, firmly 'in the head'. He maintains that a consequence of PP is that perception is fundamentally *indirect*, contrary to the position of most enactive and embodied cognition theorists that we directly perceive objects and properties 'out there' in the world. If Hohwy is correct, then the predictive mind is strongly separated from the external world by an 'evidentiary boundary', and perception is constituted by our best predictions about what is behind it.

The primary argument for the existence of the evidentiary boundary appears in Hohwy's *The Self-Evidencing Brain* (2014). The inferential processes used by PP are compared with inferences to the best explanation. The most successful hypotheses of a generative model are those which 'explain away' the most incoming data, thus stemming the upward flow of error signals. In this sense, the hypotheses of a generative model are 'self-evidencing' as described by Carl Hempel (1965). That is, if a hypothesis, *H*, successfully explains away some phenomenon, *E*, then *E* becomes evidence for *H* (to the extent that *H* explains *E*). This kind of reasoning is prevalent, but also risks circularity, if our justification for believing in the occurrence of *E* is not independent of our assumptions about it occurring. Hohwy suggests that the inferences made by generative models do risk this kind of circularity, and that the possibility of circularity in these cases rests on the existence of an evidentiary boundary – with the hypothesis (*H*) and data about the phenomenon (*E*) on one side of the boundary and the hidden causes of *E* on the other. The causes of the phenomenon can only be inferred. The generative models of PP systems are shaped by input, constantly evolving, modifying its hypotheses to fit the data, and thus 'it maximises the evidence for itself' (Hohwy, 2014).

Not only does Hohwy embrace the circular reasoning of PP, he also embraces the more radical epistemic consequence of this view, Cartesian scepticism. With perception hidden behind a veil of transduction, the evidentiary boundary, the possibility that all our sense data is being manipulated to preserve our illusion of the world becomes very real. Embodied and enactive cognition rejects or blurs the mind-world boundary, allowing for direct perceptual contact with the external world, thus banishing Descartes' evil demon from the roster of serious philosophical problems to be solved. Hohwy thus concludes that 'scepticism becomes the canary in the coal mine, indicating whether an account of mind and cognition is orthodox and internalist, or not' (Hohwy, 2014).

I will defer my discussion of whether PP really does entail a strong evidentiary boundary in principle until after I have summarised the view of Andy Clark, who pursues a line in direct opposition to Hohwy's internalism. However, we can at least ask Hohwy *where* to draw the evidentiary boundary. Hohwy maintains that the boundary is the organism's sensory periphery, which separates its nervous system (the generative model) from the rest of the world (including the organism's own bodily states). *Prima facie*, there seems to be some principled basis on which to extend the model to encompass the whole organism, and thus the boundary to the outer layers of skin. Friston (2010) seems sympathetic to the notion that the whole organism embodies a model of its world, in the way that the whole system must work together to resist the second law of thermodynamics by minimising free energy. There is also perhaps a more effective, but less important argument that the evidentiary boundary could be pushed inwards, and that in fact we could draw it at any layer of the hierarchy, taking that layer as the evidentiary boundary for the higher cortical areas. Though I won't pursue this line of argument, it does demonstrate that the positioning of an evidentiary boundary is strongly interest-relative, which justifies our strong demand for Hohwy to provide some principled reasons his proposed frame of reference.

Thus we find Howhy defends his thesis in terms of explanatory interest: 'the agent worthy of explanatory focus is the system that in the long run is best at revisiting a limited (but not too small) set of states. It is most plausible to think that such a minimal entropy system is constituted by the nervous system of what we normally identify as a biological organism: shrinked agents are not able to actively visit enough states, and extended agents do not maintain low entropy in the long run' (Hohwy, 2014). Hohwy's reference here to 'extended agents' is in the context of Clark and Chalmers' (1998) Otto's notebook thought experiment. Thus, it is not at all clear that Hohwy's claim does rule out the possibility of his principle applying more fruitfully to *embodied* agents, in the sense just discussed – organisms embodying a generative model of their environment. We might also call into question his assertion that these 'minimal entropy systems' are the agents worthy of explanatory interest. Indeed, the lessons of embodied, extended and enactive cognition give us good reasons to think that our scope should be wider, so Hohwy's argument here begs the question against such positions. In chapter nine, we will explore the possibility that in order to fully understand how a system minimises entropy, we must to take the extended frame of reference.

5.2 Clark's Radical Predictive Processing

While Hohwy's richly reconstructive, internalist interpretation of PP is the intuitive philosophical response, Clark (2015) offers us an alternative, embodied and enactive slant which he argues aligns more closely with PP once the implications of certain, key elements are acknowledged. However, Clark falls just short of committing himself to a fully enactive, non-representational view of PP. Rather, he claims the nature of representation is far leaner than Hohwy suggests, with these systems being heavily world-involving and world-creating in the way they seek to maintain their structural integrity over long periods of time. While Hohwy appeals to analogies of being behind a leaky dam, Clark highlights the way some robots take advantage of their bodies' passive dynamics, and drawing links with Gibsonian ecological psychology, construing PP networks as constantly and actively working to maintain a grip on their environment using 'quick and dirty' methods in processes he describes as 'productive laziness'.

Quick and Dirty Cognition

> 'In a certain sense, the brain is revealed not as (primarily) an engine of reason or quiet deliberation, but as an organ for the environmentally situated control of action. Cheap, fast, world-exploiting action, rather than the pursuit of truth, optimality, or deductive inference, is now the key organizing principle.'
> (Clark, 2015)

The quote above from *Radical Predictive Processing* succinctly summarizes Clark's analysis of recent work on action-oriented cognitive processes which serves to illuminate everyday ways our brains keep a grip on the world. Motivating this controversial claim is a large corpus of experimental work spanning robotics (passive walkers), perceptual psychology (eye-tracking studies), and developmental psychology (sensorimotor involvement in concept acquisition). As these examples are instructive, I will briefly summarize their main points here so we can get a better sense of Clark's perspective.

The ability to walk may not sound like a classic problem for cognitive science, but in fact it is incredibly difficult to build a machine that walks naturally and with high energy efficiency. The Japanese robot ASIMO, created by a team at Honda is a humanoid robot that is able to walk and run, with a fairly natural gait, but expends huge amounts of energy and computational resources to perform the feat. Clark compares the capability of the energy intensive ASIMO to 'passive walkers', which have no internal computer. These machines are instead constructed so that when they are placed on a small inclined plane, they gracefully descend with a simple walking motion that takes advantage of the natural dynamics of the construction and depends on characteristics of the materials used for success, e.g. the friction between the 'feet' and the walking surface. Given that our brains evolved in tandem with our bodies, the argument goes that the most efficient use of cognitive resources is not to pre-program every detail of motion, but to piggyback on the existing dynamics, spreading the computational load across both brain and body.

Ballard et al. (1997) engage with a more recognizably cognitive phenomenon – a visual sorting task. In this experiment, subjects were asked to move coloured shapes across a computer screen using a mouse to recreate a pattern. The experimenters used eye-tracking technology to observe the saccadic movements of the subjects' gaze. Rather than discrete processes of memorization and reconstruction, Ballard et al. found that subjects constantly referred back to the original pattern, typically several times during the movement of each shape. To account for this they hypothesize that the brain is employing a highly efficient computational strategy that minimizes the amount of information stored in working memory at any given time, offloading the necessary processes, where possible, into the task-domain itself. In other words, subjects needn't remember the colour, shape and orientation of a block all at the same time, but can check and re-check each requirement many times directly, using their eyes. In this case the brain negates the need for a rich, reconstructed mental model to generate complex inferences and subsequently guide action, instead exploiting its capacity for action to maintain an energy-efficient grip on its task.

Some of the most compelling work demonstrating the brain's quick and dirty cognitive abilities is produced by developmental psychology. The theme of distributing the cognitive load between brain, body and environment is echoed in research by Lungarella and Sporns (2005), who show that concept acquisition in infants is aided by continual active engagement with their surroundings. They create multi-modal experiences through self-generated action that serves to build up sensorimotor know-how, speeding up and enriching the infants' learning processes. These results cohere with a PP model of neural processing, which leads us to expect the system to build up detailed, action-involving generative models of the world, but with the emphasis very firmly on the active and embodied part of that process, rather than the strongly inferential, computationally expensive model suggested by Hohwy.

With these examples, among others, Clark motivates his alternative conception of the PP paradigm – a parsimonious, enactive, world-involving, computationally distributed understanding of how the human brain maintains a homeostatic grip on the world.

Two arguments against reconstruction

However, these examples don't provide any concrete reasons to suppose that PP is friendly towards this parsimonious, 'quick and dirty' formulation. Rather, they constitute empirical evidence of the kind of lazy cognition that Clark argues for, thus establishing its pedigree as a scientifically respectable notion. Indeed, they are *compatible* with the PP framework, but nothing has yet been done to establish that PP systems necessarily exploit quick and dirty methods, in the way that Hohwy's formulation offers a convincing account of the reconstructive, rich model-building nature of PP. However, Clark does put forward two arguments drawing from details of the mechanisms of PP networks, which serve to establish his claim more firmly.

I touched on the first of these arguments in chapter five, when I outlined Rao and Ballard's error minimisation constraints. A feature of error minimisation in PP systems is that this process not only maximises the accuracy of predictions, but also minimises the complexity of the generative model. Clark finds reference to this also in Fitzgerald et al. (2014). The reason for this *prima facie* unexpected relationship is the way that error in a PP system is cashed out mechanistically. Rao and Ballard (1999) note that '*E* [where *E* is their formal notation for the sum of squared prediction errors at each level in the network, weighted by inverse variances] can be interpreted as representing the cost of coding the errors and parameters in bits'. Karl Friston (2010) provides an extensive mathematical formulation proving that the free-energy is equal to the accuracy of a generative model, minus its complexity[11]. Rich models have many parameters which are expensive to maintain, and this cost is realised in the overall calculation of error. 'Error' minimisation thus pushes the system towards an optimal *balance* between parsimony and accuracy.

So perhaps the long-term error minimised by PP systems is better understood as 'overall system costs', and we can break this down the way an economist might. In business terms, we can understand the cost of the model parameters as long-term overheads of making any predictions at all and upward flowing error signals as the direct unit costs of making each individual prediction. The directors of a company may minimise costs in either way, but do well to maintain a balance that will minimise overall cost in the long term, just as an error minimising system does. In the case of PP, this means making a trade-off as discussed, between detailed and accurate models with a low volume of upward flowing error, but a high operating cost in the form of a rich model with many parameters; and more parsimonious models that suffer greater amounts of error flowing through the system. Clark rightly highlights this aspect of the PP mechanism to demonstrate that these systems are 'in no way committed to the conservative (richly reconstructive) reading that would render it incompatible with [productively lazy] solutions. On the contrary, one of the fundamental principles of PP renders it opposed to that reading' (Clark, 2015).

[11] Friston (2010) relates this to the better known 'infomax principle', of which the FEP is a generalisation – where the infomax principle applies when there is no uncertainty in the data, and it is represented using point estimates of the causes, the FEP applies to probability densities of unknown causes. The lesson for us is the same however – the mathematical principles underlying PP ensure that the generative model is coded efficiently.

Clark's second line of attack is to draw our attention to the action-involving nature of predictive processing, which I hope has already been made somewhat clear in preceding sections. Hohwy pays lip-service to this important dimension of PP, yet remains wedded to a firmly internalist picture. The inferential, neo-Helmholzian view of PP is strongly suggestive of a 'passive perceiver' understanding of our sensory systems. However, the notion that an organism is thrust into the world and bombarded with inputs and tasked with gradually inferring the hidden causes of the data captures just one way these systems might work to detect salient features in their environs. The other way is to act to bring the flow of sensory stimulation into line with expectations, and the fact that PP is itself a mechanism for unifying perception and action gives us a powerful tool for conceptualising how this works. The same system that minimises error by inferring hidden causes is also able to, instead of inferring hidden causes, thrust itself out in ways that minimise error (and trim its model parameters) even more effectively. The case studies Clark points to at the beginning of the article are just a proof of principle, that it is often computationally cheaper to involve the body and surroundings in solving these problems, thus mitigating the need for reconstructive perception, and in some cases eliminating it altogether.

Expanding the impact with precision modulation

Together, the two arguments above provide a strong foundation for resisting the internalist vision of PP. However, Clark goes further and builds on this, demonstrating the extensive reach this conception has over the explananda of cognitive science. He does this by shifting our focus once again, this time to the flexibility afforded by the function of precision weightings in PP, which at root provide a robust way of building context sensitivity into the process of solving any problem. This kind of context sensitivity give PP the power to account for many kinds of problem-solving strategy that involve adapting to particular circumstances, or taking account of new, or missing variables.

The mechanics of precision or *gain* in PP effectively permits a single system to instantiate a bag of tricks, each trick being the most effective strategy in some small subset of circumstances. Clark notes the well-known example of the outfielder catching the fly-ball which I covered in some detail in part one. Cancelling the optical acceleration of a specific object in the visual field isn't typical visual processing, but it is a trick the brain has learned which is extremely effective at catching fly-balls. The way precision modulation allows us to fill a network with tricks like this is to 'implement fluid and flexible forms of large-scale gating among neural populations' (Clark, 2015). Contextual cues prompt the generative model to recruit a specific learned strategy that will be best for dealing with the current problem, resulting in a high precision value being quickly assigned to that strategy. The expectations then generated produce a cyclical flow of highly-directed action and perception that result in effective problem solving, crucially without the need for a rich inner model – remember the system just needs to select the right trick for the situation at hand.

Enacting worlds

All of this scientific work also dovetails nicely with the enactive philosophical tradition. Returning to the way in which PP networks can make use of action to bring about expected sensory stimuli, Clark notes the similarities with the enactive tenet that organisms 'bring forth' their own worlds. An organism with a nervous system describable as a predictive processor exhibits a high degree of control over the way the world appears from their perspective, and so in some sense, what the world *is* for them.

PP also provides a neat implementation of Alva Noë's sensorimotor theory of perception (Seth, 2014). According to Noë, perception is constituted by a skilful knowledge of the sensorimotor contingencies the world presents. As we have discussed at length now, skilful active engagement with the environment is a core element of perception in PP, and accounts for much of the evidence that Noë draws on to support his enactive thesis.

Apart from action being constitutive of perception, there is the related enactive thesis that the primary function of perception is to enable action, in contrast to the *reconstructive* view that takes the processing of perception and action to be distinct – the function of perception being to reconstruct the external world before facilitating action planning. The enactivist predicts that perception and action are functionally intertwined, and this is exactly what we expect if PP is correct. As demonstrated by the Ballard et al. (1997) case study, and in the outfielder example, action is modulated quickly and efficiently using highly directed saccades and tracking eye-movements. Perception is hijacked for the task at hand, and must have close neural links with motor control areas in order to provide the necessary speed needed for each task.

Clark also suggests that PP offers an explanation for the broader world-creating claims of enactivism. Thus, 'we humans do not merely sample some natural environment. We also structure that environment by building material artefacts (from homes to highways), creating cultural practices and institutions, and trading in all manner of symbolic and notational props, aids, and scaffoldings' (Clark, 2015). Human technology and infrastructure can be interpreted as a way of making our worlds more *predictable*, thus creating a friendly and energy efficient environment for organisms aiming to reduce their prediction errors in the most parsimonious way possible – reducing uncertainty out in the world so our models require fewer parameters and our occurrent predictions can flow with higher certainty.

Another way to consider how a PP system might use technology to increase its predictive grip on the world is as way to bootstrap more and more active inference. Technology increases our access to the world, which in turn allows us to exert greater control over it, giving us new ways to interact and thus reduce prediction error. The lightbulb for example, allows us to confirm visual predictions during the night or underground, when previously we would be left with uncertain, imprecise perceptions, vulnerable to illusion. The telephone drastically increases possibilities for interaction on a social level, reducing uncertainty (manifesting as worry and stress) in the interpersonal domain. As technologies grow familiar we find new and interesting ways to exploit them to our advantage, and thus the enactive thesis that mind and world are co-creating and mutually self-sustaining (Varela et al., 1991) doesn't seem so far-fetched.

What’s left of representation?

There is however a tension between cognitive science based on PP and the enactive paradigm that Clark acknowledges. While enactivism shuns the notion of inner representation, PP appears to embrace it in the form of generative models, predictions and error. While enactivists talk of organisms’ survival in terms of a dynamic, homeostatic relationship with their econiche, rooted in a direct autopoietic coupling with the environment, PP is wedded to a framework reliant on the processing of information-theoretic inner states, weight and activation vectors that constitute predictions and knowledge about the world. Though Clark takes great pains to show that PP uses these tools in an extremely minimal and ultimately unproblematic way, he insists (Clark, 2016, ch. 9) that we must resist the fully non-representational conclusions of enactivism.

Clark is careful to distance his take on mental representation from the richly reconstructive notion endorsed by Hohwy: ‘instead of simply describing ‘how the world is’, these models - even when considered at the ‘higher’ more abstract levels - are geared to engaging those aspects of the world that matter to us. They are delivering a grip on the *patterns that matter* for the *interactions that matter*’ (Clark, 2015). The import here is that representations are not aiming at truth, or even accuracy, necessarily, but that they are needed for the system to get a pragmatic ‘grip’ on its situation, which is fundamentally action-involving and computationally parsimonious. Reconstruction demands too much of perception because we don’t *need* to reconstruct a highly-detailed model of the hidden causal nexus to get about, however, we do need to keep track of salient features here and there, and we do need a way to understand how the brain can deploy the right kind of sensorimotor strategies at the right times. For this, Clark argues, the computational understanding of PP with all its representations, inferences, and model-talk, is an indispensable tool.

In Part Three, I will nuance this final claim. In particular, I will show how the broader free-energy minimisation formulation shows us that talk of inference and models may be so weak as to require a re-understanding in terms of dynamical systems theory. In *Radical Predictive Processing*, Clark has laid some persuasive groundwork that will help to motivate an understanding of PP that eliminates the need for representations in many areas of cognition. Indeed, he also indicates an important feature of the contents of generative models – that they are almost certainly impossible to capture with natural language, because generative models are in a very different business to that of natural language, they are in the business of survival, not of clear communication. We might question then whether talk of contentful states can really help us understand and explain the workings of a system when, by the advocates' own lights, the contents are too complex to be expressed. When science comes to a point like this, perhaps an alternative interpretation of the model is necessary, in order to further our understanding and bestow clarity on our explanatory projects.

5.3 Embodied or Encapsulated Predictive Processing?

Jakob Hohwy has issued several forceful arguments against Clark's claims. He rejects the suggestion that PP is compatible with enactive cognition; he denies the role of action is grounds for extending the mind beyond the generative model; and he also rejects Clark's appeal to quick and dirty cognition as a way of deflating the role of representation in PP. First however, I want to take a brief look at Hohwy's own conception of the mind-world relation.

Hohwy's position is that our connection to the outside world is no more direct than the connection of any object to its immediate surroundings. That is, we are simply linked causally. He states that 'this plea for a causal conception of the mind-world relation [is] compelling because it places us squarely as just elements in the overall natural, causal nexus of states and events in the world' (Hohwy, 2013, p. 229), and goes on to re-iterate the claim that the predictive mind is a fragile mirror of nature: 'some causal connections are very fragile, even if they do work reliably within a circumscribed set of conditions. It doesn't take much to upset them. I have been arguing that the mind is fragile in this sense' (Hohwy, 2013, p. 229). So although we are connected to the world, and ‘shaped to recapitulate’ it, ‘this is essentially a passive and conservative process’, and ‘in that sense we are mere pawns in the causal nexus’ (Hohwy, 2013, p. 229).

The core claim here is extremely weak – PP systems, all organisms, are part of the causal order, and our interactions with the environment all boil down to causal relationships, thus in some sense we are ‘mere pawns’ just like everything else – all naturalist philosophers, Clark included, would be inclined to agree. However, it is possible to put pressure on Hohwy’s inference that this means the mechanisms through which organisms preserve themselves are passive, fragile, and reconstructive. Once we attempt to explain how systems do the things they do, and invoke causal-mechanical relationships which introduce the notion of function, we can ask interesting questions about whether or not characteristically mental functions are distributed between mind, body, and world.

To rebut this kind of argument Hohwy has analysed the classic Clark and Chalmers (1998) Otto’s notebook thought experiment. He suggests that the notebook would fail to qualify as an external mental state, because PP offers a tighter notion of ‘mental function’ than the broad functionalism Clark and Chalmers were originally targeting. Otto’s notebook looks like a close match for a memory or belief-like state, but does not perform the function of a prediction, an error message, or an error-minimising generative model.

The externalist can quite easily concede the particular case here, but maintain that there are plenty of other, more common examples of states do perform long-term error-minimisation functions. I outlined a few of these in chapter four. Others might include clothes, which serve to suppress the amount of wind and cold that might impinge on our senses, and tools which allow us to offload tasks that otherwise would need to be performed by the organism itself, thus increasing its long-term efficiency (Otto’s notebook, or indeed anyone’s notebook, could be considered part of a prediction-error minimisation strategy in this sense).

Hohwy (2014) follows up his complaint with a further worry with these examples that 'the object is both beyond one evidentiary boundary and within a further evidentiary boundary', thus creating a needlessly complex explanatory framework, where his simple, encapsulated view would be more parsimonious and perfectly adequate for accounting for mental phenomena. However, as mentioned earlier it is important to acknowledge that the placement of an evidentiary boundary is *explanation relative*. For any given explanation we only need posit one boundary (in a loose sense, to define the system we are concerned with), which is perfectly simple and parsimonious. Furthermore, a cognitive science that embraces this kind of relativity can use the evidentiary boundary construct to its advantage, allowing us to explain a wider range of cognitive phenomena with appeal to extended, embodied, and enactive crutches which properly capture the mechanisms at work. In this way we might contest that Hohwy's appeal to explanatory virtues fails, because there is no decrease in parsimony by adopting an externalist perspective, only an increase in potential fruitfulness.

Next I want to examine Hohwy's reasons for deflating the function of action to another form of internal inference. He notes the three distinct functions that action appears to play – action can bring the sense data into line with expectations; action can be used to verify a hypothesis; and action can be used to falsify a hypothesis. However, he maintains that all three of these apparently distinct functions (which appear to make action an importantly embodied part of cognition) deflate to 'prediction error minimisation in the light of the expected precision of prediction error', thus, 'acting in the world is therefore nothing to do with a direct, seamless engagement with the world… to act is just to engage in more statistical inference' (Hohwy, 2014).

Advocates of a strongly action-involving externalism would not deny that action is part of an organism's statistical inference, in the sense that it enables error-minimisation. But it is important to note that action importantly enables the system to engage in the kind of computational offloading that I have described at length in this chapter. Again, this offloading allows the system to reduce systemic 'error' by lowering the number of parameters in the generative model. Understanding action in these terms supports the view that cognition involves direct and seamless engagement with the world, because the engagement and the world both contribute to the underlying computational processes. Behind Hohwy's evidentiary boundary, action does 'reduce to an inference about different kinds of sensory input' (Hohwy, 2013, p. 220).

The third and final point of contention I will discuss is a short argument Hohwy makes regarding 'quick and dirty cognition'. Though Clark's examples of, and appeals to quick and dirty, world-involving cognition appear to offload computational complexity to body and world, 'this overlooks the fact that the kind of selective, sparse sampling in play here requires heavy, explicit modelling of external causes… in other words, the sampling can be simple and efficient in the way highlighted by Clark only because countless aspects of the causal order of the world are already being modelled internally' (Hohwy, 2014). This response misses the import of Clark's examples. Rather than necessitating a rich internal model that thus permits the organism to engage in selective sampling, the generative model can instead direct patterns of action directly, that exploit the structure and dynamics of body and world, to avoid the baggage of a parameter-heavy model. The generative model will adopt this action-centred strategy precisely because it is more parsimonious than the alternative that Hohwy insists is necessary.

5.4 Conclusions

In this chapter I outlined two philosophical responses to PP. Jakob Hohwy holds that PP heralds the return to an internalist, indirect philosophy of mind and perception. On the other hand, Andy Clark argues that PP provides a strong mechanism that accounts for all the lessons from embodied and enactive cognition. Here I have endeavoured to support his case by responding to arguments from Hohwy, appealing primarily to the built-in parsimony requirement that the full description of prediction error mandates. I have also hinted that Clark doesn't go far enough in his conclusions, and that fully non-representational enactivism may also be compatible with PP's mechanism.

Chapter 6: Evidence for PP

In this chapter I will summarise the evidence for several elements of the neural mechanism suggested by PP. These elements are: functional asymmetry between hierarchical layers in perceptual processing (Friston, 2005); precision weighting modulated by the pulvinar in the thalamus (Kanai et al. 2015); functional asymmetry between forward (driving) and backward (modulatory) connections in granular layers of the cortex (Friston, 2005) (Bastos et al., 2012) (Kveraga et al., 2007) (Huang and Rao, 2011); neural representation of uncertainty in the lateral intraparietal (LIP) area (Knill and Pouget, 2004). While there is also a growing corpus of psychophysical evidence that corroborates the predicted behaviour of an embodied predictive processor, I will not focus on this source of evidence, as I wish to focus on the neural underpinnings of the mechanism. Some good summaries of this work include: Bubic et al. (2010), Friston (2009), Yuille and Kersten (2006), Weiss et al. (2002).

The implications for our ongoing understanding of perception and action are far-reaching if PP is an accurate model of neural function. Especially when considering Friston's active inference hypothesis, we can note several observables that would go some way to confirming the theory. Adams et al. (2013) make a detailed case for the neural plausibility of active inference. In particular, they note that the functional organization of the brain supports the idea, insofar as perceptual and motor control areas share many lateral and vertical connections which is strongly implied by the theory, and suggestive of its accuracy. Furthermore they note that "as much as 25% of the corticospinal tract originates from postcentral, sensory areas of cortex" which they hypothesize indicates a substantial pathway for downward flowing predictions of self-actuated action in order that we are less sensitive to sensations resulting from our own movements.

6.1 Hierarchical Organisation

Underpinning all deep learning and the PP framework in particular is the use of a hierarchical architecture that facilitates abstraction. Higher layers operating on lower layers abstract relevant information in order to respond adaptively to spatially and temporally extended stimuli. If the brain isn't organized at least somewhat hierarchically, then these research programmes are dead in the water. However, as noted by Friston (2005), it has been well-known for some time now that it is highly probable that the brain is loosely hierarchical, albeit with a great deal of parallelism and neural shortcuts. Some direct evidence for this comes from Van Essen and Maunsell (1983) who mapped neural connectivity in the macaque monkey and found a distinct hierarchical structure, and Felleman and Van Essen (1991) who performed a similar analysis, finding that both the visual and motor cortices are arranged in hierarchies, and that the visual cortex is organized into 14 discrete hierarchical layers. Further to this there are telling details within the hierarchical organization. First, backwards connections appear to be more abundant than forwards connections (Friston, 2005), which is suggestive of considerable top-down influence in neural processing. Second, backward connections tend to have much longer range than forwards connections, which are typically limited in scope (e.g. Salin and Bullier, 1995). This is also compatible with PP, which takes advantage of high-level areas' ability to influence prediction at lower levels, but also expects prediction error to be processed by the layer directly above. These factors help to show that PP has a strong start with respect to the neuroanatomical structure of the brain, as not only do we find a hierarchical structure, but the *right kind* of hierarchical structure.

6.2 Precision weighting

Recall that prediction and error are modulated by precision estimation. Precision estimation is theorized to be responsible for a wide variety of phenomena and functional effects, chief among them is attention. High precision will be assigned to upward flowing error signals when the system has low confidence in its hypothesis, and so needs to be more sensitive to incoming data. Biologically this is realized by increasing the post-synaptic gain on prediction errors that are estimated as precise.

A brain area whose function is to *modulate* precision weighting in certain tasks should implement a mechanism for increasing or decreasing the impact of specific prediction errors when performing that task. Komura et al. (2013) found that the pulvinar was highly active when monkeys exhibited certainty about their ability to perform a categorization task, but when the pulvinar was suppressed, the monkeys exhibited uncertainty despite no change in their ability to perform the task. Kanai et al. (2015) note that this finding provides compelling evidence that the pulvinar modulates precision in visual categorization tasks when combined with the findings of Saalman et al. (2012), who demonstrated that the pulvinar is instrumental in facilitating information transfer between V4 and TE by synchronizing alpha wave oscillation in those areas. When the alpha wave oscillations are closely synchronized, information transfer is stronger between the two areas, but when alpha waves are out of sync, information transfer is disrupted. This is exactly the kind of physiology we expect to find in an area responsible for precision modulation.

Kanai et al. (2015) go further and note that the pulvinar, and the thalamus more generally, is responsible for modulating oscillations in many cortical regions. Barth and MacDonald (1996) link the thalamus to modulation of gamma oscillations in the auditory cortex, Contreras et al. (1996) demonstrate that the reach of thalamic influence is extremely long-range, with signals being transmitted through deep corticothalmic connections. Han and van Rullen (2017) note the role of frontal theta-oscillations and occipital beta-frequency oscillations on prediction. Shipp (2004) provides a theoretical overview of this long-range thalamic connectivity with respect to its modulatory role.

The thalamus is well-placed to perform this vital function, sitting at the crossroads of incoming sensory signals, and responsible for ‘sorting’ these signals and forwarding them to the appropriate cortical regions for further processing. Considering in addition its deep interconnectivity with deep cortical areas, the thalamus has all the anatomical and physiological features one would expect of a region tasked with modulating the relative precision of low-level sensory signals with respect to the high-level cortical activity.

6.3 Top-down and bottom-up asymmetry

Another key element of PP is the functional asymmetry implied by the distinction between top-down flows of predictive information, and the bottom-up flow of residual error. This provides the neuroscientist with an empirically testable claim that falls out of the theoretical PP framework. If it were found that top-down and bottom-up processes were largely symmetrical, performing largely the same function across the brain, PP would be roundly falsified. It is thus a great relief that there do seem to be important physiological differences between forward and backward connections, and further, that these differences map neatly on to the functions hypothesized by PP.

Kveraga et al., (2007) provide a thorough analysis of the role played by the orbitofrontal cortex (OFC) in a PP architecture. According to PP, high-level cortical regions, such as the OFC, would be responsible for generating predictions that influence the activity of many other areas, orchestrating appropriate responses. They found that several physiological and anatomical properties of the region confirmed this hypothesis. Pathways leading to the OFC are delivered by 'fast magnocellular pathways' carrying 'gist' information, which is responsible for triggering top-down responses. In this case, gist information seems to be playing the role of long-range error signals, providing contextual cues about the current situation. The OFC also seems to activate early on during recognition tasks – an unusual result for a feedforward understanding of perceptual processing, but a key prediction of PP that puts a strong emphasis on high-level influences in perceptual processing. A more speculative finding that Kverga et al. (2007) offer, is evidence that activity levels in the OFC correlate positively with image ambiguity, which suggests that the area is working harder when there are multiple interpretations of an image, and thus could indicate some kind of probabilistic inference being performed.

If PP is correct, then we can say more about the functional differences between top-down and bottom-up pathways. We expect error signals to propagate in a usual feedforward fashion, excitory signals responsible for driving processes (e.g. probabilistic inferences and learning procedures). On the other hand we expect predictions to inhibit or modulate error pathways, the core of their predictive task is to 'say' what they expect to happen at the level below by inhibiting the flow of error, but also have some driving connections in order that high-level predictions can influence low-level predictions. Friston (2005) and Huang and Rao (2011) provide several examples of studies showing that we find precisely this kind of physiology in the visual system (Girard, et al., 1991; Büchel and Friston, 1997; Dong and Atick, 1995; Dan et al. 1996) finding that backwards connections from the middle temporal area (MT) to V2 were primarily modulatory, while forward connections were driving. These confirmations are further strengthened by the nature of the receptors involved in backward versus forward flows of information in high cortical layers (supragranular layers). Forward connections use extremely fast receptors AMPA and GABAA, which decay between 1.3-6 milliseconds. Backwards connections on the other hand are effected with NMDA receptors, which have a much longer decay time (around 50 milliseconds) and also are sensitive to voltage, which – according to Girard et al. (1991) - allows backwards connections to display nonlinear dynamics, another crucial aspect of the PP framework.

Bastos et al. (2012) provide an excellent summary of the predictive physiology of backward connections. They cite a series of neuroimaging studies that not only demonstrate that expected stimuli result in lower overall levels of brain activity, indicative of predictive inhibition of error, but also that these low activity levels are not the result of habituation or adaptation[12] which goes some way to ruling out competing explanations of these findings. Bastos et al. (2012) also provide some recent work that helps establish the second function attributed to backward connections – their driving influence on lower layers. For example, they note that Covic and Sherman (2011) found these downward flowing driving influences in the auditory cortex (A1 and A2) and De Pasquale and Sherman (2011) found corresponding data in the visual cortex (V1 and V2).

[12] Murray et al., 2002, 2006; Harrison et al., 2007; Summerfield et al., 2011; Alink et al., 2010.

The growing store of evidence for systematic functional asymmetry in forward and backward neural pathways is a great boon for the PP advocate. As mentioned above, this is one of the core testable elements of PP that differentiates it from other high-level theoretical work in cognitive science.

6.4 Neural representations of uncertainty

Of utmost relevance to the philosophical aims of this thesis, there are some studies that examine the neural mechanisms responsible for probabilistic decision making. Knill and Pouget (2004) cite three neuroimaging studies that illuminate these representations of uncertainty in monkeys. Platt and Glimcher (1999) showed that activation of a subset of neurons in the lateral intraparietal area (LIP) was proportional to the probability of a certain eye-movement being rewarded. The LIP is responsible for sensorimotor integration of visual stimuli and saccadic movements and these results confirmed that there were neural correlates for the Bayes-optimal behaviour exhibited by the monkeys (whose task in this case was to move their eyes left or right for a sip of juice, experimenters varied the amount of juice over time, observing behaviour patterns and also measuring activity in the LIP area). Knill and Pouget (2004) note that this is consistent with probabilistic representations being implemented with distinct populations of neurons encoding probabilities for each possible outcome.

An alternative method of encoding uncertainty is to instead have a single population of neurons encode a likelihood ratio. This form of representation is hypothesized in a report from Gold and Shadlen (2001) who develop a computational model of neural encoding of likelihood ratio, and resultant decision-making activity. Which, combined with recent experiments, 'have suggested that the [logarithm of the likelihood ratio] is accumulated and represented in neural structures that are involved in planning actions'. Similar methods were used by Anastasio et al. (2000) to identify representations of likelihood ratios compatible with computation of Bayes' rule in the superior colliculus.

These efforts provide good evidence that certain processes are able to reflect the statistical regularities present in a task-domain. There is still work to be done however, as the more conclusive Platt and Glimcher (1999) experiments are focused on fairly low-level visual processes, in a situation with highly limited choice. It would be interesting to see how the proposed mechanism of encoding separate probabilities for each hypothesis was able to scale up in higher cortical areas, when complex decision-making procedures and greater uncertainty of outcome are at play. The studies presented in favour of the encoding of a likelihood ratio so far only provide circumstantial evidence in support of their computational models. However, the framework is robust, and may be more computationally versatile in scaling up to higher layers of the network, responsible for more complex uncertainties.

6.5 Conclusion

In this chapter I summarized a small but representative subset of the kind of empirical work which confirms (or at least, is visibly consistent with) the neural mechanism proposed by PP. The structure of the cortical hierarchy seems to map well on to the PP architecture, and with the anatomy supporting the right kind of physiological relationships we would expect. There is a strong case that precision estimation is carried out by modulatory feedback connections that can 'adjust the volume' on error signals, and thus account for attentional effects. At the heart of PP lie the distinct functional roles played by forward and backward neural connections. I went into detail on several studies that not only confirm a functional asymmetry, but also indicate the kind of functions that PP predicts for these pathways. Backward connections modulate long-lasting effects and display non-linear dynamics, and forward connections convey residual error, Finally I examined some examples of neural representation of uncertainty. These provided an interesting starting point for this kind of research, but it is too early to draw any firm conclusions.

Part Three: Function, Dynamics, and Intentionality

7. Mechanism, Representation, and Function in PP Systems

7.1 Representational Mechanisms in PP

Which components of the mechanism schema offered by PP are supposed to represent? Do these components pass Ramsey's job-description challenge? Our discussion of the first question will be relatively brief, though some ambiguities will be noted, and I will attempt to resolve these. Analysis of the candidate components will reveal whether their role-function within the PP schema is properly representational, or better understood as performing some other mechanistic function.

Clearly, as a computational theory, PP treats these elements as representational at the symbolic level of description. However, as we learned in chapter two, it is not always obvious whether the symbols in a connectionist architecture can be properly grounded to their content. I will thus follow the strategy advised by Ramsey (2007). Recall also the power of the teleological theory of intentional content to account for representational error in a way that its main competitor – the causal theory – fails to do. In light of this, the second task of this section is to take a step back and consider how to properly understand function, and explore an alternative to the etiological theory of function (Millikan, 1989; Neander, 1999; Garson, 2012) and consider whether it supports the possibility of intentional content.

Which components of the mechanism schema offered by PP are supposed to represent?

For a first pass at the first question, we can consider the way representational language is used when explanations are given by PP theorists to a lay audience. A good example here is Jakob Hohwy's characterisation of the explanation given for the phenomenon of binocular rivalry. In this case, each eye is presented with a different image, e.g. the left eye is shown and house and the right eye is shown a face. Subjects do not report seeing a ‘house-face’ as one might expect, but rather an image that oscillates between being a house or a face, suggesting that each eye is ‘fighting’ for the brain’s approval, hence the term binocular rivalry. Hohwy offers an explanation in the Bayesian language of predictive processing:

> “To put it in the Bayesian vernacular, the prior probability of such a mishmash cause of my perceptual input is exceedingly low. Instead, very “revisionary” hypotheses are selected, each of which effectively suppresses a large part of the incoming sensory signal. It is as if when a face is seen the visual system says, “it is most probably a face, never mind all the parts of the total input that the face hypothesis cannot explain”; and *vice versa* when perception then alternates and the house is seen. How exactly this inferential process proceeds is a further matter but it is difficult to see how we could even begin to explain this effect without appealing to some kind of inference.” (Hohwy, 2013, p. 20).

In this paragraph, Howhy presents what he considers to be the minimal requirements for handling cases of binocular rivalry. Some kind of inference is taking place, and so representations are required as we use inferences to move from knowledge about something to knowledge about something else. Fulfilling this role, Hohwy speaks of hypotheses and perceptual inputs. The suggestion is that the system represents both of these in a way that makes inference about them possible. Also relevant is knowledge of prior probabilities; a core constituent of this explanation is that the subject’s perceptual system knows that the presence of a house-face is absurd, that is, that it has an extremely low prior probability. It is thus more likely that one eye is providing erroneous information to the system, and that there is in fact either a face or a house in front of the subject. However, the system lends almost equal weight to each eye, hence the reported perceptual oscillation, as the information from each eye is given preference in turn, at first supporting the house hypothesis, and then the face hypothesis. The system must be representing several things – the prior probability that there is a house-face, the prior probability that there is a house, the prior probability that there is a face, the image from the left eye, the image from the right eye, the confidence it has in the left eye, the confidence it has in the right eye, and finally the posterior probability of each of its hypotheses once it has performed the Bayesian inference based on the information it has.

These beliefs about the prior probabilities of any given state of affairs would be encoded in the knowledge nets (Rao and Ballard's 'predictive estimators' (1999)) instantiated at each level of the hierarchy in a predictive processor. They are an extremely important part of the inferential mechanism and are clearly supposed to be about the hidden states of the external world, which implies that they represent those states so that the system can make its inferences.

Also important are the predictions generated by these Bayes-approximate inferences, which are passed down as predictions to layers below, and their companion, precision which provide a confidence weighting for the prediction, taking into account broader contextual factors.

Finally, we might consider the upward flowing information provided by our sense organs. This is the driving informational force from the external world, that classical models of perception use to construct percepts and concepts. In contrast, predictive processors seek to suppress this flow with precision-weighted predictions, so the only information that propagates is error, the information that wasn't predicted by the system. It is not immediately clear whether this driving signal is representational – on one hand it seems to be merely transduced information which is not decouplable from its source without further processing, on the other hand these seem to be the informational building blocks of perceptual representations, and so seem to have some intrinsic intentionality, and thus may be candidates for representation-hood.

We can see that the way prediction and precision are represented will be closely linked to the way error is represented and to the form of incoming sensory data, as the error signal is generated by comparing a prediction (modulated by precision) to the actual sensory signal, and calculating the differences. Even if the precision, and prediction and error signals are not immediately comparable, they must be commensurable in some way, whether they represent or not.

It is worth noting the possibility that although precision must necessarily interact with prediction signals, precision weightings may be properly representational even if it turns out that prediction signals are not. That is, precision may represent something about predictions, even if the predictions themselves are not representing anything. In this case, precision weightings would be mere representations, rather than meta-representations, but interesting for our discussion nonetheless.

Until now we have been considering the naïve view suggested by the loose use of representational language when giving example PP explanations to a lay audience. It is worth testing this view by taking a look at some case studies within the PP literature, which might shed some light on whether this language is metaphorical, whether it is a gloss on the proposed mechanism, or whether it accurately describes the mechanism. Indeed, there can be no argument that representational language is ubiquitous in the way PP is articulated, so looking a little more closely at the way the framework is invoked, at both an abstract, computational level and at a neurobiological level, is necessary in order to take any sceptical concerns seriously.

As it turns out, more detailed discussion of the underlying mechanisms also uses representational language extensively. For instance, Adams, Shipp, and Friston (2013) use the term 'representation' to refer to the total knowledge or model of the task domain possessed by the system; in contrast they use the term 'prediction' to talk about the highest likelihood estimated cause of input, as we have been so far. It is interesting that they separate the terms in this way, as it suggests that although the sense of the term 'prediction' seems to involve the representation of some content, that Adams et al. do not necessarily consider these signals to be representing, or at least hints that they are playing an importantly different function from the generative model itself.

Friston (2003) discusses a slightly more abstract, computational-level analysis of PP systems, and addresses each candidate in representational terms. However, given the computational nature of the analysis, his use of representational terminology is somewhat weaker than that of Adams et al., that is, he takes variables in the equations that determine the dynamics of the system to be representing specific activity vectors implemented by neural networks. This is a fairly innocent use of the terminology, not entailing that the state-vectors themselves are content-bearing. When it comes to discussing this kind of function, Friston (2003) is in broad agreement with Adams et al., suggesting that the system's model represents the causes of its sensory inputs, and learns a better system of representation based on error signals. This point is also echoed by Rao and Ballard (1999), who provide a little more detail. In their model, they 'assume that the cortex tries to represent the image in terms of hypothetical causes', and this representation is encoded holistically in the weights of the connections between units. As Bogacz (2017) clearly illustrates, precision is also implicitly encoded in connection weights as part of the model of the task domain. The variance of the probability distribution supporting the most likely hypothesis, and the variance of the probability distribution assigned to the incoming sensory evidence modulate the prediction and incoming data to calculate the resulting error signal.

In this section, I have examined the way the PP schema is used in explanations from those given in simple, lay terms, to the more technical offerings given in research papers. I found a slight dissonance between these two styles concerning the status of the system's prediction signals. In Hohwy's simplistic explanations these are styled as hypotheses that carry content about the world, the result of Bayesian inferences. The technical literature uses these signals in a slightly different way, as the output of a representational system – the knowledge nets – and whether this output is itself representational is not specified. It doesn't follow necessarily that the output of a representational system is representational. However, whether these prediction signals are representational is an important question due to their commensurability with error and precision signals. If prediction signals are representations, then it would strongly support a representational understanding of precision and error also, resulting in PP systems that are replete with representations. The degree to which PP systems are representational thus depends deeply upon whether predictions are representations, so it would be foolish to set these aside purely because research hasn't given a clear indication on the matter one way or another. In the following section then, I will consider first whether the functional characterisation of the prior-encoding knowledge-nets as representational is correct, and then secondly whether predictions and the associated error and precision signals function as representations in the mechanistic schema offered by PP.

Do these components pass Ramsey’s job-description challenge?

High level descriptions of PP architectures tend to say little about the implementation of the prediction-generating knowledge nets that do the heavy lifting of the generative model. When the implementation is discussed, or a model has been built by the authors, knowledge nets are typically instantiated using connectionist networks, or neural networks (e.g. Rao and Ballard, 1999). Each net is linked hierarchically with others in the system via strict functional relationships (the calculation of prediction and error). So, the component which is taken to be representing is a connectionist network, with the stored contents taken to be encoded in the activation vector or weight vector (a complete description of the strength of the connections between each node in the net) of the network. There are well-trodden dialogues on this topic that we considered in chapter two (Smolensky, 1987, 1988, 1995a, 1995b; Fodor and McLaughlin, 1990; Clark, 1989, 1993) which have demonstrated the ways in which aggregate network states can form symbolic representations. So let us consider in a little more depth Ramsey’s (2007) doubts concerning whether connectionist representations can be properly grounded, which are still being discussed (Sprevak, 2011; Shagrir, 2012).

The vehicles of representation in connectionist networks are taken to be only recognisable at a level of description abstracted from the individual nodes and connections of the network. These basic elements are instead taken to encode information ‘subsymbolically’. Representations are identified in either the weight vector of the network, the state-space of the activation dynamics (e.g. Shea, 2007), or a combination of both (Smolensky, 1995). This abstracted view suits PP. As discussed in part two, their knowledge nets are described in terms of a weight vector. However, of the three ways of making sense of connectionist representations, identifying them with the weight vector is perhaps the least sophisticated and most vulnerable to Ramsey-style criticism.

Recall that Ramsey (2007) argues that representations in connectionist systems are either mere causal mediators (receptor representations), or merely dispositional properties of the system (tacit representations). Representations encoded in the weight vector (the complete set of long term connection weights) of a network are tacit representations – they are not explicitly coded by discrete components or processes in the system. Instead their presence is tacit – their existence is only evidenced by the selective and intelligent behaviour of the system which suggests a representational explanation. Ramsey's claim is that whether or not the system behaviour can be explained in representational terms, the elements that we call tacit representations are not functioning as representations. That is their function is not best understood as 'standing in for' some content, rather they opaquely modulate the system's response to input, and thus function as encoding a set of dispositional properties. We can distinguish nets functionally in terms of the kinds of dispositions they encode, e.g. net *x* encodes speech-related dispositions, and net *y* encodes action-related dispositions. But it is important to note that a system that can be carved up like this doesn't equate to a system possessing organised representations, no matter how finely we can carve it (e.g. even if net *u* encodes horse-related dispositions and net *w* encodes biscuit-related dispositions).

Applying this argument to PP systems specifically, we might argue that hierarchically organised knowledge nets can legitimately be carved up in terms of functional sensitivity to greater or lesser abstraction, and to longer or shorter periods of time, as is commonly done. However, we can recognise this (very attractive) feature of PP without admitting that the network is encoding representations that stand in for more or less abstract concepts, or action plans. Instead we can insist on a more accurate characterisation of the function of knowledge nets as encoding dispositions whose function is to mediate interaction with the environment at certain degrees of spatio-temporal abstraction.

A natural response to this kind of argument is that the representational ascription helps us understand what is going on inside the knowledge nets, and that without such ascriptions Ramsey takes us back to an uninformative behaviourism about cognition. Indeed, it seems to be true that treating knowledge nets as disposition nets removes any guide to discovery regarding their inner workings and perhaps promotes complacent and uninformative science, which is content to describe these important parts of our cognitive system purely in terms of inputs and their associated outputs.

While powerful, this kind of argument doesn't knock-out the sceptical argument. The symbol-grounding problem is a metaphysical problem – and appealing to explanatory virtues fails to solve the question at the heart of the problem: in what concrete sense does the system possess stand-ins for some content? If we argued that the brain is one large, opaque, connectionist net, then treating it as a very complex disposition machine would indeed be antithetical to the computationalist project. However, as we explored in chapter three, there are alternatives to the computational paradigm, and in the next chapter we will see how PP is able to leverage the tools of DST. We'll see that treating knowledge nets as dispositional rather than representationally merely undermines the practice of ascribing representational content to the nets, only recommending a re-understanding of these components, not an overhaul of the entire project, which is not as wedded to computationalism as perhaps is first supposed. Further to this, it may be that such an approach does in fact offer an alternative guide to discovery, by treating knowledge nets as complex self-organised systems, which allows us to bring to bear a new set of analytical tools. I will go into that approach in more depth in chapter nine.

Rather than challenging Ramsey's argument in terms of explanatory virtue, we might instead challenge it empirically. Connectionists frequently equate weight and activation vectors with representations. For instance, Geoffrey Hinton (2007) trained a generative model to classify handwritten numbers. He identified representations in the hidden layers of the network by considering the weight of the connections between a given unit in the hidden layer with an input unit. After training, most units in the hidden layer developed sensitivity to different points of the system's visual field, and some became sensitive to straight lines or broader distributions of input. Following this analysis, it doesn't make sense to call the network opaque. We can interfere with the mechanism precisely and stop it recognising some part of its input. So it makes sense to say that such and such a hidden unit is standing in for such and such an input, as the output layer of the network which decides which number is being presented on the basis of the activation of the hidden units, which are standing in for the raw input data.

This kind of precise analysis provides a strong case against the sceptical claim that all these analyses do is provide a mapping of the dispositional sensitivities of generative models. However, the key point at the centre of the debate here is whether or not the hidden units are functioning as stand-ins for the world. The sceptic can argue here that looking simply at the connectivity of the hidden units with the input units is not enough to determine that the hidden units are representing a certain input pattern. That is, from a mechanistic perspective, whether or not the units are playing the standing in role depends on how they are being used by the output units. A unit in the hidden layer may, for example, be sensitive to a vertical line in the visual input but have very weak connections with the output layer, i.e. it may be functionally irrelevant to the output of the system. In this case, the unit would have the kind of weight vector that would enable it to function as a representation of vertical lines, but is not functioning for the system, that is, for the mechanism, as a representation of vertical lines and so is not a representation of vertical lines. To make sense of whether the system represents vertical lines, we must include an analysis of the connectivity between the output layer and the hidden layer alongside the connectivity of the hidden layer and the input layer. Importantly, focusing on just the connections between the output layer and hidden layer does not provide enough information about the functional structure of the network to justify representation talk. So again, in order to have a chance of identifying representations, we are forced to tackle the entire weight vector of the system which provides only a description of its complex sensitivities and dispositions, and not of components standing in for classes of percepts or concepts.

Work such as Reddy et al. (2011) claims to decode neurally-stored representations using fMRI with subjects engaging in imaginative exercises. Thus we seem to have evidence of mental representation that truly functions to stand in for contents. In their study, Reddy et al. (2011) showed subjects pictures of four different objects and then asked them to imagine the objects. By comparing neural activation patterns during visualisation with patterns during veridical perception, the researchers found that they could identify which object the subject was imagining with 50% accuracy rate. This rate is significantly higher than the probability of chance identification, which was 25%. So Ramsey's claim may be mistaken, as we have good evidence that perceptual representations are invariant with respect to the presence or absence of a stimulus. That is, these neural responses are decouplable from their stimuli, which suggests that they really do function as stand-ins for their stimulus, or at the very least have the potential to function as stand-ins.

The main problem with this argument in this particular dialectical context is that PP identifies the system's long-term knowledge representations with the weight vector of the knowledge nets, not with their activation patterns. So, strictly speaking the work done by Reddy et al. (2011) does not conflict with Ramsey's claim. However, it does suggest that we should take seriously the possibility that neural activation patterns represent.

Whilst the mathematics of PP refers to knowledge nets in terms of their weight vector, that weight vector does define a set of possible activation patterns (or vectors) within the network. Another (decompressed) way of describing a knowledge net might be using a complete list of possible activation vectors. The sceptic might endorse this kind of description as it results in a complete understanding of the network's dispositions. However, it also provides the representationalist with resources. With access to these activation vectors, we can identify similarities and patterns between certain kinds of input and certain kinds of output. This kind of thinking results in us identifying neural representations using techniques such as cluster analysis (e.g. Clark, 1990).

Cluster analysis of a well-known connectionist network, NETtalk, which was trained to perform grapheme-to-phoneme translation, demonstrated that the network's activation patterns were clustered into readily identifiable categories (Sejnowski and Rosenberg, 1986). The network had clearly differential responses to vowels and consonants for example, with all vowel responses being more similar each other than consonants, and all consonants being more similar to each other than vowels. Within these categories, further categorisation was evident. For example, plosives were clustered within the consonants category. Representationalists claim that this research indicates that NETtalk has knowledge of these categories, and this knowledge functions as part of the mechanism which allows the system to perform correctly. Strengthening this claim is that the NETtalk experiment was performed several times, and in each case, the system developed these categories, identifiable as clusters of similar activation vectors.

Stich and Warfield (1995) provide a simple response to these claims. We cannot extrapolate from the NETtalk example, they argue, because the analysis performed was only done on systems with a particular architecture. It is possible that there are other 'NETtalkers' out there that could perform accurate grapheme-to-phoneme translation without succumbing to the same kind of cluster analysis, purely because their architecture is structured differently. Whilst this is a valuable observation, as it ensures that we remain liberal in our search for biologically accurate models of cognition, it also fails to refute the broader lesson of the representationalist's argument. That is, whatever the details of a network architecture, we may be able to perform a kind of cluster analysis and thus identify the organisation of a conceptual schema being used by the system to perform its task. If for instance an alternative architecture was used for grapheme-to-phoneme translation, and different categories emerged in the organisation of the system, then perhaps we would learn something new about the relationships between the written and spoken word. Whatever the result, the fact that we are able to perform cluster analysis demonstrates that if we look in the right way, we can find meaningful patterns in network activity that can help us understand how it works. Indeed, subsequent work has been done to demonstrate just this point (e.g. Laakso and Cottrell, 2000).

I believe the stronger argument against this kind of approach is to cast doubt on the functional role of patterns of activation vectors. Activation vectors are descriptions of the flow of energy through the network given a certain input. From a mechanistic perspective these are a complete description of the system's process in response to its given input, but this plays no role in explaining, mechanically, how this happens. In order to provide a mechanistic explanation, we need to know the weight vector, as it is the weight vector that provides the functionally relevant properties of the system components, and thus is all that is relevant for the explanation of system responses. Things become even more tenuous for representations when we note that they are not even to be found in activation vectors themselves, but in patterns of activation vectors, another step of abstraction away from functionally or causally influencing the mechanism. From the strict perspective of mechanistic explanation, representations emergent in patterns of activation vectors, discoverable through cluster analysis, play no role.

Nicholas Shea (2007) offers an eloquent rebuttal to this kind of argument. He considers the capacity of classification systems like NETtalk which is so often a benchmark for their success or failure – the ability to generalise by classifying new inputs, ones not contained in the training set, successfully. For instance, if NETtalk had never read the word 'fox' before, but nevertheless output the correct sequence of phonemes, then NETtalk is successfully generalising. Shea argues that to explain these cases connectionist researchers usually appeal to the clusters learned by the system, and importantly, that they are right to do so. Successful generalisation occurs because the system's response to a new input falls into the cluster that produces the correct output. When a system has not clustered properly, or indeed, failed to cluster at all, it makes mistakes when attempting to generalise: 'in that case, misrepresentation at the hidden layer helps account for misrepresentation at the output layer' (Shea, 2007). Clusters represent the properties shared by similar inputs – and we cannot simply explain this with reference to the inputs present in the training set, since novel samples will provide novel inputs, so Shea argues, we can only appeal to the properties of the samples themselves. That is, the clusters do not just represent the groups of inputs by the output they ought to produce, but are actually learned representations of the properties relevant to producing the correct output.

Furthermore, Shea (2007) addresses my specific demand in this section, that these cluster representations be relevant to mechanistic explanations – for this is exactly the kind of way they are used in the literature. A cluster representation does not possess causal powers per se, so will not appear in a full mechanistic explanation, but can and does function as a representation, so can be legitimately used in mechanistic explanations that are not fully reduced to causal interactions. So, Shea would contend that my previous argument is too demanding, and indeed, that treating representations as functional components is an appealing feature rather than a shortcoming – it allows us to draw similarities between networks that have different implementations, which if connectionism is a good model of real cognitive systems, would allow us to make meaningful psychological generalisations across individuals and perhaps even species.

Nevertheless, there is still reason to be cautious before accepting Shea's diagnosis. Namely, Shea's conclusion fails to avoid Ramsey's (2007) key demand – that they must function as representations by standing in for some content. Above we applied Ramsey's argument against the weight vector of the network, now we must apply it to cluster representations. Through Ramsey's lens, it becomes clear that cluster representations are also just a description of dispositions. This shouldn't be surprising, as cluster representations are nothing more than a clever way of understanding the weight vector of a network. Cluster analysis helps us group inputs together, first inputs in the training set, and later novel inputs, by the output they generate. Indeed, it does reveal to us how the system itself is grouping the inputs, and such understanding is valuable to understanding how the network works on a functional level. However, clusters are not playing a properly representational role – that is, they are not standing in for some content. They certainly seem to be relevant (following Shea) for explaining how networks discriminate between certain kinds of input, and to explain why some networks fail. The most we might want to say though, is that these clusters indicate the network's sensitivity to certain classes of input, and that the network has cleverly learned a weight vector that facilitates that sensitivity. If generalizable, that insight alone is a huge step forward in our understanding of connectionist systems. However, it would not be correct to claim that clusters are representations of their content; rather, they facilitate complex dispositions which result in a pattern of output which looks like the operation of a representational system. We might even legitimately deploy Dennett's intentional stance to facilitate predictions about these system's behaviour.

More charitably, we might argue that taking Shea's analysis as reducing to simple dispositions is not wholly accurate. Recall our chapter two discussion of Shagrir (2012), who showed how we might apply the tools of DST to root out structural isomorphisms in the phase-space of neural networks. The presence of these isomorphisms show that whilst the network might be understood in terms of dynamic dispositions (cashed out as differential equations), those dispositions are *structured* in a particular way which allows us to leverage Cummins' (1996) solution to the symbol-grounding problem. In the same way, Shea (2007) has demonstrated that at a certain level of analysis, a connectionist system possesses a certain shape, that allows us to identify structural properties in terms of clusters. Whether or not that structure can be used to ground a representational scheme depends only on whether there is an appropriate interpretation that maps that formal structure to some content (Cummins, 1996), and this is what Shea (2007) has shown.

Precision modulated prediction signals are the output of knowledge nets. So if we admit this charitable interpretation of Shea (2007), and the possibility that the dynamics of the PP system's generative model may succumb to an analysis similar to that offered by Shagrir (2012) in the case of the oculomotor control system, then predictions will be grounded in virtue of the structure of their producer in which they participate as outputs. So there is a good case to be made here that these too pass Ramsey's challenge (2007).

Error signals are a precision-modulated function of prediction signals and the raw data from the layer below or from the sensory periphery. As we learned in part two, they play several roles in the system, but their most basic role is to inform the layer above where its predictions went wrong and facilitate efficient learning to improve them in the future. This primary role is as a driving input for the system's learning algorithm, which would only appear representational if we adopt a fully representational attitude to the mechanism. That is, it is only if we believe that the learning algorithm requires representational inputs that we would think that this function of the error signal is representational. I would argue that a representational interpretation is a clear distortion of the true function in this case – a driving signal does not function as a stand in for a system, it functions as a simple input. Indeed, error's other functions are as similarly non-representational driving signals – for instance its role in long-range context modulation. In this case the error signal is delivered to a knowledge net much higher up the hierarchy to provide that net with a quick update about what is going on nearer the sensory periphery. Again these signals are functioning as simple drivers for the processes in high level network allowing them to adjust their output appropriately.

At this point there is an obvious objection – the representationalist can allow that error signals function as driving signals, but maintain that driving signals can be representations – that in the broader context of the mechanism, these driving signals only have that function because they represent something useful for the system consuming them. This is especially clear in the second example, context modulation. The reason the higher-level networks take error signals from the sensory periphery as input is because they carry valuable, real-time, information about the external world which the network can then use to better reduce the long-term incoming error.

This objection strikes the nerve at the heart of many disagreements between representationalists and sceptics. Here the information carried by error is supposed to be its content. That is, error functions as a representation because it stands in some direct, lawful relationship with external states of affairs. This is what makes error signals useful for context modulation, and it is what invites the intuition that error signals allow networks to learn about the world. This kind of representation is what Ramsey (2007) labels 'receptor' representation, and dismisses as a legitimate form of representation. Ramsey argues that receptor representations fail to fulfil his job description challenge because they function only as causal mediators, rather than playing a standing in role. One of the powerful arguments to this end runs as follows: if we allow all causal mediators to be called representations, then the universe is filled with representations, almost every object and process is a causal mediator of some sort and so almost every object is a representation. What is important to whether a signal is a representation is not whether it is a causal mediator, but whether it is used as a stand in for some content. This is where error signals' application for the representation job is turned down. They are just being used as causal mediators, and do not perform the representational function of standing in for content. The simple indication of this is that error signals are not decoupled from their source. If we detach a PP system from the external world entirely, then there will be no input, so no error flowing. We can thus admit that error signals carry information, without conceding that error signals are representations.

In this section I have aimed to provide arguments to motivate doubt towards the representational posits in the computational explanations provided by connectionist PP models. Despite trying hard to find a strong argument for Ramsey's sceptical conclusion, we found that Shea (2007) offers a strong analysis which identifies a deep structure possessed by some connectionist networks, which gives us good reason to believe, when combined with Shagrir (2012)'s concerns discussed in chapter two, that representations in generative models or knowledge nets and their output – precision modulated predictions – can be appropriately grounded. When considering error, we found that these information flows function as causal mediators, and fail to be grounded as stand-ins for contents. Here we made the important distinction between information bearing signals and representations.

I recognize that there are vast portions of literature that I have bracketed here, and numerous concerns remain unaddressed. However, in the following section, I will address the most pressing issue – that of intentional content. We have seem how representational states in PP systems might be grounded, but have not yet solved the problem of intentionality which has been left open since chapter one.

7.2 Function and Intentional Content

As we saw in our discussion of teleosemantics in chapter one, function provides the normativity necessary to address the problem of error for mental representation. That is, only a functional understanding can allow us to naturalise a representation's ability to be true or false – if a component's proper function is to represent a cow, but in fact is also activated by a horse, then that component is misrepresenting, it is incorrect, or false. The nature of representational content (in a computational, mechanistic framework) is thus inextricably bound up with our more fundamental understanding of mechanical functions. If we want to believe that talk of content is useful, we must also believe that function-talk is justified. I believe that a decent amount of the confusion around mental representations in current science stems from neglecting this issue, because without solid reasons for believing in 'proper functions', that is, functions that are natural facts, we have little reason for believing in proper, or real, content. It is important to note that the project of mechanistic explanation does not live or die based on what we think about function, as we will see, but mental representation in mechanistic systems may well do.

I will begin this section with a discussion about mechanistic function, and whether the current best arguments in the literature point us toward a belief in proper function, or whether they point to some alternative. We will consider the popular alternative to the teleofunctionalist/selected effect account outlined in chapter one, namely the *perspectivalist* view argued by Carl Craver and Frances Egan. I will argue that the arguments of the perspectivalists strongly outweigh those of the teleofunctionalist, and that the notion of function is critically dependent on the interests of scientists. In the second part of this section I will apply the import of this work to our interest in representational content. I will argue that the perspectivalist notion of function provides the fruitful and flexible basis necessary for the progressive use of mechanistic explanations. However, I will further contend that perspectivalism does not provide a solution to the problem of intentionality, and thus mental representation. The extension of this argument is that mechanistic explanation, with a proper perspectival interpretation, fails to ground mental representation adequately.

Motivating Perspectivalism

Chapter one detailed the classic objection to teleofunctionalism – Davidson's swamp man (Davidson, 1987). Recall that on the teleofunctionalist account, a spontaneously created human being, swamp man, who is capable of having philosophical conversations, would not be regarded by the teleofunctionalist as possessing mental representations, whilst his physically identical interlocutor *would* succumb to a representational analysis. In this example, we would have two physically identical systems, and one would be thinking, but the other would not.

However, one important motivation for etiological theories is that in order to make any sense of biological systems, and to develop effective interventions, we need a theory of what biological components and processes are supposed to do. However, it is this fixation on biology that lets these theories down as a suitable foundation for mechanistic function, understood more broadly. Biological systems are just one class of mechanism, and despite Garson's laudable defence of a more pluralist understanding of selective effect theory (Garson, 2012), not all mechanisms derive their function in virtue of some selective process, and certainly not in virtue of the small class of well-worked out selective processes Garson develops as suitable for grounding proper function.

Carl Craver (2001, 2013), drawing on the work of Robert Cummins (1989, 1996), has developed an influential alternative to the notion of proper function – perspectival function. Perspectival function consciously accounts for the strongly recursive, hierarchical nature of mechanical systems, as the function of a component or process depends on the explanatory perspective we are taking – each component or process can itself be thought of as a system with its own components and processes. As these elements derive their function from the overall function of system, they can then be treated as a system themselves, and the elements that make them up derive their functions in virtue of the function of the system from this new perspective. In an attempt to maintain some clarity in this discussion I will call a system-function a capacity, and reserve function talk for the elements that make up the mechanism.

For example, our system of interest might be a fridge, and we might ask "how does this fridge cool down food?" We have identified a system (fridge) and a capacity of that system we are interested in (cooling down food). We identify components of the fridge – door, seals, cavity, radiator, air pump etc. and then explain a) the role-function of each component and how all the components are related so as to bring about the cooling of food. However, we might then be interested in asking "how does this air pump pump air?" to derive a full mechanistic explanation. In this case the function of the air pump is the explanandum, and so the system capacity, even though it is part of the explanans of our main question. The workings of the air pump - the role-functions and causal profiles of its components – explain how it is that the air pump pumps air, but do *not* help justify the presence of air pumps in fridges. That is, if someone were to question whether the air pump was in fact an air pump, one only needs to consider its role-function for the larger system, not the underlying mechanics[13].

So far, so uncontroversial. The perspectivalists' contention is that both relevant functions and capacities are dependent on the perspective of the explanation. That is, there are no privileged capacities of a system – there are just a set of phenomena that the system brings about, any one of which we might be interested in explaining, and none of which have primacy. Perspectivalists are thus attempting to battle the intuition that any system has some function it is supposed to perform, as the function being performed depends on one's perspective, especially if one is attempting to provide an explanation, as one has to carefully delineate exactly what capacity of the system it is that one is providing an explanation for.

From this perspectivalism about capacity one can derive the more radical conclusions about the role-functions of components in a system. For instance if we ask, 'how is it that the car races so fast?' The rear wing on the car might have the function to generate downforce (*x*) when the capacity of interest is to race as fast as possible. However, in some cases the question might be 'how is it that the car impresses people?' In this case the capacity of interest is to 'impress people', and in that case the rear wing's function might be to look cool (*x'*). So whether the rear wing performs *x*, or *x'* is thus dependent on which capacity of the car we are interested in explaining.

[13] It is worth noting how this point feeds back into the earlier discussion regarding knowledge nets. Though Hinton (2007) provides a way of identifying contentful structures in the hidden layers of generative models, this does not help us answer the question 'does the generative model represent?', as the answer to that question is concerns the generative model's functional role in the rest of the system, not about the functional arrangement of its components.

But what justification can we offer for rejecting the idea that some functions are privileged over others? It still seems as if the car's proper function is to go as fast as possible, and not to impress people. Just as it still seems justifiable to say that the heart's proper function is to pump blood, not to indicate an animal's level of physical exertion. I argue that these intuitions about proper function are motivated by biologists' concerns. Once we step out of a biologist's mind set, and think more broadly about the application of mechanistic explanation, we find that to privilege one function over another is to limit the applicability and thus fruitfulness of the framework. The simple demonstration of this is that not every system that one might wish to explain mechanistically is susceptible to an etiological, selective reading. One might give a mechanistic explanation of the process of covalent bonding in chemistry, here there is no selective process at work – the only function that allows us to ascribe function to the elements of the mechanism is the capacity of interest to the scientist, viz. the capacity to bond covalently. We do not lose any of the normativity of function – we can explain the failure to bond covalently in terms of some part of the system failing to play the right role, insofar as it is a part of a covalent bonding process, that is, from the perspective of covalent bonding.

A second point to demonstrate is that biology doesn't need proper functions. This is a major contention by etiologists for whom the necessity of proper function in biology is a primary motivation for finding a satisfactory theory. However, once a mechanistic framework is adopted, one can view biological systems as just one (albeit especially interesting) kind of mechanism. But from the perspectivalist's point of view the functions that etiologists believe are special, are special just because we find them especially interesting. This kind of approach is more than strong enough to justify the claims of biology, because they are just that, the claims of biology, not of social science or psychology. It is thus understandable that biologists find their inspiration for which functions are especially interesting based on evolutionary concerns – evolution being the force that has driven the formation of the phenomena they are interested in. Nevertheless, they can afford to be perspectivalists, as perspectivalism leaves them free to choose their own approach whilst not belittling other approaches which appeal to different kinds of function.

Thirdly, once a proper function has determined by the scientific community (presumably as a result of a particularly convincing evolutionary just-so story), an intellectual barrier has been created for alternative ways of approaching the system. In the language of the perspectivalist, once perspective has become dominant and research from other perspectives is underfunded and understaffed. Once we adopt the perspectivalist view however, we see how absurd this attitude is. Proliferation of ideas and methods is one of the most important drivers of scientific progress. Perspectivalism about function encourages proliferation, while defending proper function hampers it.

So far in this section I hope to have motivated some doubt regarding the necessity and theoretical credibility of proper function, and I also hope to have provided some support for a perspectivalism regarding function, which is a superior way to understand function for use in mechanistic explanation as it permits a broader use of function-talk, whilst preserving normativity, so not diminishing the notion at work in biology. In the following section however, I hope to show that the perspectival view of function does have consequences for how mental representation is understood in mechanistic explanations.

Perspectival function and intentional content

Aside from the differences outlined above, perspectivalism and selected effect theory have one other important difference: the ontological status of function. For the perspectivalist, there is always the acknowledgement that functions are only fixed once a subject has taken a specific perspective on the system, whereas for the selected effect theorist, the proper functions are always the same, regardless of when, where, or by whom the system is being examined. As I have mentioned, whether one is a perspectivalist or a selected effect theorist has little to do with the quality of the explanations one can provide, and is usually just an indication of the domain that is the focus of one's work. In this section I will argue that whilst this is 'just' a philosophical difference, and despite the fact that mental representation is a central posit in cognitive science, that it is nevertheless an unobservable theoretical posit, and so this difference between the teleofunctionalist and the perspectivalist has dramatic ramifications for the status of mental representations in mechanistic explanations. On one hand the teleofunctionalist boasts an observer-independent notion of intentional content, on the other, the perspectivalist admits that intentional content is dependent on a subject.

My primary goal for this argument is to show that neither the teleofunctionalist theory of proper function can ground intentional content, nor the perspectivalist. We have already discussed at length the problems for the teleofunctionalist, so here I will argue the point against the perspectivalist. The conclusion will motivate two attitudes. First, an openness to alternatives to mechanistic explanation in cognitive science. For instance, dynamical systems theory offers covering-law explanations, and in chapters eight and nine I will be leaning on this kind of approach. If the reader is strongly wedded to mechanistic explanation as the only appropriate approach to cognitive science, then I doubt they will find my arguments convincing – the purpose of my arguments is not to convince this reader, but rather to present my own motivations for taking a different route as reasonable. The second is that once we are open to alternatives, we might, paradoxically, be in a better position to understand function. Once we have seen the importance of dynamic processes for making sense of cognition, a robust notion of function does emerge, and with it, intentional content. I will call this *proper perspectivalism* and will argue for it in detail in chapter nine.

To this end, allow me to note once more that mental representation is a *functional kind*. The term functional kind refers to a category of mechanical elements that are defined in terms of their functional properties rather than their causal properties. Functional properties are typically identifiable as they refer only to the way the element contributes towards the overall system capacity, rather than the causal properties possessed by the element irrespective of the system capacity. That a component or process possesses intentional content is a claim about the function of that component or process. This embeds the element in a mechanistic milieu which provides the explanatory context for it to play this function. In another milieu it may play a different function. For instance a neural structure may function to represent the smell of wet paint when positioned at the right point in the olfactory bulb, but we can imagine that if we artificially transported that structure into the right place in the visual cortex, it might represent the appearance of wet paint[14]. Thus an element represents in virtue of its functional profile, rather than its causal profile, and its functional profile is partly dependent upon the relations it possesses with the broader system.

It is worth fleshing this out through comparison with other functional kinds, e.g. pumps, vessels, energy sources, switches. Imagine an object that is typically used as an air pump, but in another context can be easily repurposed to function instead as a water pump. The brute causal powers of the object remain the same, but its functional properties are what are relevant for our explanation of how they help solve a certain problem. Indeed, what we call the air pump can fail to function as a pump at all, if say, it is only ever used as toy gun as part of child's play, in which case it is functioning as a token in a kind of game with certain rules governing its use (e.g. if I point it at you and say bang, then you should fall over and drop your own 'weapon'). This is an instance of the same physical object playing three different functional roles, each of which is a functional kind – the object is an air pump when it is used to pump air, it is a water pump when it is used to pump water and it is a toy gun when it is used as such. Indeed, any object used to pump air is an air pump, and any object used as a toy gun is a toy gun precisely because the facts about object implementing that function are not relevant beyond the facts that enable them to fulfil their function. So, when thinking about intentional content, we can ask 'does this object have what it takes to possess intentional content?', but that is a separate question from 'does this object function as a representation in this mechanism?'. We may answer yes to the former while answering no to the latter. That is, a component may have the requisite causal properties to be a representation (passing Ramsey's JDC (2007), and thus avoiding the symbol-grounding problem), but does not bear the right kind of relationship to the rest

[14] Though this example is clearly absurd, I mean it to demonstrate the principle that the same physical structure can have different functions in different mechanisms, or even within the same mechanism – e.g. a cog in a car may function to drive the differential, but may by coincidence be the right shape to also function as a gear in the gearbox.

of the system that permits it to function as a representation of intentional content (that is, it fails to avoid the problem of intentionality). Thus, a *mental* representation is a functional kind, as a whether a state possesses *intentional content* is defined in terms of its functional, rather than causal properties.

Under the selected effect theory of function, some function is a proper function if it has been selected for by natural selection, neural selection, or some other selective process. It is reasonable to think that most elements in a neural system have been selected by natural or neural processes, having been determined either by genetics or by the brains own learning processes. Thus the selected effect theory would support the claim that the proper function of representations is to be about some content. They thus malfunction when they fail to stand in for those contents, and this would be a case of misrepresentation. Crucially, because all these functions derive from biological selective processes, they are objective facts about the workings of the brain. In this way, selected effect theory purports to provide a solid ontological foundation for mental representation. That is, if there were not independent philosophical objections to this claim, then selected effect would be a perfectly good way of identifying a state's intentional content.

On the other hand, perspectivalism offers no such security for bearers of intentional content. For the perspectivalist, the function of an element in a mechanistic explanation does not derive from a selective process, but derives purely from the capacity of the system and our interest in explaining it. There are no limitations on the perspectivalist about what capacity is a proper capacity of a system, though in biology and cognitive science, this capacity will usually be determined by how the system in question operates to contribute to a larger system such as the organism or the brain. However, tacit in these explanations will be an assumption about the broader function of the organism/brain. Despite being tacit, these assumptions are nevertheless perspectival – we might be interested in the brain as a regulator of hormones, or as a control system, or as a mind; we might be interested in the organism as a unit of natural selection, or as a part of a society, or as part of an ecosystem. The perspectivalist does not assign special status any of these perspectives, as each may be of legitimate scientific interest – indeed, it also allows us to welcome new and radically different perspectives whilst preserving a familiar mode of explanation.

An important consequence of declining to assign special status to any perspective is that every possible function that we might ascribe to a given element in a system is equally real (or equally unreal). Which function is relevant is purely dependent upon the perspective of an explanation – so their ontological relevance is radically relative to explanatory concerns. The function of the air pump in my example above depends on how it is being used at a given time, or even how it could imaginably be used by a subject. What the perspectivalist recognises is that in many cases an object can be fulfilling multiple functional roles at the same time, when considered from different perspectives. So one unattractive option is to conclude that the element in question instantiates all these functional kinds at once and we embrace a kind of relativistic pluralism, in the case of intentional content, this would amount to radical indeterminacy of meaning – a representation's intentional content would encompass all the possible interpretations that a perspectivalist could possibly ascribe to that state. On the other hand, we might deny that any functional kinds are real, (representations with intentional content included) and instead are only useful epistemologically, for explanation, as a way of understanding how a system works, which would lead us to *fictionalism*.

Fictionalism takes the conclusion that functional kinds do not refer as a given – instead treating functions as useful fictions that permit us to provide explanations of phenomena and guide research. Relativist pluralism is the position that functions are real from a given perspective. That is, an element in a mechanism is a representation in virtue of its function for the system; its function for the system is entirely dependent upon the explanatory perspective. However I believe this line of reasoning has radical consequences, leading us to, at best, the fictionalism which they are trying to avoid. This is because function that is entirely dependent upon explanatory perspective is observer-dependent, and things that are observer-dependent are not objectively real; therefore kinds defined in terms of function, functional kinds, such as mental representations are observer-dependent and not objectively real. We also might take the radical indeterminacy of content that results from the perspectival view, outlined in the paragraph above, as a *reductio* to the same conclusion. So, perspectivalism about function, whichever way we interpret it, entails that functional kinds such as mental representations invoked by mechanistic explanations are not real.

The perspectivalist might object that we are attacking a straw man. Rather, once a perspective has been fixed, that is, once the system capacity of interest has been carefully and explicitly identified, then the role-functions of elements in that system are fixed, indexed to that particular capacity. So, with respect to a certain (real) capacity of the system, the elements of the system have a singular, definite, and real function. This position is importantly different from the fictionalist. The fictionalist maintains that functional kinds are useful fictions, whereas the perspectivalist holds that functional kinds are real, but depend on what interests us about the system. It is arguable that such dependence seriously undermines the reality of the ascribed function. What other kind of scientific reality depends so heavily on the interests of the examiner? Even if this is palatable, it certainly suggests that there might be a better way of carrying out our analysis, that we are able to further divorce from human perspective. On the other hand, the perspectivalist might respond that once the explanatory perspective has been fixed, the physical structure of the mechanism will serve to radically constrain the set of legitimate functional analyses. In this sense, perspectivalism is not making a radical claim – not all ascriptions are fair game, there are right and wrong answers, once the perspective is fixed. However, fixing that perspective is the tricky part, from the point of view of the sceptic. It might be true that once a perspective is fixed that everything else follows, but on what principled basis can we privilege a certain perspective? Of course, this privileging is exactly what the perspectivalist wants to resist. But if all perspectives are legitimate, then we fall inextricably back into radical pluralism because every capacity of a system is equally special, and so all functions deriving from all possible capacities are equally real. In chapter nine, I will speak to both the sceptic and the perspectivalist by providing an account of function that provides naturalistic criteria for legitimate perspective taking. This will endow the perspectival account with a principled defence for taking certain perspectives, and thus avoid the sceptical conclusions.

For now, the sceptical position results in a bold conclusion – that functional kinds are just useful fictions. But there are all sorts of functional kinds that are clearly real: pumps, valves, cogs, transistors, computer mice, etc. I argue that this is a special problem for mental representations because they are not only functional kinds but also unobservable, theoretical entities. Other functional kinds, such as those listed above, have an observable extension that allow us to verify their reality. We might give a mechanistic explanation of a road traffic network, and this might identify road signs as a functional kind, indeed as a kind of representation, but their intentional content is grounded in the convention of users, we do not require a naturally grounded account of intentionality to account for this sort of derived meaning. Mental representations do not have this quality. Mental representations with intentional contents are theoretical constructs made as part of computational explanations to help us make sense of what cognitive systems are doing and how they bring about the phenomena we are interested in. The problem of intentionality requires *determinate underived* content, so what we are searching for is a *determinate functional kind*.

The way philosophers of science usually justify belief in theoretical posits is by appeal to inference to the best explanation. For example, we believe in electrons despite not being able to observe them because they are a crucial part of our best theories in physics which explain a huge number of observable phenomena. These are good explanations not just because they account for a large number of phenomena, but also because they have been instrumental in engineering new technology, and in the development of new theories. Unfortunately for states with intentional content, the fictionalist position entailed by perspectivalism about function is an obstruction to our inference to the best explanation. As the fictionalist holds both that mental representations are useful posits for explanation, and that they are not real, the realist's inference to the best explanation fails. Mark Sprevak (2013) addressed this uncomfortable result of fictionalism about mental representation by noting that holding this position requires either rejecting that representations are part of our best explanations in cognitive science, or rejecting the use of inference to the best explanation. Neither of these options are attractive, but it seems undeniable that mental representations do play a central role in explanations, so we must reject our inference to their reality. In either case, the representationalist is in trouble.

There is an objection the representationalist might make to my reasoning. They might reject the necessity of IBE for drawing realist conclusions. Most famous among entity realists pursuing this line is Ian Hacking, who developed experimental realism. His main contention is that that the success of a theory is not important to our realist attitude towards entities that appear in that theory. Rather, we should be concerned about the degree to which we can manipulate our theoretical posits. Crucially, if we can manipulate these entities in such a way as to investigate the claims of other, unrelated theories, then we are justified in believing in their existence. Of electrons, Hacking famously said, 'if you can spray them then they're real' (Hacking, 1983). It is clear that we do manipulate mental representations in this way in neuroscience (Bechtel, 2016), so we are justified in our belief in them, without having to use a problematic inference to the best explanation.

A classic objection to Hacking's experimental realism is that it is in fact an inference to the best explanation in disguise (for a review of possible interpretations of Hacking's argument see Miller, 2016). The lynch pin in Hacking's argument – the use of unobservable entities in experiments that test other theories – is just the stipulation of a very specific kind of theoretical success. That is, the reason this is such a strong indicator of entity referral is because only a really good theory would posit entities that can be used in this way. The success is still the success of the theory, not of the entity, and thus the reality of the entity cannot be torn from the theory in which it plays an explanatory role. Another way of thinking about this is to consider the question 'why did experiment *x* about theory *y* using entities *z* yield useful results?' for this we must provide an explanation, and an important part of that explanation will flesh out the details of the entities, *z*, which derive from alternative theory *y'*. In this fully spelled-out version of Hacking's argument, the belief that *z* are referring entities depends upon the virtues of this explanation, and thus we haven't avoided inference to the best explanation at all.

In this section I have briefly introduced the debate about function in mechanistic explanations. I endorsed perspectivalism as the most attractive option, due to grave problems with etiological theories, and the virtues of perspectivalism as a liberal approach that doesn't hinder conceptual development, and offers a strong foundation for all sciences making use of mechanistic explanation, not just biology. I went on to draw out the implications of this view for mental representation, and concluded that perspectivalism entails that determinate functional kinds such as states with intentional content do not have a place in our ontology. Having also provided some rebuttal to prima facie objections, I intend that these arguments together provide some reasons to doubt that mechanistic explanations support representational claims in cognitive science, and thus motivate the search for an alternative explanatory framework.

7.3 Conclusion

In this chapter I have discussed the nature of and problems with representation in a mechanistic understanding of PP on two fronts, the first was highly focused, the second was extremely broad. Together, I hope to have provided a way of understanding representation in PP systems, but shown that the deep problem of intentional content still remains. In section 7.1 I used William Ramsey's job-description challenge to dispute that any of the candidates for representation-hood in PP systems were really fulfilling a representational role. However, after lengthy discussion, it seems that knowledge-nets encode structured dispositions that can be understood as grounded representations. On the other hand, errors seem to fail to be decouplable from the environment, thus cannot operate as stand-ins for their content, but rather as causal mediators. In 7.2 I showed the extent of the challenge posed by the problem of intentionality. This challenge applies not only to teleological notions of function, but also perspectival notions. As a determinate functional kind, mental representations do not find support for their reality in virtue of their role in good mechanistic explanations, and thus the usual inference to the best explanation to support the reality of a theoretical posit is not available for advocates of mental representations.

In the remainder of this thesis I will move on from mechanistic explanation. In chapter eight I will consider the Free-Energy Formulation of PP. I will follow work by Allen and Friston (2016) and Bruineberg, Kiverstein, and Rietveld (2016) arguing that this conceptual foundation of PP suggests a happy marriage with 4E cognition and the explanatory strategies that they tend to employ. These strategies provide us with a new way of interpreting the representation talk used by PP theorists, and we will look at the implications of this interpretation for cognitive science and philosophy. In chapter nine I will use Bickhard's work on representation to dig a little deeper into the embodied and enactive approach to PP, and how we might actually use the covering-law explanations they give to provide a fruitful and justifiable way of using representation talk to explain PP processes.

Chapter 8: The Free Energy Formulation

We have come a long way with the project of investigating the nature of mental representation in PP systems, having gained a working understanding of the systems themselves in part two, and having considered the prospects of working out an explanatory strategy that is at once mechanistic, representational, and applicable to PP systems. My emphasis on the 'big picture' should also be clear, as I draw many of my argumentative strategies from zooming out and viewing the context, both empirical and philosophical, of the current problem. In this section I will use this technique again, and analyse the way representation-talk is actually employed by proponents of the Free-Energy paradigm in biology and cognitive science (e.g. Karl Friston), which I will be referring to as the FEF (Free-Energy Formulation).

FEF is in some ways the conceptual grandfather of PP, and is used to both motivate and ground PP in a broader scientific context. With section 8.1, I will justify this broadening of focus by following Jakob Hohwy's arguments to the effect that 'if you believe PP, you better believe FEF'. Further to this, sections 8.2 and 8.3 will go some way to anchoring FEF to our understanding of PP, as I will outline several ways in which FEF informs the current problem of representation in PP, and also suggests that DST is a viable alternative to the usual computational expression of the framework. In 8.2 I will be examining FEF's close ties to 4E cognition, and attempt to resolve some tensions between the two frameworks caused by their differing attitudes towards representational language (Allen and Friston, 2016). In 8.3 I will argue that the watered-down concept of representation used by FEF is too weak to ground its function for cognitive science, and that acknowledgement of this demands a reinterpretation of the framework in non-representational terms. To this end I will demonstrate how FEF lends itself to the covering-law explanatory strategy favoured by dynamical systems theory and radical embodied cognition, leaning on the work of Bruineberg and Reitveld (2016), and Chemero (Chemero, 2011; Lamb and Chemero, 2014). Finally, in chapter nine, I will argue that a useful notion of function emerges from an ontology of dynamic, structured processes, which we can then use to make sense of intentionality.

8.1 If you believe PP, you better believe FEF

During his 2016 talk at the Centre for Cognitive Neuroscience, Dartmouth University, Jakob Hohwy gave two arguments. With the first he concluded that if you advocate the PP framework, then the tools provided by FEF strengthen your position – if you believe PP then it's better for you to also believe FEF. In the second he argued that FEF situates PP as part of a powerful, unifying framework in biology that's on the rise – so if you believe in PP, then you better believe in FEF, or get left behind. In this section I will rehearse these arguments with my own focus in order to motivate an important link I want to draw between the conclusions here and the issues concerning representation in PP systems.

Avoiding the dark room

Let us recall the dark-room problem from part two. This is probably the most obvious and most common objection to PP theory. If we are just in the business of making accurate predictions about sensory data, the sceptic asks, why don't we just sit in the corner of a dark room and never move? After all, if we are alone in a dark room, we will very quickly learn to make perfectly accurate predictions, it's a predictive processor's dream scenario. The sceptic continues – so, either human beings are terrible predictive processors who consistently fail to minimise prediction error in the most efficient way, or PP is simply wrong-headed, and we should search for alternative theories. This nasty problem clearly requires an urgent reply from anyone who takes PP seriously as a theory of cognition.

Now let us imagine that we are advocates of PP who reject FEF. In order to respond to the problem, we would likely invoke some set of priors, residing deep in our cognitive system that encodes the expectation that we will not sit in a dark room for an extended period of time (or experience hunger, thirst, loneliness, boredom, etc.). These deep priors, presumably protected in some way from the system's learning processes, ensure that we will never predict a dark room situation for very long, thus keeping us up and about, processing error often, but in relatively small amounts, which is preferable to the large prediction errors which would be suffered if we acted against the deepest ‘hyper’ priors in the system.

This does help the advocate avoid the dark room problem, and hypothesising hyper-priors in a PP system can help us make sense of many related puzzles. However, this habit is ultimately degenerative for the PP research programme. It highlights another common charge against PP – that it is unfalsifiable. By explaining away these problems by hypothesising convenient hyper-priors, we seem to be making ad-hoc additions to the theory, for the sole purpose of avoiding falsification of the theory. Whilst philosophers of science have debated back and forth regarding exactly how bad this practice is (c.f. Popper, 1963; Lakatos, 1978; Feyerabend, 1975) they all agree that it is somewhat irrational, if not downright unscientific. The prevailing consensus today is that consistent ad hoc modifications are a good indicator of a research programme in decline.

The PP advocate who also buys into FEF, and takes PP to be a local instantiation of a global free-energy minimising process entirely avoids the charge of making ad-hoc additions. Their response to the dark-room problem will be essentially the same, that deep unchangeable aspects of the generative model will result in large amounts of prediction error if the organism were to stay in a dark room for too long. However, the FEF advocate has good in-principle reasons for these claims – that not just the brain, but the whole organism is a free-energy minimising system, and that it is the organism’s phenotype that imposes these constraints. In other words, the whole organism is the ‘generative model’ and that includes all the low-level evolutionary adaptations that demand that we eat, drink, and explore. Of course, FEF is not the only high-level theory that the PP advocate might appeal to in order to explain the existence of appropriate hyper-priors. In fact, one might find it more natural to appeal to evolutionary theory to account for the installation of effective priors. Indeed, the goal of this chapter is not to unequivocally demonstrate that FEF is true, or even that it is necessary to understand PP, but that it enriches our understanding of PP. To this end, FEF is richer than evolutionary theory for explaining these hyper-priors. That is, free-energy minimising systems helps us understand *how* PP systems have been evolutionarily successful: because they help to ensure that the system remains within its tight bounds of viability, which are tacitly encoded by the basic physical properties of the organism itself. One of our conditions of viability is that we require food, water, exercise, and sanitary living conditions. These needs, or 'hyper priors' are quickly violated in the dark-room situation, leading to error, which signals that if we remain in the situation for much longer we will be outside our bounds of viability, and more than likely die.

FEF has broader horizons than PP, applying to all biological systems, or indeed any kind of self-organising system. It thus situates PP in a way that explains why, in the case of a given cognitive system, it is rational to make assumptions about certain kinds of hyper priors. Without this broader context, PP looks as if it must make increasingly ad hoc additions to its theory in order to account for particular behaviours that do not result naturally from the PP framework[15].

This is the first prong of Hohwy's case – that PP is a bad scientific theory when it is not embedded as part of FEF, and that it is a good scientific theory when it is coupled with FEF. Despite the fact that FEF doesn't directly improve PP's empirical credibility, it does improve PP's conceptual credibility. Under FEF, PP is a progressive research programme whose theoretical development is driven by certain guiding principles, rather than by ad hoc modifications made in the face of otherwise crippling shortcomings.

Unification

Hohwy's second prong goes further. FEF's role in the PP research programme is much deeper than merely the supply of a friendly theoretical framework. The core principles of PP models can in fact be derived from FEF. This relationship doesn't forcefully

[15] A serious potential objection to this argument is the FEF also derives from evolutionary concerns, however this is a concern that runs through several arguments I make both here and in section 3.3, so I will address this fully in part four.

oblige the PP advocate to also agree with FEF, there are other virtues of PP that make it attractive apart from its relationship to FEF – it is an efficient and powerful schema for implementing Bayes-optimal reasoning, for example. The FEF sceptic might hold that PP is just one of the brain's 'bag of tricks' that helps it approximate Bayesian inference, that it is a useful but not a unifying theory.

The key manoeuvre Hohwy performs to motivate his argument is to, again, contextualise this attitude. The bag of tricks approach neglects the fact that we are not just agents that are able to approximate Bayesian reasoning, but we are systems that approximate Bayesian reasoning in a "complex, changing and uncertain world with relatively stable agents". Hohwy recommends that a system in this situation will have or develop processes that reflect the scientific method of inference to the best explanation.

Explanatory virtues, the argument goes, all help reduce prediction error in the long run. That is an important part of what makes them virtues for scientific explanations, and it is also what relates them to PP. For example, a PP model that is simple is unlikely to be overfitted to the data, a PP model that is integrated with broader knowledge will be more able to account for confounding causes that may generate long term prediction error in narrower models, and most importantly, a PP model that is fruitful, that is, a model that recommends actions that in turn generate good predictions, will best minimise error in the long run.

It is these explanatory virtues that fall neatly out of FEF. System complexity is minimised automatically by the mathematics of FEF, thus ensuring that the model is as simple as possible; free-energy minimising systems are context sensitive, ensuring that they integrate the knowledge of their current state with long term knowledge; and finally the process of active inference that free-energy minimising systems use to directly reduce surprisal, corresponds precisely with what we expect from a system that favours fruitful models. These points all evidence the way the formal architecture of FEF scaffold an effective prediction-error minimising, that is, PP, system.

So, although one might believe that PP is just a trick in the brain's bag of tricks, Howhy presses that PP's trick is precisely described by FEF – the difference is that FEF encourages us to treat this as a more universal theory of cognition. Whilst this move might be resisted, I argue that this is foolish – for the very same reasons that FEF makes for good PP systems. That is, FEF is a good explanation – it is simple, fruitful, and aims to unify a broad range of domains. Accepting the FEF framework as the theoretical foundation for PP thus improves its explanatory status – it would be part of an extremely virtuous explanatory framework, rather than a theory with relatively narrow scope in an already extremely crowded domain.

The point of this second prong is thus itself twofold – on one hand, the structure of a PP system is well fulfilled by FEF, that is, the mathematics of FEF provide a ready-made framework for a prediction error minimising system as they closely mirror the scientific inference to the best explanation process. On the other hand, FEF brings with it the theoretical weight of an extremely attractive explanatory framework in which PP can situate itself and improve its credibility. Combining these conclusions with those deriving from the problems raised by dark-room thought experiments, we have powerful motivations for believing FEF if we believe already believe PP. Not only does FEF offer a powerful framework with which PP neatly integrates, it also solves serious conceptual difficulties that threatens PP's status as a progressive research programme.

8.2 FEF is friendly to Embodied and Enactive Cognition

In Part Two I very briefly touched on the idea that a theory of cognition grounded using FEF would have natural ties to embodied and enactive cognition (often known as 4E) approaches. In this section I am going significantly expand on that suggestion. As the FEF framework encompasses domains as broad as cognition, biology, and ecology, it will come as no surprise that it naturally pushes our attention outward to these areas when thinking about mental processes. The notion of active inference also clearly requires an active and interactive element to be core to any theory of cognition derived from FEF. The first job of this section is to detail in full the ways in which a free energy minimising mind is embodied, extended, embedded, and enactive. To do this I will follow the work of Allen and Friston (2016), The second job of this section is to address some of the obvious tensions between FEF and the more radical facets of 4E cognition. These tensions are focused on the use of representation talk in FEF. The resolution I offer to this tension is a recognition of the extremely minimal sense in which Karl Friston and other advocates of FEF use representational language. This will set up the discussion in part 3.2.3 which will consider the pros and cons of using representation talk in this minimal way.

Similarities

FEF and 4E cognitive science share similar interests. Both are keen to unify the sciences of life and mind, employing a single explanatory framework that can unify the currently separate sciences in these domains. This is due to their shared belief that the phenomena investigated by these sciences are fundamentally interrelated, and that important lessons are being neglected due to the current state of separation. 4E has typically done this by either borrowing concepts from the life sciences, as in Gibsonian ecological psychology, or by extending mental phenomena out into the world, as in Clark's functionalist extended cognition. FEF on the other hand is a formal framework developed to help us understand systems that are able to resist the second law of thermodynamics, maintaining their order and organisation in the face of cosmic and environmental forces. In this way it may seem closer to theoretical physics, but it just so happens that the systems which the sciences of life and mind are interested in fall into this category of self-organising systems that are capable of maintaining homeostasis with their environment.

Here I will provide a summary of the striking way that details of FEF resemble some of the core tenets of radical branches of 4E cognition. I will consider how free-energy minimising systems are autopoietic, which mirrors the enactivist contention that minds co-determine and co-create with their environments. I will then look at active inference in greater depth, and provide detail on the ways in which action has been invoked as a mental process by 4E theorists, and how this compares to the way FEF suggests it is involved. We will see that the work of Alva Noë on his enactive thesis that perception is based on sensorimotor dependencies, as discussed earlier, can be derived neatly from FEF. Though there are interesting similarities between representation in 4E and FEF, I will expand on these in the next section, as it will be the differences between the notions that are most interesting for the purposes of this thesis.

In order to ground the discussion, Allen and Friston (2016) offer a broad definition of living systems, which helps us demarcate the space of phenomena we are interested in. "By definition, living beings are those that maintain an upper bound on the entropy of their possible states" (Allen and Friston, 2016). This upper bound on entropy demarcates the point at which a living system will no longer be viable, and it dies, succumbing to the second law of thermodynamics. Also crucial to this definition is the notion of maintenance, as it is systems that are involved in the maintenance of their entropy that are living – we are able to create conditions under which certain structures are able to remain within a certain bound of entropy, but we are responsible for maintaining that system using energy – e.g. we can prevent food from decaying by putting it in a refrigerator. It is only systems that are capable of harvesting energy, metabolizing it, and using it to maintain a bound on entropy that are properly understood as living systems.

This definition raises a question that is important in relation to our understanding of the embodied and extended nature of minds and organisms: how do we identify the border between system and environment: what is internal to the system and what is external to it? To make sense of this, FEF makes use of the concept of Markov blankets, borrowed from machine learning. A Markov blanket divides two or more systems that are statistically independent – that is, the probability of system A being in a certain state does not change when system B is in a certain state. In other words, the Markov blanket separates phenomena resulting from external causes from phenomena resulting from internal causes. It is the ongoing maintenance of this boundary that permits a system to remain organised in the face of entropic forces and distinguish it from its surroundings – without a Markov blanket, a system would very rapidly 'dissipate' into its environment (Allen and Friston, 2016). Identifying a Markov blanket then allows us to identify the systems to which we might apply the principles of FEF. However, it also highlights the necessity of a sharp distinction between internal and external, which on the face of it is antithetical to 4E cognition.

Allen and Friston transform this apparent difference into the conceptual foundation of understanding FEF systems as 4E systems. For whilst the Markov blanket operates as a strict theoretical boundary, it is developmentally porous in important ways. In order to maintain its internal organisation and its Markov blanket, a self-organising system must be able to transfer resources from the external world to the internal system. This transfer of energy requires interaction with the environment – the system interacts with the environment in order to maintain its separation from it. Furthermore, there are a limited number of environmental conditions that allow the system to maintain its integrity. That is, in some environmental conditions, the Markov blanket will be rapidly compromised, and thus the system must work to avoid those conditions. For human systems, these are conditions such as extreme heat, extreme cold, lack of oxygen, lack of desalinated water etc. Other organisms may have very different conditions they must avoid, resulting in the diverse web of living systems that flourish all over the earth. The common factor shared by all self-organising systems is they must have ways of interacting with their environment such that they remain within the small set of states in which they are capable of maintaining their Markov blanket. Living systems are only separate from their environment in the special way that allows us to demarcate them from their environment. They are thus separate in only this most fundamental way.

The way that living systems maintain their internal and external states is known as autopoiesis – self creation. So, if the brain is also describable under the FEF, then it too is autopoietic, and this point is a central tenet of the enactivist tradition (Varela et al., 1974). Autopoietic systems involved with their own ongoing creation and maintenance, but also, through their necessary interaction with the environment, are heavily involved with the creation and maintenance of their surroundings, which is necessary for their ongoing survival. That is, autopoietic systems craft their environment to suit themselves. If the brain is autopoietic then it is deeply enactive.

Another way in which Markov blankets are importantly porous is through the abstract process of synchronisation. Statistically independent systems can come to synchronise over time by interacting through a Markov blanket. The process of synchronisation results in high mutual information between the systems, and thus erodes their statistical independence. However this process is not automatic, and depends on both the nature of the Markov blanket and the systems. If the Markov blanket does not permit any interaction between the two systems, then synchronisation cannot occur. The Markov blanket must provide a causal link between the systems. Secondly, the systems must be sufficiently similar in order to synchronise, if there is a large discrepancy between the structure of the systems, then synchronisation will fail. The classic example that brings out these conditions comes from Huygens' C. Horoloqium Oscilatorium (1673).

Huygens describes a set-up in which two pendulum clocks are hung from a wooden beam. The clocks are statistically independent – their pendulums can be swung at different speeds and started at different times. Here we have two systems, the clocks, separated by a Markov blanket, the beam. The surprising result is that the pendulums housed by the clocks tend to synchronise, over time. If, that is, the required conditions are met. When the beam is held fixed, there is no possibility of interaction between the clocks, and their dynamics do not change. However, when the beam is allowed to oscillate freely, synchronisation is possible. We also find that when the pendulums are set in motion with radically different periods, synchronisation does not occur – the systems have to be similar enough that they are able to synchronise.

This process of synchronisation is a kind of free-energy minimisation, and thus the systems are easily described by the FEF and fall into the class of systems we are interested in. Synchronisation demonstrates an important kind of porousness that is formally describable, and yet sets the stage for a deep kind of autopoiesis – the co-evolution of the dynamics of two systems, which through synchronisation, engage in a process of feedback and response that results in both systems shaping each other, and thus the explanation of how each is the way it is importantly involves the other. Thus if the mind and environment are related in this way, as the FEF would have it, then our understanding of the mind is crucially interwoven with its relationship with the environment and the environment itself. Widespread recognition of this would be a major victory for advocates of 4E cognition.

I will now look more carefully at Karl Friston's notion of active inference, relating it to more formal talk of synchronisation. The capacity for active inference is vital for a free-energy minimising system, and this, unsurprisingly, has strong links to the way enactivists understand the mind – as being deeply action-involving, and being constituted by a capacity for action.

Active inference is the process through which a free-energy minimising system changes its environment to bring it into line with its expectations. A simple example to demonstrate this is that babies expect to see faces in the centre of their visual field. Thus when a face is present in their visual field, they will move their eyes and head so that their sense-data conform to their expectation. The mind is thus changing its environment – the body's position relative to a stimulus – in order to verify its prediction. Through active inference then, the FEF says that systems will display extreme confirmation bias – they will rigidly seek out 'evidence' that confirms their 'theories'. This has led commentators such as Bruineberg et al (2016) to note that 'if my brain is a scientist, then it is a crooked and fraudulent scientist – but the only sort of scientist that can survive in an inconstant and capricious world'. It just this caveat at the end that renders this kind of behaviour consistent with our discussion above that treats a free-energy minimising brain as using rational scientific methods.

A scientist's only job is to bring her theory into line with the world by inferring truths from her data. A mind however has a different goal, as we have discussed at length already in this section, viz. to stay alive. It is interesting that FEF should imply that a mind exhibits scientific reasoning, but to insist that it also acts in a way that is consistent with its theory, that is, by confirming its expectations, is absolutely necessary. Recall that a FEF system has the goal of minimising its free-energy in the long run. While changing its internal states to maintain a theoretical bound on its free energy is an important part of this process, the only way it can immediately reduce the amount of free energy (or surprisal) in the system is to change its relationship to the environment through action. A free-energy minimising mind is thus the most innocent kind of crooked scientist, the kind whose life depends on constantly confirming their theories as well as on having good theories.

Action is thus at the very centre of the FEF. It explains how a free-energy minimising systems' capacity to act is both absolutely crucial for survival and also is neatly integrated with our understanding of internal processes and the state of their surroundings. Action is also at the very centre of 4E cognition. Viewing the mind as an entity whose main purpose is to control action, and whose processes are, as a result, action-involving all the way down, enactivists offer an alternative foundation for cognitive science to the classical, passive understanding of cognition. On this initial basis, FEF and enactivism look set to be a good match.

However, this friendship is not just superficial. I am going to now show how FEF and the enactivist theory of perception advocated by Alva Noë are extremely closely related. With this analysis I aim to provide a detailed example of how deep the connection is between FEF and enactivism.

For an advocate of FEF-grounded predictive processing, perception is commonly identified with the process of generating predictions – a prediction is what the system expects to sense at the next time-step, and thus constitutes the system's percept of the next time-step. For an enactivist such as Noë, perception is constituted by sensorimotor contingencies, relationships between percepts and possible actions which enable us to perceive the rich world around us despite a poverty of stimuli (Noë, 2004). There seems to be little in common here. That is, until we dig a little deeper into the enactivist view.

What we find is that sensorimotor contingency is all about hypothetical action, and how the system would expect that action to alter what is perceived. That is, it is about predicting how a scene would change if one were to move oneself in relation to it. Noë's typical example is of a tomato – why do we see the tomato as a spheroid, rather than just as circular? It’s because our perception is constituted by the sensorimotor contingencies which encode the expectation that if we were to rotate our visual perspective around the tomato, its visual shape would remain roughly circular, thus we see the tomato as spheroid. Compare this with our perception of a red stop sign in the road, which looks circular to us because we expect that as we move around the sign, its visual shape will become more and more elliptical, so we see it as a flat object, rather than as a sphere. Noë (2004) argues that the sensorimotor contingencies elicited by an object in our visual field thus constitute our perception of that object, resulting in the strongly enactivist position that action is right at the heart of perception.

We can now clearly see the tight relationship that FEF has with Noe-style enactivism. A system engaging in active inference is doing exactly the kind of hypothesis testing to confirm its percept that Noë’s theory suggests. Both theories require the system to generate expectations about the incoming sensory signal. Both theories require that those expectations are closely bound up with the system’s capability to act upon and interact with its environment. Both theories predict that perceivers will work to confirm these expectations as part of their active engagement with the environment. Not only do FEF and enactivism have superficially similar agendas in terms of putting action at the heart of cognition, but they also seem to suggest almost identical ways of going about this. It would not be a stretch to argue that PP grounded in FEF provides an attractive schema for a formal understanding of how Noë’s sensorimotor contingency theory might work in biological systems (Seth, 2014).

In this subsection I have highlighted two ways in which FEF and 4E cognition are closely related. Firstly by understanding cognitive systems as autopoietic, with a porous boundary between themselves and their environment, through which they interact to co-determine and co-evolve with that environment. Secondly by relating FEF’s commitment to active inference with the details of enactive proposals. In the following subsection I will be focusing on the explanatory role of representation, why this may be a sticking point between the two movements, and how we might make progress in this debate.

Differences

The similarities above help us understand how the work of Karl Friston and his colleagues might influence the research of cognitive scientists as it grows in credibility and popularity. Though it does seem to encourage a paradigm shift towards 4E, FEF is not the same as 4E. Its origins and focus are quite separate, as are its methods. FEF has developed out of information theory and computational neuroscience, and 4E has grown mostly at the intersection of psychology and philosophy. Indeed, it may be the differences between FEF and 4E that make FEF and its PP cousin so attractive to the dominant, cognitivist paradigm in contemporary cognitive science.

The rise of radical enactivism, flying the flag for 4E cognition, has kept the heated debate about mental representation going well into the 21st century. While this is generating a certain amount of attention and is inspiring a growing number of labs to research cognition in an alternative way, the representationalists still dominate. Perhaps one of the features of FEF that makes it so attractive to the establishment in cognitive science is that it seems to ground representation talk in hard information-theoretic science, and the fact that it accounts for the concerns of enactivists means there may be some hope of bringing the rebels into the fold if FEF takes off.

By basing an understanding of life and mind in a computational, statistical framework, FEF has inherited the inferential language of statistics. Even the process of synchronisation discussed above is understood by FEF advocates as inference – the system's internal states come to reflect the states of its environment, and in this perfectly concrete sense, the system thus contains knowledge about its environment. Inference and representation here hand-in-hand, and talk of priors, predictions, and precision seems to reinforce the cognitivist understanding of FEF. Given that this aspect of FEF is attractive to the cognitivist (that is, FEF literally 'speaks their language'), a positive feedback loop is instantiated, perhaps encouraging FEF advocates to continue using it.

This is anathema to enactivists. They eschew representation talk, preferring to explain cognition in terms of dynamic couplings with the environment, not requiring there to be any content in the head, as the head is perfectly capable of coupling with and exploiting regularities in its surroundings without constructing internal symbolic systems. Instead, the mind has been shaped by the environment, and in turn creates an environment friendly to its own continued survival. Rather than explaining how the mind works by explaining how it reconstructs the world through perception, creating representations, we can explain better how it works by finding out precisely how it leans on the world, letting the world itself do the jobs that representations are supposed to be doing. For the enactivist representations are an unnecessary problem for cognitive science, and we'd do better by doing without them.

Despite this core disagreement, I will argue that the inferential understanding of FEF can be reconciled with a radical enactivist position. To do this, I will first explicate how the notions of 'inference' and 'model' are used by FEF theorists, following their own reasoning. We will see that these terms refer only very loosely to a common-sense understanding of this kind of talk, which may be acceptable to the 4E advocate. I will attempt more sensitively to characterise the positions of 4E theorists, which I have thus far been treating as radicals, hell-bent on eliminating representations altogether. I will show that this is not quite right, and that many 4E theories make room for certain kinds of representation, thus bringing them closer to FEF. In this way, I hope to show that there is a way forward for a complete and harmonious union between the two paradigms.

The first clue that inference-talk in FEF is not particularly intuitive should be apparent from the discussion above. Inference in FEF can be understood in such a weak sense that a process of synchronisation between two systems counts as a kind of inference. For classical cognitive scientists, this kind of equation is very attractive – it is hard to conceive of a bundle of dumb neurons somehow being able to make complex inferences, but it is extremely simple to understand how two systems, however simple or complex, might become synchronised. Leveraging our strong theoretical and practical understanding of synchronisation to make sense of mental inference looks like an exciting and fruitful way forward. The sceptic will however want to stop us at this early stage – what on earth do inference and synchronisation have in common that permits us to compare them, let alone equate them?

The first thing to do is to understand the two systems in the synchronisation story as internal states and external states. Instead of system A and system B separated by a Markov blanket, we have a system's internal states separated from the environment by a Markov blanket. Through the process of synchronisation, the system's internal states fall into synchrony with the environment, that is, they come to resemble the environment in some abstract way that results in an equilibrium. This kind of formal resemblance can be described, in the technical language of statistics, as a model – so through development of better and better couplings, the internal system states gradually become a better and better model of the environment. This is a kind of statistical inference, approximating Bayesian inference. As time goes by, the way the environment acts upon the internal states of the system through the Markov blanket, that is, providing evidence about itself, allows the internal states to organise themselves into a better model of the environment. The sceptic's early question is answered – by translating the notion of formal resemblance into statistical model-talk, synchronisation becomes understood as a process of statistical inference.

At this point, the FEF advocate can anticipate further worries, and end the discussion. "This is what representation and inference is for us," they can say, "it *is* weaker than the classical understanding but that is because it is developed out of the concepts of pure mathematics, physics and statistics, and that's exactly what will help us to unify cognitive science with all these other disciplines." A radical enactivist hearing such a speech thus comes to understand exactly how weak these notions are in the FEF framework. They can see that the benefits arising from FEF's formal approach via statistics and information theory are precisely what makes it friendly to their own paradigm. The notion of representation that is built out of this is understandable in their own terms, it can be related directly to world and body involving processes and autopoiesis. It might be that they will avoid representation talk as a matter of personal taste, but allow that other cognitive scientists employ them, and be able to translate that research into their own preferred mode of understanding. Representationless and representational cognitive science can become a distinction without a difference.

The hope for this kind of reconciliation is amplified when we recognise that 4E theorists are not all absolutely opposed to representation talk. Goldman (2012) provides a useful taxonomy of the positions held by 4E theorists with regards to various metrics, including representation. Goldman identifies 'conservatives', who are what I have been calling defenders of classical cognitive science, who believe that the mind is full of mental representations and that they are involved in all cognitive activity. At the other end are 'radicals', who eschew all representation, but who are, it is worth noting, are a minority. In the middle are defenders of 'light' embodiment. These moderates defend the use of something along the lines of what Goldman terms a 'b-representation', viz. a body-representation. Alva Noë, as discussed above, falls into this category – his sensorimotor contingencies may be encoded as representations, but representations that are critically body-and-action-involving. These moderates are certainly true advocates of 4E, but still recognise the usefulness of representation, albeit in a drastically reduced and specific domain when compared with a conservative.

B-representations also link interestingly with FEF in two ways. In FEF, the organism itself is a representation of its environment, insofar as it instantiates a statistical model of a limited set of environmental conditions – the conditions in which the organism remains viable. That is, the phenotype of a given species has, through evolution, come to reflect its environment, sharing some abstract, formal relationship with that environment grounded in statistics. In this sense we might treat the organism as a b-representation, not in the intuitive way that a b-representation carries some content about a body, but insofar as the body itself is a vehicle for representation, a representation whose content is the environment. Granted, this doesn't sit well with classical cognitive science, but it offers an interesting new avenue for moderate 4E researchers looking for alternative ways in which a cognitive system might represent.

The other way that FEF links with b-representations is through the internal dynamics of anticipation and prediction. When an organism engages in an interactive exchange with the environment, in order to be successful it must take account of its own abilities and limitations. This is reflected in the similarities between FEF-grounded PP and Rick Grush's work on forward models (2004). Clark and Grush (1999) highlight the necessity for an expectation about future input, which mirrors PP's own structure. But they also make clear the implicit encoding of the system's own physical state and capacities in performing this computation. This is a b-representation. So whether or not the representation is explicit, it seems clear that a system that needs to anticipate its sensory input, as an FEF or PP system does, must contain some kind of B-representation. As a free-energy minimising system becomes a model of its environment through inference/synchronisation, the shape of that model, that is, the conditional expectations it encodes, will be tailored to the physical shape and capacities of that particular system. Though the system may not encode anything about its own body, it would be possible to infer the properties of the body from the model it has learned, given knowledge about the environmental niche.

The moderate 4E theorists that endorse Goodman's 'light' representation, have some common ground with the FEF advocate. However, from the discussion above it is clear that navigating this ground is not altogether straightforward. B-representations are comparable to elements in FEF, and if we look sideways at the system we can see that they are definitely at least implicitly encoded. There is enough here to provide hope for a reconciliation between the two, but it is also tenuous enough that we are justified in remaining sceptical. In the next section, 8.3, I will be pressing this sceptical line, pushing FEF away from a moderate reading of 4E cognition, and towards a more radical reading.

8.3 Statistical models and inferences do not involve representations

In the previous section I laid out the way that inference and models are understood by FEF. It was revealed that these notions are used in a highly technical sense and that this sense is much weaker than the sense in which representational language of this sort has traditionally been used in the cognitive sciences. In this section, I am going to extend Ramsey's job-description argument to FEF's concepts. Ramsey's challenge, as I have repeatedly applied it over the course of this thesis, compares the functional role of a representation to a common-sense understanding of the functional role of representation, boiled down to a 'standing-in-for' relation. In this section I am going to alter and strengthen this challenge. FEF is not necessarily a mechanistic theory that provides functional ascriptions, so rather than examine the function of statistical inference and models, I will be focused only on the linguistic senses of these terms, as they relate to the theory. I am going to strengthen the challenge by similarly analysing the use of representational language in cognitive science, in particular for psychology, and philosophy of mind. Having shown that there are at least two distinct uses of representational language at work, I will go on to argue that representational talk as employed by FEF is not used in the same way in other areas of cognitive science, and a result of this is that treating FEF's concepts as representational causes confusion in these other sciences. We therefore require an alternative to this understanding of statistical models and inferences, or at least widespread acknowledgement of this important distinction in order to prevent further confusion. I will close off this section by advertising the emerging, healthy body of literature that offers alternative ways of understanding cognition without appeal to representational language. I will show how these proposals map on to FEF-grounded PP, and also highlight the substantive content of making such a shift, that is, the important ways that this goes beyond mere terminological choice and should have a tangible impact on research as we move forward.

Two senses of representational talk

Talk about statistical inference and statistical models in dynamic, self-organising systems is highly specialised. In 8.2, I provided a defence of understanding these concepts representationally, from the perspective of a conservative cognitive scientist. This defence characterised the notions of inference and model in FEF as an attractive, minimal way of understanding representation and representational processes, neatly grounded in mathematics, physics, and biology. This remains the case, but it does not change the foundation of that understanding of representation in FEF, which is an understanding based on high mutual information, and the process through which two systems come to have high mutual information. High mutual information is typically the result of synchronisation – as two systems synchronise it becomes possible to observe the state of just one of the systems, and reliably infer the state of the other system. If I observe steam in the air above a coffee cup, I can reliably infer that the coffee cup will be hot to the touch. If two systems have high mutual information, there is a strong correlation between their states.

All this is grounded in Shannon information theory, and it is fairly uncontroversial that Shannon information is not semantic information – that is it is not content bearing (pace Isaac, forthcoming). So these statistical notions employed by FEF are not in the business of representing content, talk of models is a linguistic convention used to refer to formal structures that bear high mutual information with their target. This may be amenable to interpretation as a structural representation, of the sort discussed in part one, but this is not something that advocates of FEF have endorsed. Instead they tend to stress, as I have demonstrated in this section, the ways in which this process of inference and model building is related to dynamic, physical processes not amenable to representational interpretation.

On the other hand, representation talk in other areas of cognitive science often has quite a different meaning. Let's consider some research in psychology and in philosophy. Philosophers of mind invoke representational states in a folk-psychological way. That is, the mental states they tend to appeal to in their discussions are generally beliefs, desires, and thoughts. These are often couched as propositional representations, mental entities that are capable of carrying propositional content, usually understood as having a truth value, or correctness conditions. For instance, in his recent paper on thought insertion, Matthew Parrott considers the implications of his theory on the kinds of belief we might expect subjects to adopt (Parrott, 2017); in philosophical work on social cognition, a major debate between 'theory' theorists and 'simulation' theorists centres on whether or not humans have a set of beliefs about what others think and believe. Again, these are, *prima facie*, propositional representations about other people's propositional attitudes. We might weaken this kind of representation talk to model talk, claiming instead that we have a mental model of how others think that doesn't necessarily submit to propositional analysis. Nevertheless, the model certainly carries some content – it is about how other people think and behave.

This particular example also avoids a potential objection to other examples I might raise to demonstrate philosophy's use of content-bearing mental-states. Someone might object that philosophy of mind is concerned with conscious, personal-level states and processes, whereas FEF and neuroscience more generally is interested in sub-personal processing. But, the mental processes underpinning our theory of mind under discussion are sub-personal, they aren't involved in agential reasoning, as the faculty being explained – our ability to think and reason about other people's thoughts, is a personal-level phenomenon. So there can be no question that philosophers are often engaged in theorising about what is going on under the surface of conscious thought, in just the same area as FEF theorists and neuroscientists, albeit with a rather different set of methods.

Psychological research is similar in its use of representational states. Lancaster and Homa (2017) contribute to the literature on mental categorisation tasks. They talk about processes such as 'feature inference' and 'category inference', specifying different forms of inference by identifying different kinds of content being reasoned over. They claim that feature-to-feature inference is improved once a subject has learned the 'internal structure of a category' about which they will be thinking. Here extremely high-level, abstract contents are being ascribed to thoughts, learned knowledge, and inference. Barzykowski and Staugaard (2016) consider the effect of intentional effort on memory recall. Memories are entities that have often been talked about as being 'stored', 'retrieved', and 'corrupted', by psychologists. Contemporary science is no different, and in this study subjects were tested by being asked to form intentions to recall a certain memory. Not only do psychologists use content-laden representations in their explanations, but their entire experimental paradigm assumes their existence.

These examples help us get a grip on the way mental representations are used by researchers outside neuroscience. It is clear that these differ radically from the representations being invoked by FEF. The primary demarcating factor is that the statistical models of FEF are systems whose states have high mutual information with states of the environment, and are not necessarily content-bearing, whereas philosophy and psychology are interested in intentional, content-bearing states.

Of course, there is nothing incorrect in the way advocates of FEF and researchers in other areas use their technical terminology. However, there is potential for confusion here, and I believe we see this confusion propagated in work seeking to apply the FEF framework to an understanding of the mind through PP. In chapter seven, I quoted Jakob Hohwy's lay summary of an explanation for binocular rivalry. In this explanation and others like it, we see talk of statistical inference being transposed into talk of psychological inference, and talk of statistical models being transposed into talk of represented knowledge. In light of my discussion above, this kind of move is liable to cause confusion by equivocating concepts which are not equivalent, and have importantly different meanings and implications.

I want to highlight a real danger here. The threat is that this kind of conflation of terms will not just expose lay readers to misunderstanding, but also psychologists, philosophers, and other cognitive scientists who read this kind of work for an introduction to the framework. This misunderstanding thus percolates through the research community, resulting in misguided research and faulty conclusions. This danger is amplified by the enthusiasm with which FEF advocates such as Karl Friston push their agenda. Friston sees FEF as a potentially unifying theory of life and mind, and encourages application of his work to all areas of cognitive science, evident from his own forays into the philosophical arena e.g. Hobson and Friston (2012). While I too am enthusiastic about the potential scope of FEF and PP, and welcome attempts to consider how it might improve our understanding of the mind in all its aspects, this project must be undertaken with extreme care. We must be careful not to distort the fundamentals of FEF, acknowledging its foundations in statistical and information theory, and accounting for them when investigating new applications of the framework.

It might be feared that I am overplaying this issue. It may be contended that the use of statistical inference and statistical models in FEF provide a formal framework we can use to ground the richer psychological and philosophical notions. Though precisely how this might be achieved has not been shown, might there be scientific gains to be made by pushing forward with this research programme whilst an answer is worked out? I dispute this kind of response on two grounds, the first is simply that no such problem has been identified, until now, so no answer is being worked out. The second, more serious, difficulty is that the notions at play are so fundamentally different that I do not think it is reasonable to believe that there is a way to translate our familiar content-bearing understanding of representation into an information-theoretic understanding of models, in any direct and satisfying way. As we will see in chapter nine, I do believe there is a promising way forward for representation-talk in FEF, but that way forward first requires us to fully appreciate that these notions cannot be directly translated. The crucial property of mental representation – content – is simply missing from the information-theoretic story. As a result of this, the two do not play the same explanatory roles in their respective domains, and any attempt to equate them is simple equivocation, as I have argued above.

Alternatives

It is all very well to raise these objections, but such destructive arguments are not particularly fruitful without subsequent provision of good alternatives. Here I will sketch some of the ways in which work done by 4E theorists can be used to understand FEF's application to PP in a way that respects the fundamentals of the theory, and the distinctions that I highlighted in the previous subsection. The feature that all these alternatives share is the recommendation that cognitive science move towards a covering-law model of explanation, and away from a mechanistic, decompositional methodology. This not only is compatible with FEF, but can also help ensure that cognitive science respects the role of the body and environment in mental activity. I will start with the work of Bruineberg et al. whose analysis of FEF provides some concrete advice on how to go about applying the framework to cognitive science. Then

I will consider the proposals of Anthony Chemero (2009) who has presented a vision of the science that fits very neatly with FEF and PP.

The alternative conceived by Bruineberg, Kiverstein, and Rietveld (2016) (hereafter BK&R) situates the cognitive system firmly within its ecological environment. They follow the Gibsonian paradigm by focusing on affordance-based cognition. Affordances are properties of the environment that afford the possibility of interaction by a system. For example, a smooth, flat stone affords skimming; a gnarly tree affords climbing; the still water of a lake affords swimming. Affordances are also dependent on the abilities and skills of a system. As science develops we can see the number of affordances available to humans grow: a vein of rock might afford mining; a plant affords eating, or smoking for medicinal purposes; a tidal river affords the development of an environmentally-friendly power plant. Technology further enhances our affordances – a hammer affords hitting; a chair affords sitting; a pistol affords shooting; and computers afford a boggling number of possible actions. Cognition is the process of recognising and exploiting these affordances. According to BK&R, what FEF and PP offer is an exciting new way to explain how this kind of cognition occurs, sensitive as it is to the organism's embodied, situated state.

> "We proposed that the generative model is best thought of as a dynamical system of (affordance related) states of action readiness that reflect the hierarchical and temporal organization of the changing environment. As the animal develops skills, the generative model becomes more and more sensitive to the relevant particularities of the situation, and opens the animal up to the relevant affordances available in the environment."
> (Bruineberg et al, 2016).

In this quote we see clearly that BK&R are advocates of this affordance-based understanding of cognition, but what is also interesting is their identification of the system as a dynamical system. Dynamical systems theory (DST) lies in direct opposition to the mechanistic paradigm in cognitive science. It eschews the systematic deconstruction and reduction of mental processes, and seeks instead to understand the laws that govern the dynamics of the whole system. Discovering and understanding these laws allows us to provide explanations that are fully sensitive to environmental states, as they can take those states as working elements in the laws – the mechanistic approach can ultimately only appeal to internal states through their process of greater and greater reduction. They can account for external states only as components in the mechanism, but this insulates areas such as neuroscience from the implications of 4E research. The covering law explanations offered by DST can identify and exploit law-like relationships and patterns between internal and external states.

These kinds of covering law explanations also fall neatly out of the work being done on FEF and PP.

> "The organism's internal structure and organization is then understood as multiple simultaneous and coupled affordance-related states of action-readiness that together shape (through top-down precision-modulation) the salience of solicitations in the environment. The self-organization of these states of action-readiness allows the animal to tend towards an optimal grip on the multiple relevant affordances in the situation."
> (Bruineberg et al, 2016).

The hierarchical organization posited by the technical literature on PP allows us to identify different levels of aggregation at which we can analyse the system behaviour. At the organism level, we can get a grip on the patterns that emerge from the interaction between these interacting neural layers. Understanding these patterns and the law-like relations they stand in with environmental states is exactly what the FEF mathematics provides for us. The equations developed by Friston and his colleagues define the dynamics of a cognitive system's relationship with its environment. They provide us with a set of variables, which, when understood properly, scaffold a powerful and potentially fruitful new nomological framework for explaining cognitive phenomena. Holding the lessons of 4E cognition close, we reach the position of BK&R, which treats the predictive mind as a complex organ built for action in an environment, sensitive to the affordances that appear through a dynamic process of learning and interaction that describes how the system gradually organises itself into an emergent configuration that skilfully navigates its surroundings and thus preserves itself.

But what of statistical inference, models, and mental representations?

> “Crucially, once we dispense with misleading/distracting talk of probabilistic inference, it is no longer necessary to understand past experience and learning as encoded in the form of representational knowledge structures. Instead we understand past learning and experience as manifest in the skilled animal’s anticipatory dynamics to act in ways that improve grip on the affordances on offer in the situation.”
> (Bruineberg et al, 2016).

Learned knowledge structures don’t play a role in this dynamical understanding grounded in nomological relationships. The weight vectors posited by PP theorists as encoding statistical models can be treated as what they really are for cognition – an important variable. When we see FEF applied to cognitive system in this way, the talk of content-bearing models and inference falls away, it is not useful. To properly understand the mind we have to look at the aggregate dynamics of the system, in relation to the dynamics of the environment, and attempt to identify patterns in this relationship. FEF and PP help us understand these dynamics, by providing a way of conceptualising the relationship, and the structure of cognitive systems that permits them to maintain a grip on their situation. Talk of representation of external states of affairs, or even of the action-oriented kind is just subsumed into talk of skilful systems capable of latching on to and exploiting various affordances in the environment, the skills of the system are expressed through the dynamics of the system and its anticipatory, precision modulated action.

BK&R thus introduce us to a model of cognition that is grounded in the Gibsonian tradition. They treat the system as one governed by its dynamics, and thus best understood in terms of its dynamics, which allow us to explain its behaviour as a complex self-organised system that minimises its free-energy. This characterisation suggests that a covering law approach to explanation of cognitive phenomena is to be preferred to a reductionist mechanistic paradigm. One who shares this view of cognition is Anthony Chemero, who has provided a detailed defence of this methodology (Chemero, 2009).

Recall Chemero’s two main arguments for the dynamical stance (2009) from chapter three, Chemero argues that “the representational story depends on the dynamical story about the control system, not vice versa. … So the representational description is dependent upon the dynamical one." (2009, p. 77). Following a discussion of van Gelder's Watt Governor, Chemero notes that we can only construct a representational explanation of the governor's workings once we have a full dynamical account of its behaviour. This point generalises to cognitive systems – it is only once we have an accurate understanding of the dynamics of a system's behaviour that we are able to begin the process of finding and identifying the contents of representations. So in this sense the dynamical stance is primary – we need a dynamical understanding first.

Second, "the representational description of the system does not add much to our understanding of the system. Once we have the full dynamical story, we can predict the behaviour of the robot in its environment completely… without making reference to the representational content of any states of its control system" (Chemero, 2009, p. 77). This is the point that Chemero takes to be most damaging to the project of representational explanations. They can make no further predictions about the behaviour of the system that is not already accounted for by the dynamical stance. Representational ascriptions are thus empirically and theoretically redundant. Chemero follows this up by citing a slew of literature to pre-empt the common objection that our current understanding of cognitive dynamics only accounts for simple behaviours. The long-term success of the dynamical stance depends completely on the continued generation of good dynamical models that provide non-representational accounts of cognitive systems. If it begins to fail to generate models, or if those models clearly involve representational elements, then the dynamical stance would have to give way to a representational stance. I hope that this section has provided evidence that though PP is hailed as a representational model, that there is at least a strong claim to be made on this model by dynamicists. Rather than view PP as a theoretical strike against the project of DST, that by grounding PP in FEF, there is a continuity between dynamical accounts and PP, and crucially one that doesn't involve representational posits (at least, not posits strong enough to count as representations as they are usually used in cognitive science).

Chemero (2009) also highlights the implications for mechanistic explanation. In fact, he notes that the rejection of mechanistic explanation in favour of covering law explanation is a double-edged sword. On the positive side, covering law models reject teleological ascription. Chemero points towards the shift from Aristotelian mechanics to Newtonian mechanics. While Aristotle relied on teleology to help move parts of the universe, Newton did away with such problematic talk of functions and ends by providing a set of universal laws that governed, or described, the way things move. One of Hume's goals was to see psychology succumb to a similar revolution, and find itself described precisely using mathematical models and laws. On the other hand, history also predicts a problem for covering law explanations – their failure to provide an adequate guide to discovery. One function of a scientific explanation is, in addition to explaining the phenomena, to give us some idea where to look, and how to find, new phenomena to be explained by the framework. The charge is that mechanisms provide this guide to discovery – they invoke new posits that are used to explain the phenomena, which we can then go and look for. Covering law models do not, they just provide a mathematical description.

According to Chemero, there are reasons for us to think that DST does provide a guide to discovery, albeit obliquely. Wedded as it is to Gibsonian Ecological Psychology, DST can lean on this theoretical framework to provide a guide to discovery that doesn't involve representation, or any kind of mechanical decomposition. Ecological psychology provides a radically different way of viewing cognition, and dynamicists can use that framework to generate dynamical models for new kinds of cognitive capacity. Another way that DST can provide guides to discovery is through the use of general models: "with a model that seems to apply to a wide variety of cognitive phenomena, one can generate predictions to be tested with experiments by hypothesising that as-yet-unexamined phenomena can also be described by the model." (Chemero, 2009, p. 86). Chemero points to the Haken-Kelso-Bunz (HKB) model (1985). HKB provides a very general mathematical model that accounts for certain features of self-organisation, and has been applied successfully to numerous cognitive phenomena including some that are typically taken to be 'representation heavy' (e.g. speech production, Port (2003); problem solving, Stephen, Dixon, and Isenhower (2007); social coupling, Kelso et al. (1990); and perception and action coupling, Kelso, DelColle, and Schöner, (1990)). It seems to me that FEF is a model that fulfils a similar niche to HKB, but that hasn't yet been fully explored. Thus by taking the dynamic stance towards FEF and PP, we are not stuck for a good guide to discovery because the models themselves are so general that they require refinement for modelling of particular cases, which makes them good guides to discovery. FEF can also share in the theoretical background provided by the Gibsonian tradition as discussed above through BK&R.

Using dynamical models as explanations for cognitive phenomena challenges the importance of mechanistic analysis. This is due to the particular kind of dynamics we find in cognition. Typically, "the individual equations and variables that describe the coordination components do not map on to neo-mechanistically defined lower level components of a system. Instead the variables and parameters are defined in terms of energy expenditure" (Lamb and Chemero, 2014). This is true also of the FEF, whose variables and parameters are defined in terms of free energy. These kinds of variables are not easily decomposed into mechanistic component and process elements. Further to this, Lamb and Chemero (2014) provide other revealing analysis that applies well to the case at hand.

> "When a system's behaviour(s) can be modelled at a variety of scales such that each model has the same collective variable, we can take this as evidence that across these scales the system's behaviour is governed by the same energetic constraints. This means that regardless of the small-scale physical stuff that makes up the system, an explanation of the system's behaviour must include an account of the energetic constraints or enslaving principles that the system is subject to. This would be a clear case of a thoroughly dynamical system." (Lamb and Chemero, 2014)

In this case Lamb and Chemero refer to collective variables that apply at a variety of scales. The FEF model is applicable at any scale where a Markov blanket can be drawn. Its variables can then be applied to the internal and external states to capture the dynamics of the system that is defined by that blanket. Thus Lamb and Chemero's point holds for FEF (and for the PP models derived from it) – any explanation of a system describable in these terms must explain the constraints recognised by those models. These kinds of system are 'thoroughly dynamical' in that any way of understanding the make up of these system that ignores these concerns – e.g.

mechanical models – fail to explain fundamental features of the behaviour of that system. Lamb and Chemero (2014) go ahead and spell this out for us: “any neo-mechanistic explanation of the system’s behaviour would be lossy, an over-simplification of the dynamic systems model already provided. This is because a mechanistic model necessarily ignores the effects of non-linear coupling both in defining the system’s components and defining the system’s behavioural modes”.

In this subsection, I have presented the specific proposal of BK&R for understanding FEF and PP as part of the Gibsonian tradition, and then I went on to show how they fit neatly into Tony Chemero’s general dynamical stance framework. Taking these proposals seriously requires a rejection of a mechanistic understanding of FEF and PP, and instead treating them as mathematical models of the dynamics of a cognitive system. These models can be used in covering law explanations, which avoid the problems of teleology which I discussed at length in chapter seven, but which do not suffer the usual difficulty of covering law strategies, because PP and FEF are sufficiently general to also provide a robust guide to discovery for the development of future models.

8.4 Conclusions

In this chapter, I problematized the application of cognitive science’s favourite explanatory strategy to PP and sketched a way forward. That way draws upon the ideas of the primary advocate of the theory, Karl Friston, and the arguments of dynamicists, BK&R and Chemero, and involves situating FEF in the Gibsonian tradition and applying the related methods of DST to its ongoing development. The catch is that these methods explicitly reject representational commitments. To avoid a non-representational science then, the advocate of PP must either reject the close relationship between PP and FEF which I developed with the help of Hohwy in 8.1, or they must show that PP grounded in FEF is not properly understood in dynamical terms. In the next (and final) chapter, I will consider a third possibility – that representations can emerge from the dynamics of a cognitive system. This admits PP as a dynamical theory, but also maintains the possibility of representation without requiring mechanistic decomposition. This way forward strikes me as the most promising for all parties – preserving the attractive elements of both the classical, connectionist, and 4E paradigms.

Chapter 9: From Systems Ontology to Proper Perspectivalism

So far in Part Three I have cast important doubts on the explanatory role of representation for both a mechanical understanding of PP, and also for any understanding of PP based on the dynamical foundations of FEF. It is all very well to nay-say, but there must be some representational facets to human cognition, otherwise how are we to explain our ability for language, which is a clear instantiation of a symbolic, rule-based system or certain phenomenal qualities of consciousness, which seem to be clearly represented to us in experience, or our capacity for dreaming, in which scenes, feelings and thoughts appear present to us when we are unconscious, or indeed for our ability to reason about the abstract and the absent, which appear in our thoughts even when they are not appearing in our perceptual consciousness?

Our capacity for language is simple proof that our cognitive systems are capable of constructing and responding to strings of meaningful symbols. Given the complexity and scope of human language, which goes far beyond the languages present in other species, the obvious explanation is that at least some parts of our cognitive system must deal in representations of sentences in order to facilitate the formation of new sentences that conform properly to linguistic rules and conventions. In a different vein our conscious experience appears to us to be constructed out of representations – objects of consciousness appear separate from each other forming discrete elements in our sensory manifold. Whether or not this is a correct characterisation of conscious experience is a question for another book, but these kinds of observations give us *prima facie* reasons for supposing that the systems responsible for conscious elements of mind possesses some representational properties.

The work until now has been to deconstruct the idea of internal representation as far as possible, in general and within PP to show first, that its role is overused, overstated, or simply misunderstood in much philosophical and scientific theorising; and second so that, in light of this new understanding, we might be able to offer a better theory capable of accounting for the undeniably representational facets of our cognitive and mental lives.

In order to meet this remaining challenge, I will now lay out the foundations of an account capable of explaining the representational elements of cognition. I will begin by presenting, defending, and developing an ontology of processes, put forward by Mark Bickhard (2000) and developed further by Richard Campbell (2009) of various kinds of self-organising system and their basis in simpler processes (9.1). Following Bickhard's thought, I will offer an account of function based on the properties of a certain class of systems – self-maintaining systems (9.2). This account leaves open the possibility of representation in the dynamic understanding of PP that I argued for in chapter nine. With this possibility in place, I will lay out consider possible responses and implications, and argue that this points the way to a harmonious union of mechanistic and dynamical approaches in cognitive science (9.3 and 9.4). We will be left with an attractive vision of the place of mental representation within cognitive science, that at once enables us to explain representational features of our mental lives, and helps us to carve out a scientifically respectable space of cognitive and mental phenomena (9.5).

9.1 An Ontological Framework

Throughout this thesis I have been using the term 'system' without any formal definition. I hope that thus far this has been a fairly innocent neglect and that my usage has been broadly aligned with the reader's intuition about what a system is. The coming discussion however, demands the demarcation of different kinds of system, so it is worth taking a moment to think about exactly what we mean by the category term before we go carving it up into more useful pieces. Broadly, a system is a collection of elements, each with its own properties and causal powers, that interact with each other in particular ways. Cast this way, almost any arbitrary collection of things can be called a system, and that's exactly the kind of breadth of application necessary to appreciate the continuity between the sciences that is emphasised by Bickhard and Campbell's process ontology. Systems can be designed, evolved, brought together by happenstance, nomological necessity or some combination of all the above. A bicycle is a designed system. Like the vast majority of designed systems it is mechanical. Its component parts are heterogeneous, and have been put together so that in the right situation, the parts interact in such a way that a rider can propel themselves forward with great efficiency. An eye is an evolved system. It is also mechanical, with heterogeneous parts organised in such a way that, if in good working order and in the right situation, can be used to detect light and colour. A hurricane is somewhat accidental and brought together by the laws of nature. It is made of a large number of mostly homogeneous parts – water and air molecules – it is not mechanical, but governed by dynamics. In the right situation, the hurricane will persist indefinitely, bringing heavy rain and high winds to its local area.

Many natural systems – systems that have not been designed – are well understood as processes rather than entities in themselves. The hurricane for example, cannot be identified with the molecules that make it up at any given time, as they are constantly being swapped and replaced through rainfall and evaporation. It is better described and understood through the interactions between the molecules, and the component processes that facilitate the overall integrity of the system (rainfall, evaporation, warming, and cooling). This fact is recognised by biologists, and chemists, but has not been widely acknowledged by physicists, who prefer dealing with very stable, unchanging, particles. Biological systems go through numerous and often radical changes over the course of a lifetime – consider the familiar example of humans, and the stages they pass through: a baby, an adolescent, an adult, and an elderly person. There are many significant physical (and, if you like, mental) differences between every stage, but we recognise them all as part of the same, human process. Even more radical is the example of metamorphic insects – while human parts undergo large changes, the parts themselves (arms, legs, internal organs) are easily identifiable through the whole process, but in metamorphic insects the parts themselves change entirely. A caterpillar has eight legs many body segments, and no wings. They feed on fibrous leaves. A butterfly has six legs, large wings, and feed on sweet nectar. Nevertheless, the caterpillar and butterfly are readily identifiable as two stages in the life-cycle of a single kind of organism. Biological organisms are best individuated as processes, with the concept of change built in, rather than as entities.

Bickhard (2000) and Campbell (2009) follow the school of thought advocated by Alfred North Whitehead, and argue that this kind of process thought can provide a complete metaphysics, applicable to all sciences, including physics. They believe that many physicists make a mistake in denying that the natural world is made up of processes, rather than entities. I do not want to take a firm position on this issue – there is a great deal of work to be done in this area to make the claims both concrete, and convincing. However, Bickhard and Campbell do well to note that quantum field theory indicates that quantum particles are best understood as fields, rather than as corpuscles. Particles are thus processes that behave as we expect particles to behave, but in reality the universe may be 'processes all the way down'.

Taking process as a starting point, we can sketch an ontology of systems. It will become clear that early on in this sketch, the assumption that processes are metaphysically basic can be safely bracketed. What process gives us, in this instance, is an elegant and useful way of carving up the space of complex systems, with an emphasis on the kind of interaction-dominant dynamical systems that free-energy minimising systems fall into.

Our first step is to separate *persistent* processes from fleeting processes. Sadly, this already thrusts us into some murky water that Bickhard and Campbell fail to fully acknowledge. Campbell (2009) notes that 'a process is persistent relative to changes in its environment, or it is not'. This relativity is a little problematic, making it difficult to draw a nice, sharp line between persistent and fleeting processes. It invites the question *which* changes in the environment ought a persistent process persist through? Campbell attempts to clarify the issue: 'persistence is a relative quality; it turns on the organised process lasting for a longer time-span than the other processes in its surroundings'. I find this attempt unsatisfying, and it muddies the issue with additional questions. First 'which other processes are appropriate comparisons?', and second, 'how close in space and time must two processes be to count as being in one another's surroundings?'[16]. For now. let us continue as if nothing was the matter, these questions are largely insignificant for the present treatment.

Once we have identified a process as persistent, we can ask whether that process is *cohesive* or not. Campbell nicely defines a cohesive process as one 'whose internal processes work together to ensure… stability' (2009). The example he uses is that of chemical bonds – thus atoms and molecules are cohesive processes, as are larger objects such as rocks, and stars. Thus we enter the realm of entities. An entity might be defined as a cohesive process, and so from this point onwards we can, for the most part, bracket the question of process metaphysics' validity.

However, again there is relativity here. Stability is defined 'with respect to certain forces', thus 'any system coheres only within a limited range of conditions'. This is

[16] In these cases, it's often useful to consider examples, to help us get a feel for the idea, even if we don't have a neat distinction. Let us follow Campbell and take his example of a fleeting process: a falling apple. The organisation of the apple and the ground that give rise to the process of falling, captured by the temporally extended downward acceleration of the apple, does not last long, relative to other processes in its surroundings. So let us now consider again our questions.

What processes are appropriate comparisons? Putting aside the question of surroundings for now we can find a myriad of processes occurring in the vicinity of the falling apple, many of which are much more fleeting than the falling apple itself. There are thousands of chemical processes happening on a cellular level within the apple, there are millions of gas molecules in the air being pushed aside by the motion of the apple, and the motion of each can be separated as a single, fleeting, process. Relative to these processes, the falling apple process persists for many lifetimes. On the other hand, there are also many more persistent processes unfolding in the immediate environment – the apple itself is a biological process lasting hundreds of days, and the molecules of air are processes that may last millions of years.

How immediate are 'surroundings'? The appropriate surroundings to consider in the apple case may be very local – not extending far beyond perhaps the branch from which the apple fell, the surface that it lands on, and the intervening medium of travel. However, for the same apple, launched into orbit, thus falling in a similar kind of process, the environment is very different – the starting point is not particularly significant to the process, nor is any surface as, if left undisturbed, the apple will never land. The earth itself, whose centre of gravity is creating the gravity well constraining the apple's fall might be a more appropriate member of the process environment.

not an insurmountable problem, it is conceivable that we might precisely specify the conditions in which a process will remain cohesive. For instance, we expect our brick wall to remain a wall when we kick a football against it, or even if it is shot by a low-calibre weapon. However, we do not expect our wall to remain a wall for every long if struck by a wrecking ball, or during a strong earthquake.

Relativity of the sort Campbell requires to make sense of persistence and cohesion may not bother some readers. As in the earlier discussion of function, one might be happy to explicitly fix one's perspective on the domain, and work back from there. There are three reasons I will not adopt that stance. The first is that the motivation for this chapter is to find a working notion of proper function, which, if successful, will allow us to address the question of function without adopting a fully perspectival stance. The second is that perspectivalism does not resolve all the issues in this case. As my examples above illustrate, it will be little more than an ad hoc stipulation to describe one process as persistent and another as fleeting. In the case of functions, one can identify a capacity of interest, the target of the explanation, and identify how parts function to bring about that capacity. In the case of persistence, it is the persistence itself that would have to be stipulated at the beginning of the explanatory project – a system of interest will involve the interaction of many processes and assigning the label of 'fleeting', 'persistent' or 'cohesive' to distinct subsets of those processes will be part of the perspective-fixing process rather than a consequence of it. Furthermore, it is not clear to me that much explanatory leverage is gained with these labels, making the perspective taking both arbitrary and redundant. Thirdly, the real import of this ontology can be appreciated without the categories of persistence or cohesion. The real work is being done elsewhere by the categories I am about to sketch. In the analysis of these below I will explain how they help us make sense of the entity-process distinction, which Campbell suggested the 'cohesion' category dealt with.

Taking as a starting point the laws of thermodynamics, we can identify processes that are at equilibrium, and those that are not at equilibrium. This I take to be the relevant distinction that captures a properly non-relative notion of persistence, for a process at equilibrium will continue forever until perturbed in a certain way. We can define a process at thermodynamic equilibrium in the following way:

0. A process/system is at thermodynamic equilibrium if and only if it can continue without the input of energy.

This is what Campbell calls 'energy well stability' (2009). Some equilibrium processes may be more stable than others in that it they resist perturbation in a wider range of conditions than others, but this is not an admission of relativity, it is merely a recognition that there is a continuum of stability within the category of equilibrium processes. The moon is a system at thermodynamic equilibrium, if no energy entered the moon's system then it would remain as it is for eternity. In order for it to change, energy has to be brought to bear on the system, for example, in the form of an asteroid impact. A small impact will change the moon system's shape by leaving a crater. A larger impact might change the moon system by leaving a much larger crater and altering its orbit, perhaps sending the process on a trajectory that causes its demise as it falls into the earth. Processes exhibiting energy well stability are thus well-understood as entities but do not exhaust the class of entities.

The opposite of an equilibrium process is a non-equilibrium process, we don't need a separate definition of these processes as it is just the negation of (0). Namely:

1. Non-equilibrium processes are those that will not continue without the input of energy.

Of course the vast majority of non-equilibrium processes last only a very short time, and given enough time all non-equilibrium processes will eventually end due to the heat-death of the universe. This is what I take to be a more concrete way of roughly capturing what Campbell referred to as 'fleeting' processes, these are processes that will not continue past the heat death of the universe, and in most cases begin and end within fractions of a second (e.g. an electric impulse). Another interesting property of far from equilibrium process is that they can *change* without any input of energy. In accordance with the second law of thermodynamics they will change over time, of their own accord, until they come into thermodynamic equilibrium, possessing maximum entropy.

Some, but not all non-equilibrium processes fit neatly into the ‘entity’ category. Stars are clearly a kind of entity, but are non-equilibrium processes, all of them are doomed to die when they run out of fuel, the sub-process of a star, nuclear fusion is similarly a non-equilibrium process, but we don’t think of it as an entity in any way. As we flesh out this ontology we will explore the space of processes and systems that fit neatly into the entity category, finding that as complexity increases, a process fits more neatly into the entity category.

For instance, some systems are *stable* despite being far from thermodynamic equilibrium. These systems are those non-equilibrium processes lucky enough to receive the energy then require in order to continue. They are not qualitatively different from other far-from equilibrium processes, but we can ask question regarding *why* the system is far from equilibrium. For example, consider the water cycle. The water cycle on earth refers to the way that water molecules are transported from oceans, to landmass, and back to the ocean. From the top of a mountain, a drop of water will flow down a river, eventually reaching the sea, and then the ocean, due to the earth’s gravity. Without any energy input, it would stay there, and eventually all water would end up either in the ocean, or frozen or trapped on land. However, the sun provides constant energy input, and water in the oceans evaporates, forming clouds, which then move on air currents (also powered by the sun) over land, eventually gaining enough density that the water molecules condense into droplets and rain down again on to the mountains. The water cycle continues because the sun continues, it is entirely contingent upon the sun.

Some far from equilibrium systems reduce their contingency on the ongoing presence of their energy source. These systems are *self-maintaining*. The water cycle is not self-maintaining, the process itself does nothing to receive the energy required for it to persist. However, a system might ‘make active contributions to its own persistence’

(Campbell, 2009) and this system would be self-maintaining. Self-maintaining systems (SMSs) are still, of course, dependent on friendly external conditions, the leap here is to systems that actively reduce their dependence on favourable external conditions. This leap is neatly captured by Karl Friston's Free-Energy Principle (Friston and Stephan, 2007): "The free-energy principle states that all the quantities that can change; i.e. that are part of the system, will change to minimise free-energy". We know from the discussion in chapter nine that minimizing free energy is equivalent to minimizing the statistical surprise experienced by the system, and that understanding what it means for a system to experience surprise requires us to recognize the sense in which a system embodies a statistical model of its world.

So, systems that conform to the free-energy principle are SMSs. When free energy is growing, the system is moving towards energy-well equilibrium. In accordance with the second law of thermodynamics, the free-energy in far from equilibrium systems will tend to increase. However, systems that conform to the free-energy principle change in a way that minimizes their free-energy. This kind of changing amounts to Campbell's 'active contributions to its own persistence'.

2. A system or process is self-maintaining if and only if all the quantities that are part of the system will change to minimise free-energy.

Incorporating the free-energy perspective requires us to make an important conceptual addition to Bickhard and Campbell's ontology – the Markov blanket. As discussed in 3.2, the Markov blanket allows us to make a principled distinction between two systems, in virtue of treating one as a statistical model of the other. Applying the Markov blanket to process ontology may raise some complications, as the boundaries of processes are often fluid and changing (Clark, 2017). That said, nothing in the theoretical understanding of Markov blankets prohibits variation in the boundary, so this development presents only the empirical challenge of how to keep track of a system's Markov blanket as it changes over time.

To illustrate the core concepts of the SMS, let us use a favourite example of both Bickhard and Campbell: the candle-flame. In virtue of its heat, the candle-flame transforms solid wax into usable, vaporized fuel, which then participates in the ongoing process of combustion that contributes to the persistence of the candle-flame. These processes facilitate the system level changes that facilitate minimisation of free-energy within the Markov boundary. What the flame process *needs* in order to continue is vaporized hydrocarbons, and in virtue of its heat it is capable of transforming other forms of matter into what it needs – the flame starts with a very small amount of fuel (e.g. from a matchstick), but is quickly able to create its own supply of several millilitres of liquid wax. Further, in large pillar candles, the flame melts a deep cavity for itself, which protects it from gusts of air which might otherwise extinguish it, which makes the flame of a large candle a more robust SMS. The large pillar is especially instructive – it both reduces its dependency on the presence of vaporized fuel, and *protects* itself from external influences that would otherwise end the process.

At this point it is helpful to consider the entity question once more. The candle flame is a difficult edge case, which can be argued to be an entity or a process depending on one's assumptions. The fact that there are edge cases like this are helpful to hard-line process metaphysicians, who might argue that there is no proper distinction to be made between entity and process. I will not wade further into this debate here, but suffice to note that the self-maintaining stability of the candle-flame seems to make it more entity-like than a non-SMS such as the water cycle[17]. We are left with two firm categories of entity – energy-well processes, and SMSs – and one less firm category – far-from-equilibrium entities that are not SMSs, such as stars.

[17] Where does something like a star fit into our ontology? It is certainly a non-equilibrium process, but is it self-maintaining? I would tentatively argue that it is our human perspective of stars as extremely long-lived processes that encourages us to refer to them as entities, despite their fate to undoubtedly die and return to thermodynamic equilibrium. This just highlights the way that entity-talk breaks down once one takes a more process-oriented perspective, it is up to the reader to decide which is more accurate.

It is in the category of SMSs that we find the first emergence of life in our ontology. Simple life-forms persist in the same way that a candle-flame persists. They are typically more stable than the candle, as they are born into an ecological niche that suits them, courtesy of evolutionary processes. But their interaction with their environment is of the same fundamental kind as the candle flame (imagine a very flame friendly world - perhaps a deep vat of wax with a long wick running through it). Furthermore, the vast majority of life is more complex than a simple SMS, falling into the categories I will outline below.

SMSs and Function

I now argue, with Bickhard, that SMSs are also well-understood as exhibiting goal-directed, functional, behaviour. This may appear to contradict some of my claims in chapter seven, but I will address this apparent difficulty in section 9.3. For now I will lay out the case for this feature of SMSs. I will then go on to consider the more complex way that *recursively self-maintaining systems* (RSMSs) exhibit functional behaviour, bringing with it agential, autonomous behaviours, which are capable of being in error. Once we have secured this foundation, I will detail the elements that bring fully-fledged, world-representing mentality.

The keystone in Bickhard's case is that SMSs instantiate a system that have a clear, natural function – the function to maintain themselves. There is a sense in which this function defines them as SMSs. But Bickhard's use of the word function is somewhat ambiguous. In one sense of the word, the claim is trivially true – SMSs have this function because that is just what they do, so they instantiate this function, or capacity. In the normative sense of the word, the claim is not obvious at all – it is not clear that a SMS can malfunction, because as soon as it does, it is no longer an SMS, that is, it is no longer working to maintain itself. Perhaps a blurring of the term helps us get at Bickhard's intuition here – SMSs tend to maintain themselves, and given that self-maintenance involves continuing into the future, we can count it, even in the future, as an SMS. So when the SMS dies, we can legitimately call this a case of malfunction, because part of what it was to be an SMS before, was to *expect* to continue to be into the future.

Let's unpack this important move a little further. The key claim here is that part of what it is to be a SMS is to strive to continue into the future. Epistemologically, we might say that once we have identified a system as a SMS, then we are justified in having greater confidence that that system will persist into the future, than if it were not a SMS. However there is a stronger element to the claim, an ontological element. If a system is a SMS, then its probability of persisting into the future really is higher than a non-SMS. This follows directly from our definition of a SMS in terms of free-energy minimisation. It is this fact that provides the justification for the epistemological side of the claim. Usefully, this important aspect of SMSs is explicitly captured by FEF. FEF defines these kinds of system by their ability to manage the probability that they will be in friendly circumstances that allow them to persist, by staying in those kinds of situations, and avoiding situations that are unfriendly and threaten their viability (recall the walls built by the flame in a pillar candle). In a way, SMSs embody induction – their existence now justifies belief in their existence in the future, because if they exist now, then they will probably exist in the future, for the, by definition, take action to remain in friendly circumstances.

From this line of reasoning, we certainly do not immediately get the claim that SMSs then *ought* to continue into the future, because that is their function. However, the SMS does embody the *expectation* that it will continue to exist, in a very minimal statistical sense, as per the FEF. Given that this 'expectation' is about itself, the system embodies a proposition that will be normatively evaluable – "I'm here, so I probably will continue to exist". Being right or wrong in this case comes down to whether the system is a SMS or not, that is, whether or not the system has a self-imposed function in virtue of the statistical expectation it embodies.

An important phenomenon to account for here is the expected life-span of an organism of a given species. Death, it seems, is one thing that is certain for human beings, and most other animals. Indeed, there are good evolutionary reasons for limiting the life-span of individuals in a species so that the population within the ecological niche remains stable. One evolutionary reason for example, is to prevent a predator species' over-consumption of its prey species, which would be catastrophic for the predators as a whole, resulting in extinction of both species. Thus, from the evolutionary perspective it makes sense that a predator species ought not live forever, while continuing to reproduce, lest they all die of starvation. The problem this poses for us is that some SMSs seem to *expect to end*, that is, they embody the expectation that at some point they will cease their self-maintenance. Some paradox this is for us.

Perhaps we can say that the SMS does not expect to end *early*. This gets us on to the right path, but suffers difficulties. Once the SMS has exceeded the mean lifetime of others of its kind, do we say that it is malfunctioning from then onwards? This seems like an absurd answer, after all, a longer lived system must surely be a better SMS than one that dies earlier, no matter what the expectation.

An answer, I believe can be found in the *indeterminacy* of expectation about life spans, and by acknowledging how this expectation changes as the SMS process unfolds in time. At any given point in time, the system will embody a probabilistic model that expects to continue for a certain length of time, and this will change *depending on the current state of the system*. At birth, given the current social, physical, and technological environment, a woman in the UK, *ceteris paribus*, expects to live for 82.82 years (ONS Life Tables, United Kingdom, 2016). The individual's circumstances play a large role in what the precise embodied expectation will be. If the newborn child is afflicted by cystic fibrosis, then the life expectancy will, sadly, be radically shorter. However, the crucial point is that the expectation never bottoms-out at zero. That is, at any given point in its process, an SMS does not expect to end. Its expectation bout how long it will continue will get shorter and shorter – for instance, the ONS reports that once a woman in the UK reaches the age of 85, their life-expectancy is 6.7 years, they can expect to live well into their 92nd year of life. Once 92 years of age, the same woman can expect to live a further 3.94 years, and then, once 96, the woman would embody the expectation that she will live another 2.93 years.

The crucial thing to note is that the expectation never falls to zero – the system remains a SMS as long as it continues to self-maintain.

This expectation is co-determined by the system's phenotype and by the environmental circumstance. An interesting note from the historical record is that in the Roman world, life expectancy at birth was 21 years. But once a child reached the age of 5, life expectancy jumped to 42 years (Parkin, 1992). I take this to be a feature of the view, not a bug – the system is not just embodied, but embedded in the environment, in line with the lessons we have learned from 4E theorists. The important fact is that at any given moment, a SMS expects to continue. The precision and mean value of that expectation will fluctuate over time as a function of the system's internal states, given the external situation.

It is important to be careful here. I was careful in chapter nine to make clear how minimal the FEF notion of expectation is. What we achieve here is a very minimal notion of function from the framework. The SMS instantiates a statistical model of the world in which it is capable of persisting, and by making active contributions to remain within the bounds of that model (by minimising surprise, or free energy) the SMS is enacting, in a direct, real sense, the expectation of continuing. The minimisation of free energy provides the system a directedness – it directs itself towards those states in which it is most likely to persist. It is this kind of directedness that we are calling 'expectation' which results in the real functional quality of the SMS.

Dynamic Presuppositions

Further, Bickhard (2000) argues that once function has emerged in a SMS, its sub-processes also possess function in virtue of, what he calls, 'dynamic presuppositions'. A dynamic presupposition is a process that is necessary for a system *like that* to continue to self-maintain[18]. I also follow Bickhard to this conclusion. Take the

[18] An interesting case to raise again here are the metamorphic insects. When I write 'a system *like that*', it is important to remember that we are taking a process perspective. The system, properly understood, is a process. So, the metamorphic insect system, and

example of the candle flame. The flame is composed of a number of processes, most salient are the vaporization process which generates fuel from solid wax, and the convection process that carries away ash, which would otherwise build up around and eventually smother the flame. Bickhard's compelling argument is that these processes are functional for the flame as the contribute to its self-maintenance. Further, the FEF provides us is a general way of understanding these processes – they must minimise the system's free-energy. Because the SMS has a function, the dynamic presuppositions have functions too – so we can say that according to the FEF, it is the function of a SMS's sub-processes to help minimise free-energy. How they do this will depend on the precise nature of the system.

Thus the property of 'having the function of persisting', or 'having the function of minimising free-energy' bestows function on lower-level processes. So whilst function requires that a certain kind of system develops, once that kind of system does appear, the parts making up that system gain a new property – the function to contribute to the maintenance of the larger system by minimising free-energy. This may sound magical, but it captures what is special and fascinating about SMSs – their apparent purposefulness. Recognising that this purpose is well-grounded in the fundamental physics of self-organisation laid out in the FEF, it becomes easier to see the non-magical truth. The SMS imposes an organisational structure on the dynamics of its micro-level substrate, and that part of what is imposed is function. The micro-level is being enslaved by the macro-level, to help preserve the macro-level's integrity.

The notion of dynamic presupposition may raise some problems. A system is made up of a number of processes. In order to differentiate processes that are dynamically presupposed from those that are not dynamically presupposed, we will have to make some judgement about what sub-set of processes are essential for a system to count as a certain kind of system. For example, consider the kind we call 'flame', flames come in different varieties. It is not necessary for a flame to appear yellow, for it may be

the nature of its self-maintenance, is a process that involves radical change, but it is a singular, identifiable, self-maintaining process, and that is what is important for our treatment.

using fuel that does not result in carbon by-products. The process that results in yellow appearances (the superheating of pieces of ash) is not a dynamic presupposition of the 'flame process'. However, is it a dynamic presupposition of the *candle* flame process? Perhaps, because candles use fuel that produces carbon by products. Is this an accidental property of the candle flame, or a necessary property, i.e. a dynamic presupposition? This is a knotty problem indeed. For now, let us note only that to make sense of dynamic presuppositions at all, we require some account of *kinds* that will allow us to determine which sub-processes are dynamically presupposed, and which are not. As I continue to make use of dynamic presuppositions, I will be assuming that the notion of a kind can be worked out in a naturalistically respectable way. If this turns out not to be possible, we are left with just a proper understanding of system function, and not a system function that makes proper sense of the function of that system's parts.

Outsourcing Function

By identifying a Markov blanket in nature, we can demarcate a potential SMS[19]. The blanket helps us define the entity at the macro-level. In 3.2 I explained the porousness of a Markov blanket, and how it is a boundary that enables interaction and exchange rather than limiting it, and providing a channel for autopoietic processes, thus proving the FEF to be extremely amenable to even radical enactivist advocates. Through the blanket a SMS can outsource its functions to systems in its environment. That is, a SMS's actions, being part of its processes, have the function to contribute to its self-maintenance. If, through its actions, the SMS creates, maintains, or uses entities or systems in the environment to contribute to its self-maintenance, those systems also inherit the function of the SMS.

[19] Note that we must be careful here, for, being processes, some systems radically alter their Markov blanket over time, e.g. metamorphic insects (cf. Clark, 2017). The presents a challenge for those studying such processes empirically, but can be safely bracketed for the purposes of our, theoretical, discussion.

In summary, function emerges wherever SMSs emerge. This could be at many different levels of abstraction: chemical, biological, sociological, to name just a few. This is because function is a property of SMSs that emerges in virtue of their implicit expectation of their continued survival. This idea is captured in the statistical formalism used by the FEF. The FEF defines SMSs (i.e. free-energy minimising systems) as statistical models of their own survival in their ecological niche. Under this framework the way that a system gets to be this way is by being a free-energy minimiser. Free-energy minimisation is in this sense the function of SMSs. We can also presuppose that the micro-level processes of SMSs thus function to contribute to the minimisation of free-energy, which in statistical language means that they function to increase the system's confidence in its expectation of its continued survival. This notion of function is properly normative because the system's expectations can be correct or incorrect.

RSMSs

A complex process involves many other processes – even the simple candle flame involves combustion and convection. But some processes exhibit the ability to initiate and/or inhibit their sub processes. A SMS that does this is a *recursively self-maintaining system* (RSMS). Campbell and Bickhard stress that in order to perform this feat, RSMSs require some infrastructure, what they call a 'switching mechanism'. This is the mechanism through which the RSMS is able to turn sub-processes on and off.

3. A system is recursively self-maintenant if and only if it
    a. is a SMS,
    b. possesses a switching mechanism to initiate or inhibit at least one sub-process.

The canonical example here is the *E. coli* bacterium that either 'swims' or 'tumbles'– swimming and tumbling are distinct sub-processes, modulated by a network of proteins (tumbling is in fact 'not-swimming', so there is one process that can be active or not active, thus yielding two distinct behaviours). If the swimming bacterium encounters a path along which the concentration of sugar increases, it will keep swimming, but if the sugar concentration stops increasing, it will alternate tumbling and swimming until it finds another attractant gradient. In this way, the bacterium is always swimming towards the attractant (Alon et al. 1999).

Here I will present two further examples that highlight how minimal the shift is from SMSs to RSMSs. The first example is the *Rhizostoma*. As observed in his 1934 monograph *A Stroll Through the Worlds of Animals and Men*, Jakob von Uexküll describes the multi-cellular 'marine medusa' as a 'swimming pump mechanism'. All its internal processes are governed by the pumping motion – its ingestion, digestion, movement and breathing. Von Uexküll's description of the mechanism cannot be bettered so I quote it in full:

> "To ensure continuity of this motion, eight bell-shaped organs are located on the periphery of the umbrella, whose clappers strike a nerve end at each beat. The stimulus this produces elicits the next umbrella-beat. In this way the medusa gives herself her own effector cue, and this releases the same receptor cue, which again elicits the same effector cue *ad infinitum*. In the medusa's world, the same bell signal rings all the time, and dominates the rhythm of life." (p. 32)

The medusa is a complex multi-cellular organism, but is merely a SMS[20]. Its life consists in one reflex process, with no further infrastructure. It safely ignores its environment – the only stimulus necessary for its survival is provided directly through its own continued movement, which serves only to ensure the further continuation of that movement.

[20] The example of the sea medusa *Rhizostoma* used here is simplified. I am using the example as it was understood in 1934 by von Uexküll. Since then, the sea medusa has been observed to exhibit differential behaviours and deploy an intelligent and complex hunting strategy. However, the example is extremely neat, and I beg the reader's patience as I belittle the medusa's true capabilities to make my point.

The sea urchin is not far above the medusa in the hierarchy of life. Its parts are self-contained receptor-effector systems. There is no central control governing their movement. Indeed, it may seem as if it is just a slightly more complex SMS – what von Uexküll calls a 'reflex republic'. The urchin has spines, which are simple self-contained systems – when one encounters a stimulus it orients itself towards it. Scattered amongst the forest of spines however are claws. Claws seem to work together, but like the spines, each is self-contained. They respond to the presence of an approaching object by springing outwards, biting and releasing venom. But here there is a puzzle – the claws are constantly surrounded by objects, moving this way and that – spines – but they do not respond to them. The claws exhibit differential responses dependent on whether approaching objects are part of the sea urchin or not, despite being self-contained systems. The puzzle is solved upon discovery of the sea urchin's simple switching mechanism – a chemical called *autodermin*. Autodermin inhibits the claw's snapping response, and is secreted by all the urchin's spines and claws. However, the coat is not thick enough to inhibit the claw's response, so when an alien object is close, it attacks. But, because there are *two* coats of autodermin between a claw and a spine, the barrier is large enough to inhibit the claw's attack. Thus the sea urchin is an RSMS – the organism exhibits at least two separate processes (attack, do not attack) and has an inhibitory mechanism (autodermin secretion) that enables appropriate switching (via inhibition) between them.

This small step up in complexity brings with it the philosophically pertinent feature, *autonomy*, a RSMS 'strikingly provides for some of its own viability conditions' (Campbell, 2009). In the case of *E. Coli*, by searching out food, in the sea urchin by defending itself from potential harm. The presence of the switching mechanism allows the system to exhibit rudimentary choices – in a very minimal sense the sea urchin chooses (in virtue of its physical and chemical organization) to attack foreign bodies, but not to attack itself, similarly the bacterium chooses to swim towards food, and explore when there is none.

It is helpful to note two further consequences that follow from the autonomy and functionality we have recognized in RSMSs. The first is that an autonomous, goal-directed system constitutes an *agent*. The rudimentary movements of the bacterium are the result of a rudimentary choice, directed towards rudimentary goals. They may be thusly framed as the rudimentary *actions* of a rudimentary *agent*. We might say that this combination of features allows us to take Dennett's intentional stance (1987) towards the system, finding that agent-level explanations more easily capture some interesting aspect of the organism's behaviour that is manifest only in the real patterns brought about by its switching mechanisms.

Secondly, RSMSs are capable of being in error. Given that their actions are goal-directed, the action can fail to achieve that goal. The choice that led to the selection of that failed action was thus a mistake, an error. Importantly, this error is error *for the system*, that is, given that the system embodies expectations about the future, it is its own expectation about the future that is in error. We might argue that this is a property shared by SMSs, given that they also possess a function, and thus also can potentially be in error, viz. by malfunctioning. However, the presence of a switching mechanism is important here. Without a switching mechanism, there is no opportunity for alternative behaviour – a non-RSMS is physically incapable of taking a different course of action, and this means that whilst the system can malfunction, it cannot have taken the *wrong* course of action, from its own perspective – it only has one course of action, which is neither right nor wrong, it just *happens* until it malfunctions and stops happening. The RSMS, with its switching mechanism can take the wrong course of action because it has more than one possibility open to it.

Thus, the RSMS can be in error. Crucially, this also means that these errors are *potentially detectable by the system*, as this is error from the system's perspective. Whereas, as soon as a simpler SMS is in error, it will be outside of its conditions of viability and die[21]. Equipped with the right infrastructure, the RSMS may be capable

[21] The SMS and RSMS can be distinguished in FEF terms by the complexity of their statistical model. The SMS can suffer very little perturbation in its environment before dying, so the statistical model it embodies is tightly constrained around a very small subset of states. The RSMS is more adaptive, and thus embodies a model that expects to persist in a much wider range of states, and as such can cope with (potentially) much larger perturbations in its environment.

of detecting whether the expected outcome of its chosen action actually occurred. A bacterium doesn't have this capability, but some animals clearly do. This indicates the next step up in Campbell's ontology: systems capable of detecting their error.

EDSs

Campbell (2009) specifies a class of systems that can detect when they are in error, which I shall (perhaps a little unimaginatively) label *error detecting systems* EDSs.

4. An Error-Detecting System (EDS):
    a. is a RSMS,
    b. possesses a mechanism for detecting its own errors,

It is at this stage that a system begins to exhibit characteristically mental traits. Remember, EDSs are a class of processes/systems that work to preserve themselves within their environment by actively deploying a range of processes in response to different conditions, and are capable of telling when they have deployed the wrong process in a given situation. So, the system can use that error-information to alter their online behaviour. Being a complex kind of SMS, the EDS will use that error to help minimise its free-energy. Whilst I am trying to present minimal conditions for mentality, it is good to remind ourselves that we are now firmly in the realm of complex systems.

The most significant property possessed by EDSs for the current treatment is the potential possession of mental representations. Once a process is not only capable of being in error, but further, that error is detectable by the system, then the system can be well-understood as using representations, which are properly representations from not only our observers' perspective, but also from the system's perspective[22]. At this

[22] That is, the system itself uses them as representations, it is not just an explanatory gloss for us, external observers.

point, the process that functions to initiate or inhibit a process or course of action represents that course of action for the system, as it now possesses the normative property of correctness (or incorrectness) for the system. That is, the system is capable of judging whether the switching process was correct or incorrect to activate or inhibit a given sub-process at a given time.

The position I outline here is stronger than that put forward by Bickhard (2000). Bickhard argues that representation emerges in RSMSs in virtue of the fact that they are clearly capable of distinguishing (at least) two different types of environment, and possess differential expectations associated with each type of environment that leads to them choosing to pursue a particular action in a given circumstance. Thus, Bickhard argues, RSMSs possess primitive representations (in the form of expectations) about their environment. I agree that RSMSs do possess primitive qualities necessary for representation: discrimination of appropriate courses of action, and normativity; but that they lack the ability to leverage those qualities *qua* representations in the way that an EDS can. That is, an EDS uses some of its sub-processes as representations. Whereas in the RSMS, though some of its sub-processes are possibly interpretable as representations, that notion is not strong enough to fully naturalise our understanding internal representation. Namely, they are not being put to use as representations, so they will fail Ramsey's job-description challenge. On the other hand, the mechanisms present in an EDS make it very difficult to understand the internal workings of the system without invoking representational vocabulary, which is an excellent reason for supposing that EDSs really are making use of representations.

Most animals with a cortex seem to exhibit the kinds of behaviours we expect in an EDS, but once a system gets this complex, it becomes difficult to say definitively when it has crossed the RSMS-EDS threshold. So the representation wars may continue, but in the context of this ontology, such a disagreement will be an entirely empirical question, decided by the presence or absence of the appropriate error-detecting infrastructure and a given system.

In what then, might a ‘mechanism for detecting its own errors’, consist? Answering this question is central to the empirical search for EDSs and a strong conceptual understanding of the notion of representation I am constructing here. The requirement is quite specific, and thus rules out a large proportion of possible RSMSs – however, I will do my best to specify as minimal a notion as possible in the interests of parsimony. In order to detect error, first a system must be capable of detecting its own expectations – SMSs and RSMSs may *embody* statistical expectations, but none of those expectations need be accessible to the systems’ inner processes. However, the EDS must have some expectation to compare to actual data in order to detect error. So an EDS must possess:

1) *Expectations*. Some state or sub-process that functions as an expectation about the future.
2) *Comparison process*. Some sub-process whose function is to compare that expectation to the *target* – that which the expectation is about;
3) *Error processor*. Some sub-process whose function is to use the output of the comparison process (error).

Note that there is a difference between the way the expectations of an EDS expect, and the way that a SMS expects. A SMS *embodies* a statistical expectation of future states, in the very weak sense required by FEF. The expectations in an EDS are states or processes that *function for the system* as expectations. They might be quite minimal computational states that do not embody an expectation in the way that the whole system must, that is, they acquire their functional role in virtue of the functional milieu of the system in which they are situated (for instance, by being the output of a generative model in a PP system). In order to function as an expectation, there are some necessary elements we can identify. There must be some process that functionally relates the expectation to what the expectation is about – without such a process, the expectation isn’t about anything *for the system*, and so cannot be functioning as an expectation for the system. This brings us to the comparison process.

The comparison process is the process through which error is generated, this is a necessary element as there must be an error signal in order for error to be detected by the system. The comparison process takes the expectation and the target as input, and outputs the difference. The target is that which the expectation is about. For instance, I might have an expectation that the suit of next card to be dealt is spades. The target of the expectation is thus the suit of the next card. So, the comparison process would compare my expectation, that the suit of the next card will be spades, to the actual suit of the next card dealt. In this way we see how the functional relevance of content (i.e. the target of expectation) emerges via the presence of an expectation and a comparison process. At this point we might raise our hands triumphantly, having found the minimal conditions for grounding content in our ontology. However, in order for the comparison process itself to be functionally relevant to the system, and not just a means to satisfy the philosophical project of naturalising content, the output of the process must be useful for the system. So there must be an error processor.

What it takes to be an error processor is thus extremely weak – this condition bottoms out to mean any process that exploits the error signal and contributes to the maintenance of the system. We are presently concerned with error *detecting* systems, and we have successfully defined the necessary and sufficient conditions for such a system, saying much more about how the detected error needs to be used by the system would be over stepping the mark, and more importantly, bloating our ontology of the minimal conditions required for representational contents. However, there are some instructive additions to make to the ontology which I will now explore briefly – but for debates about representation, the primary focus of this thesis, all the action is around EDSs.

Beyond EDSs

There are two further classes identified by Campbell that it is useful to consider. These are Flexible Learners and Reflective persons.

5. A Flexible Learner

a. is an EDS
b. possesses a sub-process whose function is to make systematic alterations to other sub-processes in response to error.

The first thing to note with FLs is that the alterations made by the sub-process responsible for learning must be non-random, that is the alterations must be systematic. Without systematic alterations some alterations would necessarily damage the ability of the system to self-maintain. However, by possessing some algorithm for making alterations, the learning process minimises the risk of making such mistakes. The second note to make is that I have specified that learning must be done in response to error. So while I left the nature of an error-using sub-process open in my description of EDSs, FLs use error in a specific way (that is not to preclude them from also using error signals in other inventive ways, just that in order to count as a learner, they must use it in this specific way). The reason for this specification is to ensure that these systems remain autonomous. We can imagine systems capable of making systematic alterations in response to external stimuli that indicate desirable and undesirable behaviour, i.e. supervised learners. However, such systems are limited (inflexible) by their supervisors' knowledge and constant feedback. Flexible learners, using their own error signals, are autonomous, self-supervised learners. They can use error signals to evaluate their own actions.

Campbell (2009) notes that this evaluative capacity is one that emerges only with the infrastructure of FLs. Possession of a systematic learning process enables the system to assign values to future actions in virtue of knowledge of the outcome of similar previous actions. Compare to the RSMS, which responds differentially to various stimuli, and thus embodies its preferences, the FL is capable of learning new preferences, allowing it to adapt to novel situations and a changing ecological niche. The alterations made by the FL may tacitly or explicitly encode these new preferences in the system. For example – a classical computational system may explicitly alter the value variable associated with some action, but a dynamical system may alter the dispositions of the system by altering some higher-level variable and thus changing the attractor space of the system. So in some FLs it may be right to say that the system possesses explicit representations of value, but not in all; however, in all FLs it will be legitimate to describe them as having the capacity to evaluate actions because the alterations will reflect the normative associations of previous endeavours (i.e. the level of error generated by them in the past).

While it is difficult to identify a biological threshold between EDSs and FLs, we can instead consider computer models of cognitive systems. Firmly in the FL category is the Rao and Ballard model of PP that we examined in part two. In fact, it is clear that PP systems provide an attractive framework for making sense of FLs, given how they are built around the processing of error, and that they learn to minimise it, fulfilling conditions 5a and 5b. It seems that an RSMS (4a) furnished with a sub-process approximating a predictive processor would count as an FL. So, if evolved nervous systems are well-understood as PPs, then our analysis in this section goes some way to justifying the claim that many animals do possess internal, mental representations.

6. A Reflective Person
    a. is a Flexible Learner,
    b. possesses a sub-process whose function is to monitor some subset of its expectations and mental processes
    c. possesses a sub-process that processes the output of the monitoring process.

With this category, Campbell (2009) hopes to define the class in which we find humans. He argues that (primary) consciousness emerges once a system is capable of monitoring (6b) and reflecting upon (6c) its own thoughts. I believe this is a step too far, and find the claim under-argued. Furthermore, I wish to remain quiet on the question of consciousness which I believe to be tangential to the mental representation debate and indeed can distract from the pertinent points in the debate, which relate just as strongly to conscious and sub-conscious phenomena.

However, the recognition that we might stack, or nest, SMSs in such a way that one SMS exists as a process that functions within a larger more complex system to, perhaps, serve as a reflective process in a RP, is an attractive idea. We can imagine multiple layers of such processes that generate more and more abstract abilities, which in combination with learning processes could form a powerful cognitive engine akin to the human mind.

Before exploring these claims further, and what they mean for PP's image of the mind, I want to fully flesh out the key move in the argument. We will come back to mental representation in section 9.3, but having outlined an ambitious ontology in this section, seeking to provide some scaffolding for a naturalised account of function and mental representation that also points to a strong continuity between the sciences of life and mind, it's worth going over a few things, and how this framework fits with the frameworks developed in chapters eight and nine.

19.2 Proper Perspectivalism

We have seen that with the development of SMSs, we begin to find proper functions in the natural world. This is a significant and important step forward for our understanding of mental states and the potential for contents. This theory shares important qualities with the perspectivalism discussed in chapter eight. This derives from the process metaphysics that underpin the framework, which allows processes, and thus SMSs to 'nest', one inside the other, creating a functional hierarchy. This hierarchy is both well-defined and reflects the areas of study of the disciplines of study in the sciences of life and mind. It is hoped that this argument will provide a principled basis for functional ascriptions in these disciplines, and thus ultimately, for mental representation, as we have seen, which piggybacks on the normativity that emerges with function.

To close this final chapter I will analyse PP through the Proper Perspectivalist lens. This will reveal the important insights we can glean from PP's densely nested architecture. These insights will reveal the recursive functional structure while remaining sensitive to the dynamical framework inherited from the FEF. By doing this we can clearly see the interesting claims being made by PP about the organisation of the mind, which could form the basis of future research.

Nesting Processes

In the previous section I provided an argument that grounds function in natural processes, and treats it as a natural property of certain systems (SMSs). Our analysis so far constitutes an argument in favour of *proper function* grounded in free-energy minimising processes capable of maintaining themselves. In this subsection I am going to show how this notion of proper function also borrows important characteristics of perspectival function. This will be through the way processes nest within each other in the natural world.

Before I perform this conceptual juggling act, I want to remind the reader of the salient differences between proper function and the perspectival account of function as discussed in chapter seven. The perspectivalist holds that all function ascription importantly depends on the explanatory perspective being taken. By contrast, the advocate of proper function holds that some functions are not a matter of perspective – that they are real insofar as they can be grounded in natural processes, entities, and relations. The position I defend here is that perspective plays an important role in the ascription of function in scientific explanations, even in domains in which there are proper functions. In many cases a sub-system will have *multiple* proper functions, which one is relevant in a given context will depend on the perspective of the explanation, but nevertheless, each is fully naturalisable.

This difficult position comes about in virtue of the two routes through which a system comes to possess function, as detailed above. The first route is by being a SMSs, bounded by a Markov blanket[23]. In this case the system just has the function to

[23] Note that possessing a Markov blanket does not make a system a SMS. I am just using Markov blanket as a tool for demarcating a system from its environment. Once a system's boundaries have been defined in this way, we *then* ask what properties that system possesses, which will determine whether or not it is a SMS, RSMS, EDS etc.

minimise its free energy. The second is by being a sub-process of an SMS that is dynamically presupposed to be performing a function that contributes to the minimisation of free energy at the system level. Both of these are proper functions.

Take the candle flame as an example. We noted one of its sub-processes, vaporisation. The vaporisation process is key to ensuring the candle-flame's maintenance, it's the process that allows the flame to exhibit less dependency by enabling access to more than one type of fuel. Thus the vaporisation process performs this enabling function for the candle flame. The flame itself possesses the function to continue simply in virtue of being an SMS that thus embodies the expectation that it will continue. The vaporisation process possesses the function of generating fuel which is a necessary for the flame's continuing.

This simplest possible case falls directly out of Bickhard's notion of dynamic presupposition, which we discussed earlier. However, the picture can get more complex, very quickly. I will now consider a few toy examples that demonstrate the possible complexities here.

The first and simplest possibility is that a SMS is a sub-process of another SMS. This is a relatively common situation, but the possible function ascriptions need to be spelled out. Consider an SMS (MIKE) is a sub-process part of the micro-level organisation of the system of scientific interest, a larger, macro-level SMS (MACK).

The situation is simple for MACK. MACK has just one function – to minimise its free energy. Our interest in the system is to explain how MACK does this. On the other hand, MIKE has two proper functions. One is to minimise its own free energy, the other will label the contribution MIKE makes to the minimisation of MACK's free energy, which will be one of a number of dynamic presuppositions required for MACK to continue as a SMS. My core claim is that MIKE has *both these functions as proper functions*, but that for explanation, only one is salient, and which one that is depends on the explanatory context. If our scientific interest is in MIKE, and not MACK, then its function to minimise its own free energy is salient, but if we are interested in MACK, then MIKE's role in minimising MACK's free energy is its salient function.

Allow me to illustrate using our earlier (somewhat idealised) example of a sea medusa, which is a complex, multicellular SMS. Recall that the medusa is not a RSMS, as the animal possesses no switching process to inhibit or activate any of its sub-processes. However, it does possess sub-processes, these sub-processes have functions in virtue of being sub-processes of a SMS. One simple case would be the medusa's digestive system, which has the function of absorbing nutrients from sea-water being pumped through the stomach. The digestive system contributes to the minimisation of free energy for the medusa by fulfilling its embodied expectation that it will continue to have energy to fuel its movement and internal processes. However, the digestive system is itself also an SMS – it uses the products of its own process to help break down incoming material. It thus maintains itself in the minimal sense required by the definition – by reducing its dependency on external conditions. In this sense the digestive system minimally expects to continue, and thus possesses its own function, as well as performing a function for the medusa.

MIKE and MACK, and the medusa thus demonstrate what I call 'nesting'. Nesting occurs when one or more SMSs are sub-processes that make up the micro-level dynamics that together form a macro-level SMS. The digestive system is the MIKE to the MACK of the medusa. Organisms are the MIKEs to the MACK of an ecosystem. Human individuals are MIKEs to the MACK of a business corporation; businesses, organisations, governments and individuals are the MIKEs to the MACK of society. But note that these latter cases are clearly more complex than the simple MIKE and MACK example – for some of the systems in question are not mere SMSs, but RSMSs, EDSs, FLs, and RPs.

The complexity here derives from the possibility that a more complex system may be a component in a less complex system, which may lead to confusing function ascriptions. Imagine a sock factory with many workers. This factory is a simple SMS – it produces socks, and that's all it does. Each worker has simple and clear instructions to follow. One worker orders a roll of cotton thread every day. Another worker weaves the cotton thread into sheets. Another cuts the sheets into specific shapes. Another stiches the shapes together to form socks. Another packs and sends the socks out. Another worker takes payments for the socks and a final worker deposits the payment at the bank. The factory is a very boring place, but it is a sustainable enterprise, and successfully maintains itself in its specific socioeconomic niche. Note however that its sub-processes are much more complex – they are human beings, Reflective Persons.

The functional ascriptions in these cases are perhaps surprising, but not so different from our previous examples. The factory has a function in virtue of being a SMS – it embodies the expectation that it will continue. The workers have a function derived from their role in the factory system, as detailed above. But the workers themselves also have a function in virtue of being a (complex kind of) SMS – they also (like neurons or other cells) embody an expectation that they will continue. Their inner workings are, paradoxically, more complex than the inner workings of the factory system, containing robust switching mechanisms, error detection processes and learning processes that allow them to perform a vastly larger suite of actions than the factory. But surely, one might say, all this infrastructure is also part of the factory's inner workings. After all, these are sub-processes of the factory's sub-processes, and so by extension, part of the factory.

This is a puzzle for our ontology. How is it that a sub-process can possess properties not shared by the macro-level process? The answer is that the macro-process can constrain the micro-level sub-processes in such a way that some of their capabilities play no role in the success of the macro-process's self-maintenance. That is, it is not a *dynamic presupposition* of the factory that its sub-processes be Reflective Persons or RSMSs. In a suitably advanced future, the factory described above might conceivably be fully automated, run by systems that are not even SMSs[24].

[24] These are interesting cases because they teach us about the continuity between biology, psychology, society, and culture. It seems to me that corporations (etc.) are simpler systems than human beings, yet human beings self-organise to bring about these systems. The framework may even teach us a little about how to build successful organisations – we might want to build in an error-detection department that can

As far as each of these SMSs can be defined by a Markov blanket, they will have a well-defined function, which imbues them with normative properties. The FEF provides us with a framework to help us get a grip on the dynamic presuppositions that constrain the function of the micro-level systems. In particular, the FEF presupposes that the micro-dynamics of an SMS help facilitate statistical expectations about the future, and contribute to the system's anticipation that it continues in line with those expectations. This gives us the principled way in which MIKEs can contribute to the maintenance of MACKs – organising themselves as part of a statistical expectation, and/or involving themselves in the anticipation of future events.

Dealing with Swamp Man

As outlined in chapter one, the etiological theories that traditionally ground proper function in mechanics are not fit for purpose. Proper Perspectivalism is not vulnerable to the same criticisms. Let us first consider swamp man. These etiological theories fail because swamp man has no evolutionary history, there has been no historical selection for his organs and processes. However, swamp man is undeniably an EDS, and thus a SMS (else his low-entropy organisation would quickly return to a high-entropy state). So, swamp man has a proper function to minimise his free energy. Given that certain things must be presupposed for the dynamics of SMSs to govern swamp man's behaviour, as described by the FEF, we can establish functions for the processes present at the micro-level. So we gain a functional understanding of the entire swamp man system from the dynamics of SMSs, in virtue of swamp man being a SMS.

We can also attribute proper function to a much wider range of phenomena than etiological theories. Evolved life forms are just one class of SMS. In the abstract we can imagine an SMS in Conway's game of life counting as having proper function, and we can also attribute function to low level chemical processes that in SMSs, that may not themselves be evolved, or have the right kind of etiology. We can also more

disseminate its findings through a learning programme in such a way that makes the organisation more adaptable and thus more likely to survive socioeconomic changes.

easily make sense of the claims of 4E cognition. The environment is imbued with function for the SMS insofar as the SMS interacts with it. As the SMS involves external elements in its self-maintenance, those elements will have a function that ultimately relates to the minimisation of that SMS's free energy.

PPPP (Properly Perspectival Predictive Processing)

What the research into PP can teach us is that one effective way of structuring an EDS is to nest several EDSs inside it. This is very similar to the notion of a hierarchical organisation, but has a slightly different emphasis. From the process/systems point of view I have been pursuing, it doesn't make sense to treat each layer in the hierarchy separately, what matters is where you draw the Markov blanket. That is to say, the only layer in a PP system that can be properly studied as an EDS independently from the others is the top-most in the hierarchy, which is the inner-most on the nesting view. This segment is the sub-process of all the other processes, and the only segment that comprises an EDS bounded by a Markov blanket without including further EDSs within it. See the figure 1 below for a visual explanation.

It is worth noting that each layer may contain several sub-systems, not all of which are recruited for the task at hand. This includes the inner-most layer, which is likely to be constituted by a cluster of flexible modules, which only together form an effective error-minimising system given the complex environment of the organism. These may be constituted by any kind of process, not necessarily an EDS, RSMS, or even a SMS. When we develop technologies capable of investigating the precise function of these sub-systems (perhaps being neural modules instantiated by cortical microcolumns) we can ask what kind of system they are, and thus determine what kind of properties they possess. With our ontology sketched here, we can come to understand how properly mental processes (EDSs and above) are made up of other processes, some with their own proper functions (SMSs), but all of which will possess some kind of function in line with their role in the larger neural system.

We can understand the interactions between these nested processes in terms of the constraints they impose on each other. Each inner SMS effectively constrains the dynamics of its immediate environment through active inference. Ultimately this results in the organism constraining the environment it inhabits in order to maintain itself within the small set of conditions that enable it to preserve its integrity. This all falls easily out of the FEF formal framework. We can understand this neatly in dynamic terms using the notion of attractors. The state of the nested systems is a variable governing the dynamics of the larger system (predictions) and changing this variable alters the state-space of the larger system, forming new attractors and equilibria in that space. Of course the state of the larger system is also a variable in the equation governing the dynamics of the nested system (error), so these systems are coupled. Understanding the entire system requires taking account of the dynamics of each system in the nest.

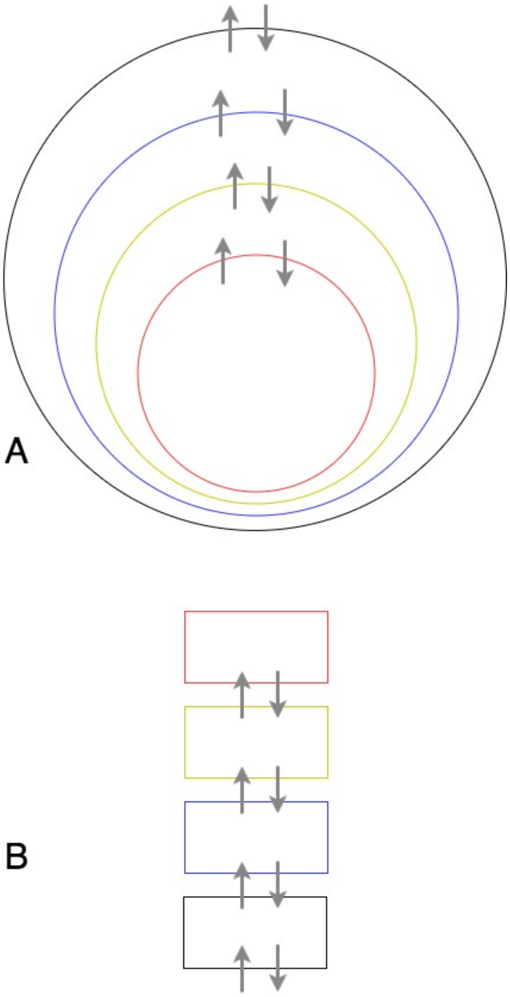

Fig. 9.1: Nesting. Diagram A represents nested EDSs. Each line is a Markov blanket. For an organism, the black line would represent the sensory periphery, the blue line might represent the neural periphery, and the yellow and red lines represent the Markov blankets of internal neural PP layers (each an EDS). The dynamics of the system demarcated by the blue line cannot effectively be treated separately from the dynamics of the yellow and red systems. Diagram B shows the usual hierarchical way PP systems are understood. Each layer is discrete and can, in principle, be isolated from every other layer in the hierarchy. In both diagrams the grey arrows represent the interaction between the systems and their immediate environment.

The function of each layer in a PP system is properly perspectival – the relevant function of a given sub-system depends on what larger system we are interested in explaining. We must identify a Markov blanket to define the system we want to understand, and then we can start to investigate the way in which the dynamics of the nested sub-systems constrain the dynamics of the overall system, which may help us make narrower functional ascriptions than the broad brush 'to help minimise the overall system's free-energy'. It is worth re-iterating that each nested SMS may have several different functions, depending on the perspective taken for the explanation. However, this is not the same perspectivalist position that appeared in chapter seven. What Proper Perspectivalism holds is that insofar as these functions are dynamic presuppositions of a SMS, they are proper functions. However, the function that is relevant for a particular explanation will be determined by the perspective taken by that explanation. So, once a perspective has been adopted, there will be a single salient proper function to be appealed to.

9.3 Explaining Minds Revisited

In part three, thus far, I have fleshed out three theoretical frameworks that can help us make sense of the PP paradigm. I argued for the importance of the FEF, and for DST – mapping out the important links between the two, and presenting the worrying implications of those links for a the mechanical-computational understanding of mental representation at the heart of PP. Then I delved into the interactivist process ontology of Campbell and Bickhard, which gives us an additional perspective on how

PP systems (as EDSs) fit into the sciences of life and mind. Out of these discussions, at the intersection of FEF and interactivist ontology, an attractive account of proper function was formed. One that builds in sensitivity to the importance of perspective for function ascription, but does not flirt with fictionalism, instead providing a naturalised foundation in systems theory and FEF.

However, all these developments were motivated, and prefaced by several arguments against the mechanistic paradigm which currently justifies the representational understanding of PP. Recall that PP is understood as a theory of computational information processing, computational explanation is a sub-class of mechanistic explanation. So, if there is a problem with mechanistic explanation then there is a problem for computational explanation, and if there is a problem for computational explanations, there are problems for PP. By demonstrating that FEF and DST provide a way of understanding PP, we saved it from this fate. The cost was the notion of representation, which is a functional kind, an outdated legacy element of mechanistic explanations.

The lynch-pin in our argument against mechanism was its failure to properly ground function. But now we have done it, albeit a little unexpectedly, through a thoroughly dynamical, processual lens. Have we inadvertently saved mechanistic explanation? That is the question I will use to close the thesis. I will look now at whether the mechanistic paradigm can take our conclusions of the last section and run, carrying on happily as before, or whether perhaps the functional notion developed is not in fact compatible with mechanical explanation at all, thoroughly drenched in dynamics, as it is (9.4); I will close with the possibility of peace, a vision of how dynamics and mechanics might work together to help us understand PP, and leave us with a more nuanced understanding of mental representation as we move forward (9.5).

9.4 Mechanics vs Dynamics

For mechanists concerned with the theoretical foundations of their explanatory project, unsatisfied with the traditional accounts of proper function, and perhaps a little unsure about perspectivalism, Proper Perspectivalism offers an alternative. It's an alternative that doesn't rely on evolutionary histories, but is tailored for biological systems. It allows us to make sense of the deeper distinctions between biological systems, designed artificial systems, and self-organising artificial systems. Self-maintaining artificial systems are functional in just the same way as biological systems, so we might now make principled comparisons, and learn something about the parallels between the two.

Better yet, radical enactivists and dynamicists who advocate eliminativism, now have a little less ammunition. Where we find systems capable of detecting and using their own errors, we can find internal representations. As outlined above, this is not a trivial requirement. A significant amount of infrastructure is necessary: expectations, a comparison process, and a system that processes the error. It is also worth noting that the explanatory value of representation talk will increase with the complexity of the EDS. A very minimal EDS capable of only a few simple behaviours would, on my view, satisfy the ontological requirements to count as possessing representations, but whether or not the representational strategy would be the most appropriate would still be questionable. However, a more complex EDS, with a wide range of possible behaviours and perhaps the infrastructure necessary to qualify as a FL, would almost certainly be well-understood as using representations. When comparing paradigms capable of making sense of representational systems, the computational strategy will win out every time. We can take the FEF-Interactivist insights about function, and forge ahead with cognitive science's favoured tools, at peace with our efforts, and able to effectively respond to the eliminativist worries that threaten to de-stabilise the collective enterprise.

Proper function, and with it, mechanics, computation, and representation, are somewhat re-imagined, but are basically safe and secure. We can enter fully into a realist belief in mental representations that have the real function to bear real contents. If only it were so simple. I believe that the dynamicist may have some legitimate grievances with the mechanists' cut and run. The Properly Perspectival notion of function is a dynamical notion. I will now address the fact that having a notion of proper function licenses mechanics, but argue that this particular formulation puts dynamics first, and underscores Chemero's argument (2009), summarised in chapter nine, that dynamics is explanatorily primary.

Using Clark's (1996) useful distinction between component-dominant dynamics and interaction dominant dynamics, we can see that the interactivist ontology used to scaffold our understanding of functional SMSs is best understood using the tools of dynamics. Clark (1996) notes that systems that are well understood by mechanistic explanation exhibit component-dominance. That is, the capabilities of the system are brought about primarily (this is not a binary distinction, but marks two ends of a continuum of possible systems) by the properties of components and their organisation. On the other hand, interaction-dominant systems (such as most living things) are best understood by DST. The capabilities of interaction-dominant systems are primarily brought about by changes in the emergent collective variables measured over the states of large numbers of the components. For instance, ecosystems are interaction-dominant, and they are captured by the collective variables of biomass and energy transfer through various biological strata.

Biological SMSs are interaction-dominant systems. The behaviours relevant to their self-maintenance are defined in terms of their macro-behaviour that allows them to interface effectively and efficiently with their environment in order to exploit energy sources. Given this, the language and tools of DST are best placed to make sense of them. The resulting account of function itself is deeply dynamical, and thus pushes against the standard theories tailored to the interests of mechanistic explanation. Understanding function is not possible through mechanistic decomposition, but instead requires understanding the dynamic properties of the macro-level that instantiate a SMS.

What this means is that explanations that invoke proper function, such as mechanistic explanations in the sciences of life and mind, must ultimately ground themselves in the dynamics of the phenomena. Mechanical understanding of brain processes and biological processes is important, most notably for intervention in the form of medicine. However, the path to full understanding of the mind lies with the dynamic paradigm. Chemero (2009) rightly notes the empirical debts that dynamics imposes on mechanics in terms of scope and prediction. Now we can add a conceptual debt that mechanics owes to dynamics – normativity through function. We have shown that normativity is necessary for understanding the function of mechanical parts, but that proper function itself is generated from interactive dynamics of simple elements that work to minimise free energy and result in a SMS. A thorough understanding of biological mechanisms must be prefaced by a thorough understanding of their dynamics.

9.5 Peace?

Contrary to the arguments in favour of pure mechanism and pure dynamics, I believe that Proper Perspectivalism provides a principled fulcrum to lever a pragmatic middle way between the two. There have been previous attempts to identify such a middle path, for instance by Clark (1997) and more recently by Bechtel and Abrahamsen (2006), and Anderson, Richardson, and Chemero (2012). The path is a pragmatic one, a demonstration that there can be a happy marriage between the paradigms, using detailed discussion of empirical work. I will here rehearse the main points of this vision: the mutually constraining relationship between mechanics and dynamics, soft-assembly in neuronal populations, and the incorporation of 4E cognition. What Proper Perspectivalism offers is a firm philosophical foundation in which these relationships crystalize to help us understand how mentality, in the form of mental representations, can be understood rigorously and used to drive the sciences of life and mind forward.

> “We propose that cognition is supported by a nested structure of task-specific synergies, which are softly assembled from a variety of neural, bodily, and environmental components (including other individuals), and exhibit interaction dominant dynamics.” (Anderson et al., 2012)

This short passage encapsulates the vision of the mind being driven at here. We have so far investigated the idea of a nested structure, providing some further understanding of it in terms of free-energy minimising processes; we have also considered in some detail the embodied and embedded approach that explains how a system can bring to bear ‘external’ elements to enhance and scaffold cognitive tasks; we have also just heard a little about interaction dominance, which links into our broader discussion of dynamical modelling and explanation. That leaves just the idea of ‘soft assembly’ unexamined. Let us consider this now.

Soft Assembly

Soft assembly refers to the capacity of a flexible, plastic system to recruit its sub-processes in a wide variety of different contexts to solve many different problems and perform related tasks. Solutions to problems and plans of action are soft assembled insofar as the system can construct and deconstruct them very quickly, and perhaps reconstruct them differently in slightly different contexts. In this way, a system with a large number of small, highly specialised units (such as cortical microcolumns) can solve a large number of problems over many different domains and modalities with the minimum expense of energy by recruiting just the right units at just the right time in precisely the right configuration.

This seems like an impossibly difficult feat, but nevertheless, it does seem to capture accurately, the behaviour of the brain (Anderson, 2010). Furthermore, PP may offer an elegant computational solution to the problem. Clark (2016) sets out a compelling vision of the brain as a PP system that continually soft assembles its resources to minimise the flow of free energy in the form of error. The most powerful enabler of soft assembly is precision estimation (see chapter four).

Precision estimation allows the system to quickly ‘turn up the volume’ of certain units and along certain pathways to allow a rapid re-configuring of the effective connectivity of the system, enabling soft assembly in response to changing stimuli. Further, because precision estimation can be used to modulate both the influence of internal elements and the influence of incoming signals, when in an appropriately scaffolded environment, cognitive duties can effectively be offloaded to elements outside the head, like a smartphone or computer, by turning up the volume on incoming data from those sources, and freeing up internal resources to tackle other problems, and better predict the unfolding of the situation. Soft assembly is a powerful tool, allowing a system to make the best use of its resources, and PP provides an elegant story about how a system might achieve these feats in practice.

The Place of Dynamics and Mechanics

Let's now look at how to forge a middle way between the mechanistic explanatory paradigm and the dynamical explanatory paradigm. The rough shape of the answer is this – we should use both styles when they are appropriate; dynamics is (as discussed) better at dealing with phenomena that arise from interaction dominant systems, and capturing the way the collective variables that emerge from those interactions change (Clark, 1997); mechanics is useful for making sense of the causal order of a system – dynamics gives us descriptive laws that often provides a detailed characterisation of the problem space, but won't help us engineer the system, or intervene in the system, those are the fruits of the mechanistic explanation.

In a limited sense the dynamical stance is primary – it helps us better understand the phenomena that the mechanist ought to be looking at. This feature is captured nicely by Proper Perspectivalism, which, requiring a dynamical approach to define and demarcate a SMS for study, then allows the mechanist to make concrete function ascriptions that allow them to look properly at the sub-processes and infrastructure of the system's microstructure. Abrahamsen and Bechtel (2006) neatly describe this idea: '[scientists] ask what sorts of cognitive mechanisms might be responsible for phenomena that could be characterized, but not fully explained, using equations alone.'

Our analysis does perhaps encourage the extension of the reach of dynamics. The complexity of the brain (estimates of the number of functional units of neural populations in the cerebral cortex range from 300,000 to 2,000,000) may mean that the tools of the dynamicist are needed to tame its massively interaction-dominant behaviour. The key point is that once tamed by mathematical laws, the mechanists will be able to get to work more quickly, armed then with a much better understanding of the phenomena they are aiming to explain. For example, we need dynamics to make sense of why a particular group of neural populations were activated in a given situation, but the mechanist can explain provide a functional context for that activation, delivering understanding of the population's role in providing a successful (or unsuccessful) adaptive response to the situation.

PP has a long way to go as an empirical research programme. However, as my discussions in part three have shown, the framework does provide us with a great deal of guidance on both dynamics and the mechanics. This is another reason to be optimistic about the future of the paradigm. The dynamics of FEF constrain the problem space, and give us an insight into what it means to be mental at all by helping us make sense of recursively self-maintaining, error-detecting systems. The computational framework of PP inherits these lessons from FEF, and goes further by providing highly suggestive indications of effective neural mechanisms that we can test empirically. PP can also account for and acknowledge the possibility of fine grained mechanisms within the structure. That is, while the system level function will be to minimise upward flowing error signals, the strategies instantiated by neural populations will certainly be extremely varied, and require a more detailed mechanical analysis than that provided by bare-bones PP. What PP does so well is provide an overarching computational structure that shows how a collection of simple mechanisms might be combined quickly and effectively to produce adaptive behaviours.

10 Conclusions

The debate about mental representations has once more been brought to the forefront of philosophy by the predictive processing framework. The stakes are high – this is arguably the closest that cognitive science has been to a grand unifying theory of cognition. Knowledge of how mental representations work is central to our understanding of the mind. They are the building blocks of thought and cognition. PP systems, as we have seen, possess all kinds of interesting elements that are extremely suggestive from the perspective of the mental representation debate. The models are deeply imbued with superficially highly representational language and normative notions, but the intellectual origins of the framework are deeply dynamical in nature, indicating a strong continuity between the sciences of life and mind through basic principles of self-organisation. Simultaneously, as a class of computational model, PP inherits the decades of past discussion regarding the legitimacy of representation talk, and the nature of intentionality.

The aim of this thesis was simply to develop an understanding of mental representation that can be applied to PP. There were several classic challenges to address. The first hurdle was to show that the representational states are grounded – that they have the right causal powers to work as bearers of content. The second hurdle was to show that these representational states could work as *mental* representations – that there is an account of how they possess intentional content. Having elucidated the major difficulty faced by causal accounts of intentionality –an inability to account for error, we considered teleological accounts. However, both the teleofunctionalist and perspectival accounts were found wanting. So a third hurdle emerged – to find a notion of function that would account for intentional content.

First then, a recap of how we tackled the first hurdle – grounding representation. Chapters one and two showed how classical computational systems ground (non-intentional) content through the isomorphism based account of Robert Cummins (1996). In chapter four we saw how PP systems might be implemented using connectionist networks made up of layers implementing generative models. This allowed us to appeal to the kind of account offered by Shagrir (2012) to show that neural networks instantiate Cummins-style isomorphisms. I made this explicit in chapter seven using the work of Nicholas Shea (2007), who showed that we can identify structured dispositions within the computational connectionist models commonly posited within computational cognitive science, of which PP models are a subset. This allowed us to confront the second and third hurdles – grounding *intentional* content and thus accounting for the problem of error by appealing to proper function.

To do so, we appealed to Bickhard and Campbell's systems ontology, coupled with a better understanding of FEF. Together, they enabled us to build an account of how proper function emerges in systems capable of making contributions to their own self-maintenance. The complex interplay of nested systems and processes shows us how these proper functions build upon each other to produce the varied and adaptable biological and cognitive systems we find in the natural world. This account makes key concessions to the perspectival understanding of function – depending on the system of interest, there will indeed be different functions of interest, but those functions are all grounded naturalistically. So we do not suffer a radical relativist pluralism of intentional contents, as entailed by Craver's (2003) perspectivalism. Rather, the sub-systems and processes of a given organism may each possess multiple proper functions, and so, depending on the level of description, may carry different contents. However, the appropriate levels of description have a grounding in natural kinds – Campbell's (1999) ontology has given us a principled way to identify these levels.

What is especially striking for us, as theorists interested in the emerging PP research programme, is its neat fit with this alternative account of function, meaning, and representation. In some ways this is not surprising. We have made explicit use of FEF in grounding out and justifying the claims here; the same FEF that is PP's intellectual grandparent. However, the systems ontology facet of the theory, which does much of the heavy lifting by providing the conditions under which a system will gain new and interesting features that contribute to its self-maintenance, was developed entirely independently of PP. Nevertheless, it is that systems ontology that provides our error-detection condition. Error is the conceptual and mechanical heart of PP. It is the long-term minimisation of error that drives the development and learning of PP systems, so the detection and processing of error is central to our understanding of the framework. This is suggestive. It indicates that PP provides a framework that fits within a robust paradigm that recognises the continuity between life and mind, and that provides attractive, minimal conditions for making sense of intentionality and properly mental representation.

Given this revised conception of PP, let us review some of the implications for our understanding of the framework itself, calling back themes from Part Two, and for cognitive science more generally, recalling the non-computational, DST paradigm.

Action-Oriented Predictive Processing

The core argument in favour of Clark's radical predictive processing as presented in chapter five is that there is a wealth of research that has been done into active and embodied elements of cognition which must be accounted for by a unifying theory of cognitive science, and which PP elegantly and parsimoniously is able to draw together. What is really being sold is a view of the mind as deeply *interactive*, rather than *reconstructive*, and this is the core disagreement between Clark and Hohwy that was discussed earlier.

What has become clear in the light of our discussion in Part Three is that the predictive mind is not very much like a robot that plugs a leaky dam. That is, unless the dam is on wheels and possesses some actuators capable of actively prodding and probing the environment – perhaps altering the water flows that lead to the dam itself. That sort of dam might do its best to find its way to an environment that minimises the amount of water leaking in, in order to further minimise the amount of ongoing work it has to do. Indeed, a very clever leaky dam robot of this sort might build some structures around its walls, or get together with other leaky dams to collaborate on ways to collectively prevent leaks.

Biological systems are self-maintaining, they have to be active in order to continue to live. PP and its close relative, FEF, help us understand *how* these active, self-maintaining beings manage to perform this remarkable feat. Taking the reconstructive, internalist view, is thus not incorrect in every case. But if taken as a hard logical implication of PP, then we lose the insights derived from a huge amount of fruitful work that is eminently compatible with the PP framework.

Further, the reconstructive view does not follow logically from what we know of PP. As I showed in chapter nine, the central representational notion of the system, a model, is extremely minimal and depends crucially upon the system's relationship with its environment. The 'grab bag' of 'quick and dirty' strategies for coping in specific situations is thus nicely captured by the FEF concept of a model. Coupled with the precision estimation infrastructure provided by the PP framework, these frugal strategies can be deployed efficiently in response to rapidly changing circumstances signalled by salient gist information found in the environment. Further, the components that make these strategies possible can be intelligently redeployed to tackle novel tasks.

Finally, we have grounded the notion of function in the way an SMS interacts with its environment to resist entropy – thus the very source of representation-talk's normativity relies upon the action-oriented, world-involving, biological view of process. It would make very little sense to imagine the human mind as an encapsulated module, simply reconstructing a world as an inner model, when the only thing that could give that model normative force as a representation for that system is to suppose that the system possesses the capability to detect error through a rich ongoing interaction with the world. It makes much more sense to conceive of the system as interactive and exhibiting interaction-dominant dynamics through and through, embodying a statistical model that takes the form of extremely plastic strategies for action, that may occasionally make use of inner representations in order to solve tasks, but the vast majority of the time just use the most efficient tricks to cope with the situation. On this 'action-oriented' conception error and malfunction are closely tied together– when the PP model tries a strategy and it fails, that failure will be processed as upward flowing error, to be supressed on the next attempt.

Our analysis of process and function in Part Three thus provides some extra support for Clark's radical, action-oriented predictive processing. Let us now turn one last time to the question of representation, and review how the notions we considered in chapter four fare in the light of our new understanding.

The Dynamical Stance

The paradigm I have presented in this thesis embraces a plurality of explanatory strategies. There are many mechanical elements to the systems ontology – the posited infrastructure underpinning process switching, error-detection, learning etc. are easily made sense of within the mechanistic explanatory scaffolding. However, the ontology is grounded in the dynamics of process. And while perhaps the conceptualisation is more difficult, it is by no means impossible to make sense of switching mechanisms and the like in terms of, for example, the dynamics of bistable systems.

However, in PP, we have an elegant account of the way a system might self-maintain using a parsimonious and flexible switching mechanism that relies on recognising and responding to its errors. We should not ignore this powerful tool simply because it has been developed by those working within the mechanist tradition, Moreover, the details of how actual biological systems are organised that enables them to operate *as* predictive processors have not yet been filled in. Dynamical systems theory has a great deal to contribute to this project.

The principles of soft-assembly and interaction dominance that characterise neural behaviour are well described by DST. Recognising that sub-processes may soft-assemble in order to minimise long term prediction errors can now be seen as part of the project of dynamical explanation, Further, the PP framework provides a robust guide to discovery for DST to work with. By suggesting a functional architecture, the dynamics of knowledge nets at various hierarchical levels can be investigated with their potential role as a prediction-error-minimisers in mind. For example, it might be that the strength of a particular flow of information (error) is a key variable to help explain the dynamics of a given subsystem. Without the framework provided by PP, that asymmetric flow may be overlooked, or appear mysteriously arbitrary.

All that said, the deep understanding of PP and mechanism that is afforded by FEF derives from its status as a dynamical theory. It is the informational dynamics that it describes which generate the useful implications discussed in chapters nine and ten, and that link the theory to broader scientific domains such as thermodynamics.

In sum, there are important explanatory virtues that derive from both the mechanical and dynamical paradigms. Mechanics give us deeper understanding of how and why the system works, with the promise of informing future interventions. Dynamics gives us precise predictions, and a unifying link with physics, and the other sciences. Philosophically, mechanics provides us with important links to the many fields of enquiry that are informed by philosophy of mind and invariably refer to mental states conceived as inner representations that are transformed in an orderly, computational (or pseudo-computational) way. But only dynamics gives us the philosophical basis for a belief in proper function that underlies the reality of such states as content-bearing, meaningful structures. Furthermore, paying proper attention to the virtues of dynamical explanation ensures that we never lose sight of the embodied, extended nature of the mind, which at the highest levels forms a coupled system with its ecological niche such that the two cannot be fully untangled without loss.

Predicting the Future

This thesis paints an optimistic picture of tomorrow's cognitive science. Much ink has been spilled debating the nature of thought, whether it is representational or not, and the respective virtues of dynamical systems theory and mechanistic explanation.

Whilst I expect these debates to continue (far too much has been staked for them to come to an abrupt end), I hope they will be less fraught, and more constructive. Given the framework on offer, the debates should become more focused, highlighting questions such as "what is the best way to make sense of *this* particular phenomenon?" or "Is this subsystem a representational EDS?"

Most generally, I have tried to show that even though the dynamical stance is importantly primary in the way that it constrains the theoretical space, it simultaneously justifies (by grounding a notion of proper function) and motivates, via FEF, the mechanistic study of cognition. In addition, I hope to have provided a novel angle on the continuity between life and mind. For systems capable of self-detectable error, for which PP supplies a ready-made framework, may mark an important – if not totally cut and dried - threshold in the natural world between cognitive and non-cognitive ways of being self-maintaining systems, in much the same way that SMSs themselves mark the important threshold between systems capable of actively resisting entropy and those that fall passively into thermodynamic equilibrium.

Abbreviations

CCTM	Classical Computational Theory of Mind
DST	Dynamical Systems Theory
EDS	Error Detecting System
FEF	Free Energy Formulation
FL	Flexible Learner
JDC	Job-Description Challenge
PP	Predictive Processing
RP	Reflective Person
RSMS	Recursive Self-Maintaining System
SMS	Self-Maintaining System

Bibliography

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643. http://doi.org/10.1007/s00429-012-0475-5

Aizawa, K. (1997). Explaining Systematicity. *Mind and Language*, *12*, 115–136.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *The Journal of Neuroscience*, *30*(8), 2960–6. http://doi.org/10.1523/JNEUROSCI.3730-10.2010

Allen, M., & Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 1–24. http://doi.org/10.1007/s11229-016-1288-5

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, *397*(6715), 168.

Amit, D. J. (1989). *Modeling Brain Function*. Cambridge University Press.

Anastasio, T. J., Patton, P. E., & Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, *12*(5), 1165–87. http://doi.org/10.1162/089976600300015547

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245-266.

Anderson, M. L. (2014). *After phrenology*. MIT Press.

Anderson, M. L., Richardson, M. J., & Chemero, A. (2012). Eroding the Boundaries of Cognition: Implications of Embodiment. *Topics in Cognitive Science*, *4*(4), 717–730. http://doi.org/10.1111/j.1756-8765.2012.01211.x

Armel, K. C., & Ramachandran, V. S. (2003). Projecting sensations to external objects: evidence from skin conductance response. *Proceedings. Biological Sciences / The Royal Society*, *270*(1523), 1499–506. http://doi.org/10.1098/rspb.2003.2364

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *Behavioral and brain sciences*, *9*(1), 67-90.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *The Behavioral and Brain Sciences*, *20*(4), 723-742-767. http://doi.org/10.1017/S0140525X97001611

Ballard, D. H., & Hayes, P. J. (1984). Parallel logical inference. In *7th Annual Conference of the Cognitive Science Society*.

Barth, D. S., & MacDonald, K. D. (1996). Thalamic modulation of high-frequency oscillating potentials in the auditory cortex. *Nature*, *383*, 78–81.

Barzykowski, K., & Staugaard, S. R. (2016). Does retrieval intentionality really matter? Similarities and differences between involuntary memories and directly and generatively retrieved voluntary memories. *British Journal of Psychology*, *107*(3), 519-536.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695-711.

Bechtel, W. (2014). Investigating neural representations: the tale of place cells. *Synthese*, *193*(5), 1287–1321. http://doi.org/10.1007/s11229-014-0480-8

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. *Synthese*, *193*(5), 1287-1321.

Bechtel, W., & Abrahamsen, A. (2006). Phenomena and mechanisms: Putting the symbolic, connectionist, and dynamical systems debate in broader perspective. *Contemporary debates in cognitive science. Oxford: Blackwell*.

Beer, R. (2003). The Dynamics of Active Categorical Perception in an Evolved Model Agent. *International Society for Adaptive Behavior*, *11*, 209–243.

Bickhard, M. H. (2000). Autonomy, Function, and Representation. *Communication and Cognition Artificial Intelligence*, *17*(3-4), 111-131.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*(2005), 198–211. http://doi.org/10.1016/j.jmp.2015.11.003

Bostock, E., Muller, R. U., & Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, *1*(2), 193–205.

Brentano, F. (1874). *Psychology from an Empirical Standpoint (Psychologie vom empirischen Standpunkt)*.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*(August), 1–14. http://doi.org/10.3389/fnhum.2014.00599

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. http://doi.org/10.1007/s11229-016-1239-1

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, *4*.

Büchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex (New York, N.Y. : 1991)*, *7*(8), 768–78. http://doi.org/10.1093/cercor/7.8.768

Campbell, R. (2009). A process-based model for an interactive ontology. *Synthese*, *166*(3), 453–477. http://doi.org/10.1007/s11229-008-9372-0

Chalmers, D. (1993). Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation. *Philosophical Psychology*, *6*, 305–319.

Chemero, A. (2000). Representation and "Reliable Presence." *Conceptus Studien 14: The New Computationalism*, pp. 9-25.

Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.

Cheney, D., & Seyfarth, R. (1990). *How Monkeys See the World: Inside the Mind of Another Species*. University of Chicago Press.

Clark, A. (1990). Connectionism, competence, and explanation. *The British Journal for the Philosophy of Science*, *41*(2), 195-222.

Clark, A. (1997). *Being there*. MIT Press Cambridge, MA.

Clark, A., & Grush, R. (1999). Towards a Cognitive Robotics. *Adaptive Behavior*, *7*(1), 5–16. http://doi.org/10.1177/105971239900700101

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clark, A. (2015). Radical predictive processing. *Southern Journal of Philosophy*, *53*(S1), 3–27. http://doi.org/10.1111/sjp.12120

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Clark, A. (2017). How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds. *Philosophy and Predictive Coding*, (1988), 1–31. http://doi.org/10.15502/9783958573031

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7-19.

Clark, A., & Toribio, J. (1994). Doing without representing. *Synthese*, *101*(3), 401–431.

Colgin, L. L., & Moser, E. I. (2010). Gamma oscillations in the hippocampus. *Physiology*, *25*(5), 319–329.

Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Contreras, D., Destexhe, A., Sejnowski, T. J., & Steriade, M. (1996). Control of Spatiotemporal Coherence of a Thalamic Oscillation by Corticothalamic Feedback, *274*(November), 771–774.

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*(3), 345–369. http://doi.org/10.1016/j.pneurobio.2010.06.007

Covic, E. N., & Sherman, S. M. (2011). Synaptic properties of connections between the primary and secondary auditory cortices in mice. *Cerebral Cortex*, *21*(11), 2425–2441. http://doi.org/10.1093/cercor/bhr029

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of science*, *68*(1), 53-74.

Craver, C. F. (2013). Functions and mechanisms: A perspectivalist view. In *Functions: Selection and mechanisms* (pp. 133-158). Springer Netherlands.

Crane, T. (2016). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. Routledge.

Cummins, R. (1989). *Meaning and Mental Representation*. MIT Press.

Cummins, R. (1996). *Representations, Targets, and Attitudes*. MIT Press.

Dahlbom, B. (1993). Editors Introduction. In B. Dahlbom (Ed.), *Dennett and his critics*. Blackwell.

Dan, Y., Atick, J. J., & Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(10), 3351–3362.

Davidson, D. (1987). Knowing one's own mind. In *Proceedings and addresses of the American philosophical association* (Vol. 60, pp. 441–458).

De Pasquale, R., & Sherman, S. M. (2011). Synaptic Properties of Corticocortical Connections between the Primary and Secondary Visual Cortical Areas in the Mouse. *Journal of Neuroscience*, *31*(46), 16494–16506. http://doi.org/10.1523/JNEUROSCI.3664-11.2011

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, *191*(15), 3639–3648. http://doi.org/10.1007/s11229-014-0484-4

Dennett, D. C. (1969) Content and Consciousness. Routledge.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. (2009). Two Black Boxes: A Fable. *Activitas Nervosa Superior*, *52*, 81–84.

Dong, D., & Atick, J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, *6*(2), 159–178. http://doi.org/10.1088/0954-898X/6/2/003

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, Content, and Function* (pp. 17–36). Oxford University Press.

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253–259. http://doi.org/10.1016/j.shpsa.2010.07.009

Eliasmith, C. (1996). The Third Contender: A critical examination of the Dynamicist theory of cognition. *Philosophical Psychology*, *9*(4), 441–463.

Engel, A. K., Maye, A., Kurthen, M., & King, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, *17*(5), 202–209. http://doi.org/10.1016/j.tics.2013.03.006

Feldman, A., & Levin, M. (2009). The Equilibrium-Point Hypothesis – Past, Present and Future. *Advances in Experimental Medicine and Biology*, *629*.

Feldman, H., & Friston, K. J. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*(December), 1–23. http://doi.org/10.3389/fnhum.2010.00215

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *1*(1), 1–47. http://doi.org/10.1093/cercor/1.1.1

Feyerabend, P. (1975). *Against method*. Verso.

FitzGerald, T. H. B., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, *8*(June), 1–11. http://doi.org/10.3389/fnhum.2014.00457

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. http://doi.org/10.1038/nrn2536

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work. *Cognition*, *35*, 305–319.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture. *Cognition*, *28*(1–2), 3–71.

Fresco, N. (2010). Explaining Computation Without Semantics: Keeping It Simple. *Minds and Machines*, *20*, 165–181.

Fresco, N. (2014). *Physical computation and cognitive science*. New York: Springer.

Frisch, K. (1967). *The dance language and orientation of bees*. Harvard University Press.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. http://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1456), 815-836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, *13*(7), 293-301.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, (April), 150217111908007. http://doi.org/10.1080/17588928.2015.1020053

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417-458.

Gallagher, S. (2008). Are Minimal Representations Still Representations? *International Journal of Philosophical Studies*, *16*(3), 351-369.

Garson, J. (2011). Selected effects and causal role functions in the brain: The case for an etiological approach to neuroscience. *Biology and Philosophy*, *26*(4), 547–565. http://doi.org/10.1007/s10539-011-9262-6

Garson, J. (2012). Function, selection, and construction in the brain. *Synthese*, *189*(3), 451–481. http://doi.org/10.1007/s11229-012-0122-y

Girard, P., Salin, P. A., & Bullier, J. (1991). Visual activity in areas V3a and V3 during reversible inactivation of area V1 in the macaque monkey. *Journal of Neurophysiology*, *66*(5), 1493–1503. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=1765790&retmode=ref&cmd=prlinks%5Cnpapers2://publication/uuid/8C35859A-93AD-4F38-9A34-70B2F9C3C89D

Gladziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, *40*(53), 63–90. http://doi.org/10.1515/slgr-2015-0004

Godfrey-Smith, P. (2007). Information in Biology. In D. Hull & M. Ruse (Eds.),*.),* *The Cambridge Companion to the Philosophy of Biology* (pp. 103–119). Cambridge University Press.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16. http://doi.org/10.1016/S1364-6613(00)01567-9

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, *3*(1), 71-88.

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. Hackett publishing.

Gould, S. J. (1980). *The panda's thumb*. WW Norton & company.

Grice, P. (1957). Meaning. *The Philosophical Review*, *66*, 377–388.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition, *17*(9), 767–773.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian Models of Cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology*. Cambridge University Press.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3). http://doi.org/10.1017/S0140525X04000093

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Han, B., & VanRullen, R. (2017). The rhythms of predictive coding? Pre-stimulus phase modulates the influence of shape perception on luminance judgments. *Scientific Reports*, *7*.

Harrison, L. M., Stephan, K. E., Rees, G., & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage*, *34*(3), 1199–1208. http://doi.org/10.1016/j.neuroimage.2006.10.017

Hatfield, G. (2002). Perception as Unconscious Inference. In D. Heyer & R. Mausfield (Eds.), *Perception and the Physical World: Psychological and Philosophical Issues in Perception* (pp. 115–143). New York: Wiley.

Haugeland, J. (1981). Semantic engines: An introduction to mind design. In Haugeland, J. *Mind Design*. MIT Press.

Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, *4*(1990), 383–427. http://doi.org/10.2307/2214199

Helmholtz, H. . (n.d.). Handbuch der physiologischen Optik. In A. Gullstrand, J. von Kries, & W. Nagel (Eds.), *Handbuch der physiologischen Optik.*

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in Brain Research*, *165*, 535–547. http://doi.org/10.1016/S0079-6123(06)65034-6

Hinton, G. E. (2009). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.

Hinton, G. E., Mcclelland, J. L., & Rumelhart, D. E. (1986). Distributed representations, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. MIT Press

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98*(1), 82–98. http://doi.org/10.1016/j.pneurobio.2012.05.003

Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP.

Hohwy, J. (2014). The self-evidencing brain. *Nous*, *0*, 1–27. http://doi.org/10.1111/nous.12062

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580-593.

Hutto, D., & Myin, E. (2013). Radical enactivism: Basic minds without content. Cambridge, MA: MIT UP.

Jansen, P., & Watter, S. (2012). Strong Systematicity Through Sensorimotor Conceptual Grounding: an Unsupervised, Developmental Approach to Connectionist Sentence Processing. *Connection Science*, *24*(1), 25–55.

Kanai, R., Komura, Y., Shipp, S., Friston, K., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*, 20140169. http://doi.org/10.1098/rstb.2014.0169

Kelso, S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behaviour*. MIT Press.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation, *27*(12). http://doi.org/10.1016/j.tins.2004.10.007

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–55. http://doi.org/10.1038/nn.3393

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and cognition*, *65*(2), 145-168.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, *13*(1), 47–76. http://doi.org/10.1080/09515080050002726

Laird, J. E. (2008). Extending the Soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications*, *171*, 224.

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT Press.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes. In J. Worrall & G. Currie (Eds.), *Philosophical Papers*. Cambridge University Press.

Lamb, M., & Chemero, A. (2014). Structure and Application of Dynamical Models in Cognitive Science. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 809–814.

Lancaster, M. E., & Homa, D. (2017). Feature-to-Feature Inference Under Conditions of Cue Restriction and Dimensional Correlation. *American Journal of Psychology*, *130*(1), 35–45.

Lungarella, M., & Sporns, O. (2005). Information self-structuring: Key principle for learning and development. *Proceedings of 2005 4th IEEE International Conference on Development and Learning*, *2005*, 25–30. http://doi.org/10.1109/DEVLRN.2005.1490938

Lupyan, G. (2015). Cognitive Penetrability of Perception in the Age of Prediction: Predictive Systems are Penetrable Systems. *Review of Philosophy and Psychology*, *6*(4), 547–569. http://doi.org/10.1007/s13164-015-0253-4

Macdonald, C. (1995). Classicism V. Connectionism. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation* (pp. 3–27). Blackwell.

Markus, E. J., Qin, Y. L., Leonard, B., Skaggs, W. E., McNaughton, B. L., & Barnes, C. A. (1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience*, *15*(11), 7079-7094.

Marr, D. (1983). *Vision*. MIT Press.

Matheson, H., & Barsalou, L. (2017). Embodiment and grounding in cognitive neuroscience. In *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

Milkowski, M. (2013). *Explaining the Computational Mind.* MIT Press.

Miller, B. (2016). What is Hacking's Argument for Entity Realism? *Synthese*, *193*(3), 991–1006.

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. MIT Press.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, *86*, 281–297.

Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, *9*, 185–200.

Millikan, R. G. (2004). *Varieties of Meaning*. MIT Press.

Morgan, A. (2014). Representations gone mental. *Synthese*, *191*(2), 213–244. http://doi.org/10.1007/s11229-013-0328-7

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15164–9. http://doi.org/10.1073/pnas.192579399

Murray, S. O., Olman, C. a, Kersten, D., Scott, O., & Spatially, D. K. (2006). Spatially Specific fMRI Repetition Effects in Human Visual Cortex. *Journal of Neurophysiology*, *95*, 2439–2445. http://doi.org/10.1152/jn.01236.2005.

Neander, K. (1991a). The teleological notion of "function." *Australasian Journal of Philosophy*, *69*(4), 454–468. http://doi.org/10.1080/00048409112344881

Neander, K. (1991b). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168-184.

Neander, K. (1999). Fitness and the fate of unicorns. In Valerie Gray Hardcastle (ed.), *Where Biology Meets Psychology*. MIT Press.

Newell, A. (1992). Unified theories of cognition and the role of Soar. In *SOAR: A cognitive architecture in perspective* (pp. 25–79). Springer.

Noë, A. (2004). *Action in Perception*. MIT Press.

Office for National Statistics (2017). National life tables United Kingdom reference tables. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental neurology*, *51*(1), 78-109.

O'Keefe, J. (1979). A review of the hippocampal place cells. *Progress in Neurobiology*, *13*(4), 419–439.

O'Keefe, J., & Conway, D. H. (1978). Hippocampal place units in the freely moving rat: why they fire where they fire. *Experimental Brain Research*, *31*(4), 573–590.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

O'Keefe, J., & Speakman, A. (1987). Single unit activity in the rat hippocampus during a spatial memory task. *Experimental Brain Research*, *68*(1), 1–27.

Parkin, T. G. (1992). Demography and Roman society.

Parrott, M. (2017). Subjective Misidentification and Thought Insertion. *Mind and Language*, *32*(1), 39–64. http://doi.org/10.1111/mila.12132

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35. http://doi.org/10.1016/j.pneurobio.2015.09.001

Piaget, J. (1932). *The moral judgement of the child*. Routledge.

Piaget, J. (1936). *Origins of intelligence in the child*. Routledge.

Piaget, J. (1945). *Play, dreams, and imitation in childhood*. Heinemann.

Piaget, J. (1957). *Construction of Reality in the child*. Routledge.

Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. *AMC*, *10*, 12.

Piccinini, G. (2007). Computing Mechanisms. *Philosophy of Science*, *74*(4), 501–526.

Piccinini, G. (2008). Computation without representation. *Philosophical Studies*, *137*(2), 205–241. http://doi.org/10.1007/s11098-005-5385-4

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. OUP.

Piccinini, G. (2017). Computation in Physical Systems. *The Stanford Encyclopaedia of Philosophy*. (E. Zalta, Ed.)

Piccinini, G., & Bahar, S. (2013). Neural Computation and the Computational Theory of Cognition. *Cognitive Science*, *37*(3), 453–488. http://doi.org/10.1111/cogs.12012

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. http://doi.org/10.1007/s11229-011-9898-4

Peirce, C. (1902). Logic as Semiotic: The Theory of Signs. In C. S. Peirce (Ed.), *Philosophical Writings*. Dover Publications.

Peirce, C. (1958). *The Collected Papers Volumes 7 and 8*. (A. Banks, Ed.). Harvard University Press.

Peirce, C. (1977). *Semiotics and Significs*. (C. Hardwick, Ed.). Indiana University Press.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. http://doi.org/10.1038/22268

Popper, K. R. (1963). *Conjectures and refutations: the growth of scientific knowledge*.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception & Psychophysics*, *76*(2), 270–9. http://doi.org/10.3758/s13414-013-0605-z

Putnam, H. (1975). What is mathematical truth? In *Mathematics, matter, and method*. Cambridge University Press.

Quirk, G. J., Muller, R. U., & Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *Journal of Neuroscience*, *10*(6), 2008–2017.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Ramsey, W., Stich, S., & Garon, J. (2010). Connectionism, and the Future of Folk Psychology. *Mind*, *4*(1990), 499–533.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects, *2*, 79–87.

Reddy, Leila; Tsuchiya, N., & Serre, T. (2011). Mental Imagery, *50*(2), 818–825. http://doi.org/10.1016/j.neuroimage.2009.11.084.Reading

Robinson, D. A. (1968). A note on the oculomotor pathway. *Experimental Neurology*, *22*, 130–132.

Robinson, D. A. (1989). Integrating with neurons. *Annual Review of Neuroscience*, *12*, 33–45.

Rosenbloom, P. S., Laird, J. E., Newell, A., & McCarl, R. (1991). A preliminary analysis of the Soar architecture as a basis for general intelligence. *Artificial Intelligence*, *47*, 289–325.

Ross, H. E. (1969). When is a weight not illusory? *The Quarterly Journal of Experimental Psychology*, *21*(4), 346–355.

Rumelhart, D. E. (1989). The Architecture of Mind: A Connectionist Approach. *The Foundations of Cognitive Science*, (1), 133–159. http://doi.org/10.1007/s13398-014-0173-7.2

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Parallel Distributed Processing*. MIT Press.

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., & Kastner, S. (2012). The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. *Science*, *337*(August), 753–757. http://doi.org/10.1126/science.1223082

Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiological Reviews*, *75*(1), 107–154.

Scarantino, A., & Piccinini, G. (2010). Information without truth. *Metaphilosophy*, *41*(3), 313–330.

Sejnowski, T., & Rosenberg, C. (1986). NETtalk: A parallel network that learns to read aloud. *The Johns Hopkins University Electrical Engineering and Computer Science Technical Report*, *1*, 663–672. Retrieved from http://dl.acm.org/citation.cfm?id=104448

Seth, A. K. (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.

Seung, H. S. (1998). Continuous Attractors and Oculomotor Control. *Neural Networks*, *11*, 1253–1258.

Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, *93*(23), 13339–13344. http://doi.org/10.1073/pnas.93.23.13339

Shagrir, O. (2012). Structural representations and the brain. *British Journal for the Philosophy of Science*, *63*(3), 519–545. http://doi.org/10.1093/bjps/axr038

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind and Language*, *22*(3), 246–269. http://doi.org/10.1111/j.1468-0017.2007.00308.x

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences*, *8*(5), 223–230. http://doi.org/10.1016/j.tics.2004.03.004

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. http://doi.org/10.1038/nature16961

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in cognitive sciences*, *7*(8), 343-348.

Smolensky, P. (1987). *On variable binding and the representation of symbolic structures in connectionist systems*.

Smolensky, P. (1988). Putting together connectionism - again. *Behavioral and Brain Sciences*, *35*, 183–204.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. http://doi.org/10.1016/0004-3702(90)90007-M

Smolensky, P. (1995a). Connectionism, Constituency, and the Language of Thought. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation*.

Smolensky, P. (1995b). Integrated connectionist/symbolic architecture. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation*.

Sprevak, M. (2011). William M. Ramsey Representation Reconsidered. *The British Journal for the Philosophy of Science*.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, *96*(4), 539-560.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. MIT Press.

Stich, S. P., & Warfield, T. A. (1995). Reply to Clark and Smolensky: Do connectionist minds have beliefs?

Summerfield, C., Wyart, V., Johnen, V. M., & de Gardelle, V. (2011). Human Scalp Electroencephalography Reveals that Repetition Suppression Varies with Expectation. *Frontiers in Human Neuroscience*, *5*(July), 67. http://doi.org/10.3389/fnhum.2011.00067

Thompson, E., & Varela, F. (2001). Radical Embodiment: neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5*(10), 418–425.

Tolman, E. C., & others. (1948). Cognitive maps in rats and men. American Psychological Association.

Turing, A. (1936-7.). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceeding of the London Mathematical Society*, *42*(1), 230–265.

Van Essen, D., & Maunsell, J. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*, *6*(9), 370–5.

van Gelder, T., & Port, R. (1995). It's about time: an overview of the dynamical approach to cognition. In *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge: MIT Press.

Varela, F., Thompson, & Rosch, E. (1991). E. (1991):" The Embodied Mind. *Cognitive Science and Human Experience", MIT Press, Cambridge. Natural Interaction, Article on Museum Practise: Http://naturalinteraction. Org/index. Php Entry= entry070224-123657*.

von Eckardt, B. (1993). *What is Cognitive Science*. MIT Press.

von Uexküll, J. (1934). A stroll through the worlds of animals and men (trans: Claire H. Schiller). *Instinctive behavior, ed. Claire H. Schiller*, 5-80.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. http://doi.org/10.1038/nn858

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. http://doi.org/10.1016/j.jpowsour.2014.09.131

Wheeler, M. (2005). Friends reunited? Evolutionary robotics and representational explanation. *Artificial Life*, *11*(1–2), 215–231.

Winkler, A. D., Spillmann, L., Werner, J. S., & Webster, M. A. (2015). Asymmetries in blue-yellow color perception and in the color of "the dress." *Current Biology*, (D), 2–3. http://doi.org/10.1016/j.cub.2015.05.004

Yokota, J. I., Reisine, H., & Cohen, B. (1992). Nystagmus induced by electrical stimulation of the vestibular and prepositus hypoglossi nuclei in the monkey: evidence for site of induction of velocity storage. *Experimental Brain Research*, *92*(1), 123–138. http://doi.org/10.1007/BF00230389

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301-308.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, *12*(10), 931–934.

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.