The Direction of Fit of Desire

Patrick Mark Butlin

Ph.D. Thesis in Philosophy

King's College, London

Abstract

It is a familiar tenet that desires and beliefs have opposite directions of fit. Our beliefs, according to this view, should be changed to fit the world – if necessary – because they are for saying how things are. Our desires give us reasons to change the world, because they are for saying what to do, or how things should be. I argue that like beliefs, desires have only the mind-to-world direction of fit.

In arguing for this conclusion, I present new accounts of both desire and direction of fit. Desires are inputs to the goal-directed system – a system for behavioural control studied in psychology and neuroscience – with the function of tracking the reward values of outcomes. In the goal-directed system these states are combined with further states representing contingencies between actions and outcomes, in order to select the actions which offer greatest reward. According to this account, desires come in occurrent and standing forms, are likely to have a wide range of outcomes as their objects, and interact with habits, emotions and intentions in familiar ways.

My account of direction of fit uses a teleosemantic framework. Teleosemantics is a family of theories of representation that aim to identify the characteristic functions of representations and the systems in which they operate, and focus on representation as a biological phenomenon. It is particularly suited to thinking about direction of fit, because representations have their directions of fit in virtue of what they are for – that is, their functions. I claim that representations have the mind-to-world direction of fit when the systems that produce them have the function of doing so under specific conditions, and the world-to-mind direction of fit when the systems that consume them have the function of behaving in specific ways, whenever the representations occur. Desires do not have the world-to-mind direction of fit, because what the goal-directed system should do when any given desire is occurrent also depends on what other desires are occurrent at the time, and on the agent's beliefs.

It does not follow that we have no reason to try to make the world fit our desires; instead, this conclusion shows that the place of desires in rational motivation is less closely tied to their properties as representations than some philosophers have thought.

Contents

Acknowledgments

My supervisors for this project were David Papineau and Nick Shea, and they have both made great contributions to the production of this thesis. I would like to thank them both for the long hours of work they have put into helping me to develop my ideas and refine my many drafts, and for the unfailing patience and enthusiasm with which they have done so.

I would also like to thank Clayton Littlejohn, Richard Holton and Maria Alvarez, who read my work, discussed my ideas, and offered very valuable advice at various stages of the project; my examiners Stephen Laurence and Stephen Butterfill, for their careful reading and very encouraging remarks; and the staff and students of the Philosophy Department at King’s College, London for the many ways in which they have supported me.

Finally, I am very glad to acknowledge the immeasurable contribution of my parents, Roger and Alison Butlin, and my wife Susannah, for their support and encouragement.

Introduction

Our desires are among the most powerful influences on what we do and what we care about. Desires control our lives from moment to moment – at least in those moments when we have the opportunity to make choices – but they also shape our projects, careers, homes and relationships. We cannot hope to understand human motivation without understanding desire.

Meanwhile, one of the central questions of philosophy is how some objects and events can represent, or be about, other things. In particular, the nature of representation is a foundational problem in philosophy of mind, since we continually appeal to the representational properties of mental states when giving everyday explanations of our behaviour and experiences, and when giving philosophical accounts of conscious experience, perception, motivation, and much else besides.

This thesis aims to say what kind of representation is involved in desire. More specifically, the question I address is: what is the *direction of fit* of desire? My answer is that desires have only the mind-to-world direction of fit. In this introduction, I give initial characterisations of desire and direction of fit, explain why I think this is a good question, and outline my strategy for answering it.

Both 'desire' and 'direction of fit' are technical terms in philosophy, although there are no widely-agreed definitions for them. Desires are mental states that interact with beliefs in motivating us to act. Paradigmatically, we are motivated to act when we believe that doing so is likely to lead to or promote the satisfaction of one or more of our desires. To fill out that idea a bit more, 'satisfying a desire' is taken to mean that the agent gets what they desire, not that they feel satisfaction on getting it. The objects of desire are states of affairs, often called 'outcomes', and desiring some outcome means wanting it to be the case. So when we are motivated by our desires, we paradigmatically believe that by acting we can make it more likely that one or more of the states of affairs that we desire will be the case.

Philosophers typically think of each of us as having many desires, for a wide range of different things – more or less all of those things we would ordinarily be said to want. For instance, I presently desire to drink some water, and to climb the *Biancograt* on Piz Bernina in the Swiss Alps, and for the civil war in Syria to end soon. Outside philosophy, it is common to attribute to people desires for things which are not states of affairs – we might say that I desire water, or that I desire the summit. But an assumption of this thesis will be that strictly speaking there are no such desires (for an argument for this view, see Sinhababu 2015). To say that I desire water is either false, or means that I desire to drink or to have some water.

As well as explaining motivation and action, desires can also explain why we feel pleased or disappointed, and why some thoughts and objects capture our attention. For example, I am disposed to be pleased when I hear of success for my old rowing club, and to wonder sometimes about how they are doing, because I desire that they succeed. These points suggest that we have *standing* desires, which persist over long periods, because I am disposed to be pleased when I hear that my old club has won even if they are far from my thoughts at the time; but it also seems that for relatively short periods our desires can become *occurrent*. This would explain why I am sometimes very strongly motivated to eat chocolate ice-cream, and yet most of the time I make no effort to get it. The idea would be that my desire for chocolate ice-cream is a long-standing feature of my personality, but that it only motivates me when something about my circumstances causes it to temporarily take a different form, by becoming occurrent.

Moving on to the subject of direction of fit, it is intuitive that two important categories of representation are those that aim to say how things are, and those that tell some consumer of the representation what to do. Many representations can be characterised by saying which of these two kinds they belong to, and giving the proposition or state of affairs which they say is the case or is to be brought about. Some examples will help to illustrate the point. The assertions 'Grass is green' and 'Snow is white' both aim to say how things are, but obviously differ in *how* they say things are. In contrast, the assertion 'The door is shut' and the command 'Shut the door!' are of different kinds, but both stand in special representational relationships to the state of affairs of the door's being shut; the assertion says that this is the case, while the command instructs an intended hearer to make it the case. These two sentences have the same *content*, but different *directions of fit*.

Direction of fit exists beyond language, however. For example, two identical scale drawings may have different directions of fit if one is intended to show what an existing house looks like, while the other is a plan to be followed in building a new house. Sometimes one representation has both directions of fit; a single drawing could be at the same time an illustration of an existing house, and a specification for a new one.

Why is this property of representations called ‘direction of fit’? First, note that when two things are supposed to fit one another, sometimes the ‘responsibility’ for achieving the fit lies solely or primarily with one of the two. We often look for shoes that fit our feet, and when doing this we take a failure to fit to be a fault in the shoe. But in the *Cinderella* story, the Prince looks for a foot to fit the lost shoe. In both cases, success would be feet and shoes fitting one another, but they differ in whether the shoe or the foot should be changed to achieve this. Returning to representation, both of the two scale drawings – the survey and the plan – are supposed to ‘fit’ the world. They both succeed if the house concerned is (or ends up) the way they show. But the *direction* of fit is different, because if the survey and the real house do not fit one another then the survey should be changed, but if the plan and the house do not fit then the house should be changed. In the terminology which is most common in philosophy, the survey has the *mind-to-world* direction of fit, because the representation (in this case not a mental one) is supposed to fit the world, while the plan has the *world-to-mind* direction of fit.

In this context, a common thought is that beliefs have the mind-to-world direction of fit, while desires have the world-to-mind direction of fit. On that view, my desire to eat ice-cream would be a representation with the world-to-mind direction of fit with respect to the state of affairs that *I am eating ice-cream*. My desire would tell me, or some part of me, to make it the case that I am eating ice-cream. The conclusion that I am going to argue for is that this is wrong. On my view, desires do not have the world-to-mind direction of fit, but do have the mind-to-world direction of fit. To this extent, desires are like beliefs or assertions. My desire to eat ice-cream has the mind-to-world direction of fit with respect to a state of affairs of roughly the form *My eating ice-cream is good for me to degree x*.

One reason to doubt that desires have only the world-to-mind direction of fit, and hence one motivation for my project, is that it matters to our well-being not only that our desires are satisfied, but also that we desire the right things. Other things being equal, a person’s life will go much better if they have a strong desire to eat fresh fruit and little desire to smoke crack cocaine than vice versa. From a biological point of view, it would be strange if the job of desires was anything other than to keep track of, and to direct our motivation towards, outcomes that it would be biologically beneficial for us to bring about. In particular, it would be bizarre if we had an unconstrained capacity to generate desires spontaneously, which then played a major role in determining how we act. These points suggest that norms of some kind apply to the circumstances in which particular desires should be produced, as well as to what we should do given our desires. In turn, this suggests that desires may have the mind-to-world direction of fit.

In addition to this, having any given desire at a time does not determine what we subjectively ought to do, because this also depends on our beliefs about the actions available to us and their likely consequences, and on the other desires we have at the time. So it is not obvious that desires do tell us what to do; this point suggests they may lack the world-to-mind direction of fit. These arguments are nothing like sufficient to support my conclusion, but they do suggest that working out the direction of fit of desire is a difficult enough issue to be worthy of detailed investigation.

Detailed investigation of the direction of fit of desire is also worthwhile because direction of fit itself is a somewhat neglected subject. Given that many representations, apparently including beliefs and desires, assertions and commands, may be characterised by their direction of fit and their content, direction of fit might be a natural target for philosophers of mind working towards naturalistic general theories of representation. But in fact the leading philosophers working on this project in the 1980s and 1990s were mostly preoccupied with the problem of content, and to some extent with distinguishing representations from non-representations. Ruth Millikan's teleosemantics (1984, 2004) is an exception to this, in that she gives explicit conditions for what she calls *indicative* and *imperative* content, and in my view she succeeded in developing a very attractive framework for theorising about direction of fit. However, it has been alleged that Millikan's account implies that all representations have both directions of fit (Artiga 2013). The topic of direction of fit is more familiar from meta-ethics, since the directions of fit of desire and belief are appealed to in a famous argument by Michael Smith (1987). But the discussion in this area has largely been confined to desire and belief, despite the evidence that many other kinds of representations also have directions of fit.

Although my ultimate aim is to determine the direction of fit of desire, I will defend a more general account of direction of fit than those offered in the meta-ethical context.

Smith's argument also shows one way in which the direction of fit of desire matters. Smith relies on the premise that desire and belief have different directions of fit in arguing for the Humean Theory of Motivation, which is the claim that being motivated by reasons requires the presence of a desire that one believes the action in question will help to satisfy. The Humean Theory of Motivation implies that beliefs alone are not capable of motivating actions taken for reasons, even if they include beliefs about what would be right, or what one has most reason to do. This is a remarkable conclusion in its own right, but it further implies that if moral judgments (or other normative judgments) are intrinsically motivating, then those judgments cannot be beliefs. In turn, this conclusion is thought to undermine moral realism. So if Smith's argument is the best available for the Humean Theory of Motivation (as Shafer-Landau 2003 claims), then the potential consequences of my view extend to the most fundamental meta-ethical and meta-normative issues.

Finally, my topic is timely because a substantial body of empirical evidence concerning the mechanisms of action-selection is now available, and there is sufficient scientific consensus to allow initial conclusions to be drawn about desires and the processes surrounding them. Empirical discoveries about how the human mind actually works can often throw new light on topics of philosophical interest, such as consciousness, perception, action, motivation, rationality, and mental representation. Further to this, the prospects of many philosophical theories about such phenomena seem to be contingent on facts that can only be adequately confirmed or denied by science, because we assume that these phenomena exist in humans. For instance, a theory proposing that consciousness requires a certain functional process will be plausible only if the brain (or perhaps the body as a whole) performs that process. Similar things could be said about theories of both practical and theoretical rational inference, although we are prepared to accept that humans fall some way short of ideal rationality. So a promising methodology for philosophy of mind in cases in which the relevant science is available and some philosophical theorising has already been attempted is to investigate how well the science and the philosophy fit with one another, and adjust the philosophy (and conceivably argue for new science) accordingly.

Many philosophers studying consciousness and perception have been following something like this method for some time, and it has certainly been advocated in other fields, such as by the naturalized epistemology movement. But in the study of action and motivation it is relatively new. Tim Schroeder's work to develop a neuroscientifically-informed account of desire, and to draw out its philosophical consequences (T. Schroeder 2004, Arpaly & Schroeder 2014), seems to me to be extremely important for this reason, and has been an inspiration for this thesis. The neuroscience of desire is also becoming known among philosophers studying addiction (see e.g. the papers in Levy 2014). There is a valuable opportunity at present for more philosophers to learn about this science and to debate its implications.

In order to reach my conclusion, which is that desires have only the mind-to-world direction of fit, I will argue for the following five premises:

I. Desires are *outcome values*.
II. The *goal-directed control system* works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.
III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.
IV. *Biological representations* with *consumers that have discretion* have only the mind-to-world direction of fit.
V. It follows from I-III that desires are biological representations with consumers that have discretion.

I will argue for premises I-III in part I of this thesis, which focuses on what desires are and how they contribute to motivation and action. Then in part II, which focuses on direction of fit, I will argue for premises IV and V. Parts I and II make up the bulk of the thesis, but since its relationship to the Humean Theory of Motivation is one of the main reasons why the direction of fit of desire is of interest, I also discuss the implications of my argument and conclusion for this theory. This is the topic of Part III.

Part I is made up of chapters 1-5, and my case for premise I spans all five of these chapters. This is because I aim to show what desires are by identifying a natural kind of psychological state that has many of the most important properties that are commonly associated with desire. States belonging to this natural kind are sometimes called ‘outcome values’. So almost all of my discussion of what desires are and what they do contributes to the case for premise I. In these five chapters, I take on the following tasks:

- In chapter 1, I explain what I mean by ‘desire’, and state and briefly defend some assumptions. I also give a more detailed outline of part I.
- In chapter 2, I introduce the *goal-directed* and *habitual* control systems, and give the empirical case for their existence and distinctness from one another. So premise II is established primarily by the work of this chapter. I also present some other relevant neuroscience.
- In chapter 3, I discuss in detail what outcome values (i.e. desires) are, how they contribute to the goal-directed system, and how they are formed and modified. Premise III is established primarily in this chapter.
- In chapter 4, I consider how the goal-directed and habitual systems interact with each other and with other systems that may contribute to action-selection, and wrap up my positive case for premises I-III.
- Finally, in chapter 5 I present arguments against Schroeder’s (2004) theory of desire. This theory is an important rival to my view, since Schroeder draws on a similar body of empirical evidence, but reaches a different conclusion to me.

In part II I argue for premises IV and V. However, a considerable amount of preliminary work is necessary beforehand. My theory of direction of fit (of which premise IV is a partial statement) uses a conceptual framework drawn from teleosemantics, which it is necessary for me to explain before I can present and defend my own view. Also, teleosemantics has been the subject of several well-known objections, some of which would be seriously damaging to my theory if they succeeded, so I take the opportunity to address these objections. Part II is therefore structured as follows:

- In chapter 6, I introduce the topic of direction of fit and describe some advantages of the teleosemantic approach to this topic. I also give a more detailed outline of part II.
- In chapter 7, I outline a version of teleosemantics, introducing several important technical terms, and defend it against a range of objections which are relevant to my project.
- In chapter 8, I argue for a new theory of direction of fit, called the Discretion View. Since premise IV is a partial statement of this theory, the argument for premise IV comes in this chapter.
- Finally, in chapter 9 I apply my theory of direction of fit to desires, drawing on the work of part I. This yields premise V, and therefore my conclusion. I also discuss the nature of reward, and summarise my argument of chapters 1-9.

Part III contains only a single chapter, chapter 10, which is concerned with the Humean Theory of Motivation. One way to argue against my conclusion might be to claim that it is inconsistent with some attractive aspect of Humeanism, so part of the purpose of this chapter is to respond to this possible line of objection. However, although I respond to several possible objections in the course of my argument, my main aim in this thesis is to present the positive case for my view.

Part I: Desire

Chapter 1: Desire as a Natural Kind

1.1 Introduction to Part I

My aim in this and the following four chapters is to develop an account of desire as a natural kind, and to describe some of the psychological processes in which desires are involved. These include, most importantly, some of the major processes that contribute to determining how we act. This account will vindicate premises I-III of my overall argument:

> I. Desires are outcome values.
>
> II. The goal-directed control system works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.
>
> III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.

Premises II and III will be supported relatively directly by the empirical evidence I will present, and premise I will follow from my account, given the plausible assumption that if there is some natural kind that does enough of what desires are commonly thought to do, then what it is to be a desire is to be a member of this natural kind.

My strategy will be to present a more wide-ranging and more detailed account of the goal-directed system and related systems than is embodied in premises II and III. I take the best reason to believe these two premises to be that they make important contributions to an attractive overall picture of action-selection, which is at the centre of a flourishing research programme. It will not be possible for me to give an exhaustive survey of this research, but I hope to show how the main elements fit together to create a compelling theory. The breadth and depth of my discussion will also contribute to the case for premise I, by showing the range of ways in which outcome values fit the functional profile that philosophers typically associate with desire.

In chapter 2, I will present some background psychology and neuroscience which is necessary for understanding the neuroscience of desire. There are two main topics to be addressed. First, there is evidence from behavioural psychology suggesting that humans and many other mammals use two systems for action control, called the *habitual* and *goal-directed* systems. I will describe the experiments that distinguish these two systems, and some neuroscientific results suggesting that they are anatomically and functionally distinct. This part of the chapter will directly support premise II. Second, the basal ganglia are a group of brain structures contained in the cerebral hemispheres, which are centrally involved in action-selection and the learning processes that affect it. They are also the primary site of action of dopamine, a neurotransmitter which is often thought to have an intimate connection with desire. So I will outline the basic anatomy of the basal ganglia and explain how it can subserve action selection, and give an initial account of dopamine's role.

Chapter 3 focuses on the goal-directed system, and gives more detail about some of the topics introduced in chapter 2. These include the function of dopamine, and the relationship between desires and *basic drives* – drives for food and water, for instance. I also introduce the crucial distinction between *standing* and *occurrent* desires, and describe the different roles of these two kinds of states. The overall goal of chapter 3 is to explain how the goal-directed system works, in enough detail to make it plausible that there is such a system working in that way. So this chapter will also contribute to premise II, and will form the main part of my case for premise III.

In chapter 4, I take a broader view of action-selection, asking how the habitual and goal-directed systems might co-operate or compete with one another, and what other systems there are that contribute to determining how we act. I also review premises I-III, and describe how each is supported by the work of chapters 1-4.

Finally, in chapter 5 I address an important concern about my account of desire. Tim Schroeder (2004, Arpaly & Schroeder 2014) has used a similar body of scientific theory and evidence to me in developing his theory of desire, and like me he aimed to find a natural kind to identify as desire. Yet Schroeder's theory is not the same as mine – he not only gives a characterisation of desire that I dispute, but the theories are not even co-extensive. So I explain why my account is preferable.

In this chapter, my concern is with philosophical preliminaries to the project of this part of the thesis. I first describe what I mean by 'desire', then state and defend some assumptions, and explain why the lengthy discussion of empirical questions that forms much of the rest of part I is worthwhile in the pursuit of philosophical aims.

1.2 Clarifying 'Desire'

The word 'desire' is understood in a number of different ways both by philosophers, and in ordinary English. It is also relatively rare outside philosophy for the word 'desire' to be used in the attribution of intentional attitudes; we more commonly use verbs such as 'want', 'like', 'hope', 'wish' or 'prefer', instead.[1] But 'want', by far the most common of these, is itself either polysemous or highly indeterminate in meaning. So in this section I explain the sense in which I use the term 'desire', and thus specify my topic more precisely than I have done so far.

In the sense in which I use the word 'desire', it refers to a particular kind of psychological state that motivates us to act in combination with beliefs about our actions, and also refers to the members of that kind. Desires in this sense are also thought to affect how we feel about events and states of affairs that happen to us or which we learn about; we are thought to experience pleasure when things turn out as we desire, in this sense of the term. In this sense, most of us are thought to have a very wide range of desires, and philosophers often think of these desires as having a distinctive role with respect to practical rationality. On one hand, it is typically thought to be rational to act on one's desires, providing one does so in the right way, and certain conditions are met – and our having desires in this sense and being able to act on them is part of what makes us rational creatures. On the other hand, it is

[1] In the 450m-word Corpus of Contemporary American English, 'desire' appears much less frequently as a verb than any of these five – around three times less frequently than 'prefer', which is in turn much less common than the other four. 'Desire' is more common as a noun, however – only a little less so than 'belief'.

also thought that the extent to which desires can be rationally criticised or justified is quite limited.

This class of psychological states can be usefully compared with a range of other, related phenomena.

First, philosophers sometimes distinguish between *intrinsic* and *instrumental* desires. According to the usual way of drawing the distinction, the things we intrinsically desire are those that we want for their own sake, whereas those that we instrumentally desire we want only as a means to some further end. We can make this a little more precise by taking instrumental desires to be desires that we might immediately lose were we to change our beliefs about how their objects relate to other things we desire, value or care about. Paradigmatically, one has an instrumental desire for A if one has that desire only because one has a further desire for B, and believes that A is conducive to B; and is also disposed to lose the desire for A immediately on learning that A is not conducive to B. Instrumental desires are therefore formed and lost by a wholly rational process, whereas the process by which intrinsic desires are formed and lost is to some extent arational. Here I am only concerned with intrinsic desires, and my view is that instrumental desires are states of a very different kind.

Talk of instrumental desires may be closely related to the fact that we often use the word 'want' in describing what we are aiming for or hoping for on particular occasions. We say things like 'I want to finish my draft today', 'I want to avoid being on the tube at rush hour', or 'I want Nadal to win this match, because I think his performance has been courageous'. In cases of this kind we seem to be describing either our intentions, or preferences which we have consciously adopted. So we are describing the *outputs* of choices or decisions, rather than the inputs – which might also be described as desires, or by talking about what we want. My concern here is with a class of psychological states that act as inputs to choices, and I will reserve the term 'desire' for this class.

A further philosophical distinction is between what are sometimes called *desires proper* and *pro-attitudes* (Schueler 1995). This is not a distinction between two disjoint classes; instead, desires proper are thought to be a subset of the pro-attitudes. Pro-attitudes also include instrumental desires, emotional urges, and possibly also intentions and normative and evaluative beliefs. What all of these kinds of states have in common is that they can motivate us to act in combination with beliefs about what the likely consequences of our actions would be, or about other ways in which our actions would promote the objects of our pro-attitudes. The distinction is often used in arguing that just because a certain mental state can be motivating, it does not follow that it is a desire. So, for example, a philosopher might argue that fear is a form of desire, because fear causes actions in combination with instrumental beliefs. The distinction between desires proper and pro-attitudes could be used to rebut this simple argument. But whether this rebuttal succeeded or failed would depend on which sense of 'desire' was relevant in the context; some philosophers, such as Michael Smith (1987), call all pro-attitudes 'desires'.

My topic here is desires proper, and I take it to be an empirical question whether there are pro-attitudes which are not desires. But it is important to note that there may be no substantive difference on this point between me and a philosopher who uses 'desire' in Smith's way. Such a philosopher could pose the same empirical question by asking whether there is more than one kind of desire. I touch on this question in chapter 4.

Finally, desires are sometimes thought of as a kind of conscious experience. On this way of thinking about desire, what it is to (occurrently) desire something is to *feel* a desire for that thing. Philosophers who thought of desire in this way might claim that part of what characterises desires as a category of mental state is some distinctive phenomenal quality that they all share, and they might also suggest that the phenomenology of desire contributes to explaining how and why desires motivate us to act. But one need not be a philosopher to think of desire as a kind of conscious experience. We commonly use expressions such as 'I felt a strong desire to…', and in some contexts this phrase would be an equivalent substitute for 'I wanted to…'. To see that we use talk of what we want both in this way, and to describe our intentions, note that both of the following descriptions could reasonably be given of a case of mild temptation: 'I didn't want to go back to bed, but I felt a strong desire to do so'; 'I had decided not to go back to bed, but at that moment I really wanted to'.

The phenomenology of desire is not part of my topic here, but that does not mean that desires as conscious experiences are a different kind of mental state from the desires I will be discussing. Conscious experiences either are, or are very closely linked to, instances of activity in the brain. So it is possible that we consciously experience some or all of our occurrent desires (in my sense of ‘desire’), and that these are the very conscious states that are sometimes called ‘desires’.

1.3 Assumptions and Objections

I will assume that if there is some natural kind of human psychological state that does enough of what desires are normally thought to do, then what it is to be a desire is to be a member of this kind. I will further assume that the following two claims follow from this basic assumption: first, that if the same kind of psychological state also exists in other animals, then those other animals also have desires; and second, that if there is such a natural kind, then desires are brain states, in a sense which I will shortly explain. Given these assumptions, it makes sense for me to use empirical evidence to investigate what desires are like, especially since my main interest at this point is in the causal role and biological functions of desire. It is hard to deny that science is the most appropriate method for studying such matters. These assumptions entail a view of the metaphysics of desire which is similar in its essentials to David Lewis’s metaphysics of mind (Lewis 1980, 1994). In this section, I will first describe some similarities and differences between Lewis’s position and the approach I am adopting here, then consider two potential objections to these assumptions.

According to Lewis, our concepts of mental states such as belief, desire and pain are concepts of states that occupy certain causal roles. These causal roles are determined by the places that the various kinds of mental states take in folk psychology, which Lewis takes to be a theory of the causes of behaviour. However, he denies that token physical states are desires (for example) if and only if they occupy the specific causal role that folk psychology associates with desire. Instead, his view is that for a token physical state to be a desire it must be a member of some kind, the members of which typically occupy the correct causal role. Lewis uses the case of pain to illustrate this point. Suppose some particular pattern of nervous system activity occupies the causal role associated with pain, in almost all humans; then even if there was some man who was disposed to behave differently from the rest of us in response to this kind of neural activity, Lewis’s view is that it would still be pain (Lewis 1980). Lewis also does not insist that a neural kind must occupy the exact causal role imagined by folk psychology in order to constitute a mental kind, as long as the match is close enough, and no other neural kind comes closer. In these two fundamental respects, my basic assumption entails a view of desire which follows Lewis's theory closely.

Lewis also writes that mental state kinds may be constituted by different physical kinds in different populations. So human desire, for example, may not be the same thing as desire in robots or alien species. Assuming that we cannot share psychological natural kinds with robots or aliens, my assumption as it stands entails that robots and aliens do not have desires. To avoid this consequence, I could adopt Lewis's approach, re-writing my assumption to specify that what it is to be a *human*, *mammalian* or *animal* desire is to be a member of the right natural kind. For my purposes here, however, the possibility of robot and alien desires is irrelevant, so I will leave open the choice between these two alternatives.

My assumptions also leave open two further questions on which Lewis takes firm views. First, for Lewis mental states are characterised by their *causal* roles, whereas an alternative view is that it is *teleological functions* that matter. A state's teleological function is, very roughly, how it is *supposed* to interact causally with other entities, and such functions can arise naturally in virtue of the contributions traits make to organisms' survival and reproduction (see ch. 7 for more detail on this topic). Lycan (1987, ch. 4) argues that there are several advantages to thinking of mental states as characterised by teleological functions rather than causal roles, and Millikan (1996, 2002) argues that biological and psychological natural kinds are defined by their functions, rather than by causal roles or anatomy and physiology. So my talk of 'what desires are normally thought to do' could be construed either in terms of causal roles, or of teleological functions. I have no need to settle this issue, because the empirical evidence I will present identifies a natural kind that fits our usual way of thinking about desire very well both in what it typically does, and in its apparent function.

Second, Lewis is committed to the controversial claim that mental state terms are *nonrigid designators*. *Rigid* designators are terms that refer to the same object in every possible world in which that object exists, and never to anything else; nonrigid designators refer to different things in different possible worlds. Lewis's view is that it is possible that different neural kinds could have occupied the causal roles characteristic of the various mental kinds, so mental kinds such as desire could have been other than as they actually are. For example, suppose that desires are instances of activity in cortical region X, defined by its position in the head; activity in some other region might have played this causal role, so as Lewis would put it, desire might not have been desire. Kripke (1980) argues that this consequence makes Lewis's position untenable.

Lewis's view is that this is no objection to his theory, because there is no good reason to suppose that mental state terms are rigid designators. But my assumption does not require that I endorse Lewis's position here. An alternative view is that what characterises desires *as a natural kind* is either their causal role, or their teleological function. If this is correct, the identity conditions across possible worlds of this natural kind might be such that it is not possible for human desire to be constituted by any other. In this case, 'desire' would be a rigid designator. And crucially, this view is plausible, because it is doubtful that anatomical or physiological specifications can capture relatively 'high-level' natural kinds in psychology, such as desire.

The sense in which desires are brain states is therefore as follows. Provided that there is a psychological natural kind that does enough of what desires are normally thought to do, token desires are identical to token brain states. The kind *desire* is identical to a brain-state kind, but what unites this latter kind may be a shared causal or functional role, rather than some anatomical or physiological feature. Even if so, it is likely that almost all human desires share some characteristic anatomical or physiological properties, and it would certainly not follow that investigation of such properties was irrelevant to understanding desire. Desires which belong to the same kind in virtue of having the same object (such as desires for ice-cream) may well also all belong to the same causally- or functionally-individuated brain state kind, but it is much less likely that this kind would be of significant interest to neuroscience than the kind encompassing all desires.

Finally, one aspect of Lewis's theory which differs significantly from my preferred approach is the place he gives to folk psychology. According to Lewis, conceptual analysis reveals what sort of thing desires must be, if there are any desires – that is, the occupants of a certain causal role – and empirical research will tell us whether there are desires, and fill in some more details about what they are like. The conceptual analysis stage here is specifically an investigation of the commitments of folk psychology. I have two reservations about this approach, which are that ‘desire’ is to some extent a technical term in philosophy, and that it is doubtful whether there is a single coherent folk-psychological concept of desire. These points mean that careful and detailed study of folk psychology with the aim of identifying a single concept of desire may be not only unproductive, but also beside the point. Instead, we should try to identify coherent strands within existing thought about desire (both folk-psychological and philosophical) and think about how these strands relate to each other, and can contribute to wider philosophical projects, such as understanding action, reasons or representation. So my reason for making my basic assumption is that it captures an important part of how both philosophers and ‘the folk’ think about desire, and focusing on this part has the potential to be philosophically productive.

I now turn to two closely-linked lines of objection to my assumptions, both of which are also related to the difference just described between Lewis’s approach and mine.

First, Scott Sehon (2005, ch. 6) has argued against what he calls the *Standard View*: the claim that it is an implicit commitment of folk psychology that if there are mental states, then they are identical to brain states. His argument focuses on the possibility that we may be unable to identify brain states that fit the causal profiles associated with mental states such as beliefs and desires. In this situation, Sehon points out, those who stick to the Standard View must accept that there are no beliefs or desires. But Sehon also claims that we should be extremely reluctant to accept this conclusion – so much so that if there turn out to be no brain states that can be identified as beliefs and desires, then we should abandon the Standard View rather than accepting eliminativism.

According to Sehon, this means that whether the Standard View is plausible is contingent on certain as-yet-unknown neuroscientific facts. Given that the Standard View is also supposed to be a claim about the commitments of folk psychology, he claims that this shows that the Standard View is highly unattractive, because the commitments of folk psychology cannot be contingent on such unknown facts. The Standard View is much like Lewis’s view, because Lewis does think that it is a commitment of folk psychology that mental states play characteristic causal roles. So Lewis must accept that if no brain states (or perhaps embodied brain states) play the relevant causal roles, then there are no beliefs or desires.

The reason this argument poses a challenge to my assumptions is that Sehon's claim that we should be reluctant to accept eliminativism about beliefs and desires is very plausible. In particular, Sehon emphasises that belief-desire psychology is explanatorily valuable even if it is not reducible to neuroscience, because it is the primary means by which we give *rationalising*, rather than causal, explanations. Something like this view is shared by a large and diverse group of philosophers. So the challenge is to show how to reconcile my assumption that if there are suitable brain states, then those states are desires, with the attractive claim that if there are no such brain states, we nonetheless have desires.

My approach can reconcile these points – unlike Lewis's – because my assumptions are only *partly* motivated by respect for the apparent commitments of folk psychology. In making them I am also aiming to contribute to the development of the most useful possible taxonomy of mental states for a range of explanatory purposes. That is, I expect my assumptions to be philosophically productive; and what it is philosophically productive to assume about the nature of desire may well be contingent on future neuroscience. The natural-kind approach to desire that I have adopted has the potential to be very productive, as I hope to show in the course of this thesis (although as I describe briefly in section 10.1, it may also yield far more of interest than I am able to address here), provided that there is a natural kind of psychological state that does enough of what desires are ordinarily thought to do. If not, then philosophers should seek another way to think of desire, and focus their attention more exclusively on its rational and perhaps phenomenological aspects.

This brings us to the second possible line of objection to my assumptions, which is that since I will not engage in conceptual analysis of the term 'desire', I risk changing the subject. This objection might be motivated by the thought that an account that purports to say what desires are must start by giving a detailed analysis of the concept of desire. Unless grounded in such an analysis, the account might have independent value, but it would not be an account of *desire*.

This objection fails primarily because, as I have already emphasised, it is doubtful whether there is a single coherent concept of desire in folk psychology, and because the term 'desire' is used in a range of different ways by philosophers. These points mean that there is no wholly determinate subject to be changed. However, we should also not overestimate the differences between different conceptions of desire. There is a very widely-shared idea that desires are mental states which interact with instrumental beliefs to motivate action, which also affect our moods and emotions, and which have a wide range of objects. So even though there is variation among those who think of desires in this way, theories of desire which start from this shared conception are likely to have broad significance.

In this chapter, I have further explained what I mean by ‘desire’, and defended my assumption that if there is some natural kind of psychological state that does enough of what desires are normally thought to do, then what it is to be a desire is to be a member of this natural kind. In the next three chapters, I will therefore seek to show that there is such a natural kind, and to give a detailed account of some of the more philosophically-significant properties of members of this kind.

Chapter 2: Two Systems: Background Psychology and Neuroscience

2.1 Two Systems for the Pursuit of Reward

Much of the empirical research relevant to understanding desire is based on the apparent result that rats and humans use two systems for action-selection, running in parallel. These are the *habitual* and *goal-directed* systems. Given the way that the goal-directed system is thought to work, showing that these two systems exist would be a major step towards showing that there is a natural kind of mammalian psychological state that can reasonably be identified as desire, and also to confirming premise II of my overall argument. In the next section, 2.2, I describe behavioural studies which provide evidence for the existence of these two systems, then in section 2.3 I give a brief account of the structure and function of the basal ganglia, which is essential background for understanding the neuroscience of action selection. In section 2.4 I give an initial account of the function of dopamine, and in 2.5 I describe neuroscientific results which provide further evidence for the existence and distinctness of the two systems. In this section, I introduce the two systems; almost all of the points mentioned in this section will be explored in more detail later on.

According to current theories, the habitual and goal-directed systems both change in response to the individual's experiences. In different ways, they both keep track of how good the apparent results are of the individual's actions, and modify their future behaviour accordingly. They therefore contrast with *reflex* systems, which tend to produce the same action in the same circumstances, regardless of how things have gone in those circumstances in the past. The two systems are also both general-purpose, in the sense that rather than being systems for helping us to get specific beneficial outcomes, such as food or healthy offspring, or to avoid specific threats, they are capable of helping us to get any of these outcomes, and to balance the demands made by our various needs. So they are thought of as systems for maximising *reward*, which we can think of as the 'common currency' by which actions and outcomes are measured. The two systems are distinguished primarily by the kinds of information about actions and reward that they store and use.

In the habitual control system – which I will also call simply the *habit system* – behaviour is controlled by learnt associations between *stimuli* and *responses*. Stimuli are features of the animal's circumstances which they are able to perceive, and responses are actions. So either the patterns of behaviour generated by this system, or the states that are responsible for causing these patterns of behaviour, are sometimes known as *S-R associations*, acquired by *S-R learning*. I will usually call S-R associations 'habits'. S-R learning takes place when stimuli are followed by responses, which in turn are followed by *reinforcement signals*. These are signals produced in the brain which vary in strength and valence according to the level of reward which is perceived as being provided by the environment. According to modern theories, reinforcement signals represent *reward prediction errors*; that is, the difference between the level of reward perceived, and that which was expected. Positive reinforcement signals strengthen associations between stimuli and rewards, and negative ones weaken them; so habits get stronger when things subsequently go better than expected, and weaker when things go worse than expected. In so far as the habit system controls our behaviour, we perform the actions that we represent as being most valuable in our current circumstances.[2]

S-R learning is the 'classic' form of *operant conditioning* (also called *instrumental conditioning*) which was studied by psychologists working in the behaviourist tradition (Thorndike 1905, Hull 1943). However, unlike the behaviourists, modern researchers typically think of habits as *representing* the expected values of responses to stimuli.

In contrast, the goal-directed system works by keeping track of two different relationships. These are the probabilistic relationships between actions and outcomes (which will be contingent on the circumstances), and the levels of reward that are associated with outcomes. In this chapter I will call the states representing these two relationships *action-outcome contingencies* and *outcome values*, respectively. An important feature of the goal-directed system is that these relationships can be learnt about independently, with the information being stored and recombined for later use.

[2] It's normal in modern cognitive neuroscience to think of the habitual and goal-directed systems as working *stochastically*, meaning that they determine the probabilities of actions, rather than fixing absolutely which will be performed. This is an important feature of the systems, but it is not essential to the idea of *a habit system* or *a goal-directed system*.

For example, an animal might learn that eating honey is rewarding on one occasion, and that breaking open beehives tends to lead to getting honey on another occasion, then use these two pieces of information in deciding how to act on a third occasion. The goal-directed system is more sophisticated than the habit system (I will describe some of its advantages later) and also correspondingly more demanding, in that it requires the capacities to acquire, store and update both action-outcome contingencies and outcome values, and to combine them in action-selection. Under given circumstances, the goal-directed system is thought to calculate the expected values of each salient action from the agent's representations of outcome values and action-outcome relationships, and then to cause the action with the highest expected value.

Although there is much more to be said about the two systems, at this point it will be useful to remind ourselves of my first two premises:

> I. Desires are outcome values.
> II. The goal-directed control system works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.

The descriptions of the two systems just given have two consequences for these premises. First, if I can show that the goal-directed system is real – that there really is a psychological system in humans and other mammals that works in the way I just described – that will establish premise II. Second, given the assumptions I set out in section 1.3, to establish premise I I need to show that outcome values form a psychological natural kind that can reasonably be thought of as desire. An initial point in favour of this claim is that in folk psychology we recognise, and tend to distinguish, habitual behaviour and behaviour caused by beliefs and desires. So if the two systems are real and distinct, then since there is a structural similarity between the goal-directed system and the way we think desires influence action, there is some reason to think of the inputs to the goal-directed system as desires and instrumental beliefs. For these reasons, my priority in this chapter is to show that the two systems are real and distinct.

2.2 Behavioural Evidence for the Two Systems

In the 1980s, clear behavioural evidence emerged that rats' actions can be controlled by states analogous to desires and beliefs, as well as by habits, contradicting previous behaviourist theories. It is now widely accepted that rats and other animals, including humans, are capable of both stimulus-response and response-outcome (R-O) learning, contributing to distinct systems for habitual and goal-directed control, a result anticipated by Tolman (1949). The behavioural techniques that were developed in the 1980s are now being widely used to probe the neural mechanisms that support these systems.

The primary source of evidence for goal-directed control in rats is studies of *outcome devaluation*, the first of which was performed by Adams and Dickinson (1981). Outcome devaluation experiments typically have the following form. On the first day, rats are given the opportunity to press a lever, and are given a food reward such as sucrose when they do so. The rats are exposed to this environment for long enough to learn an association between lever-pressing and reward. Then on the second day, the rats are divided into two groups. One group experiences *outcome devaluation*, meaning that they are allowed to consume the particular food being used in the study, then injected with lithium chloride, which induces gastric illness. The second, control group also receives both the food and the injection, but these are given at different times, with the intention that the rats will treat them as unrelated events. Some studies use different methods for devaluing the outcome, such as inducing *specific satiety* – allowing subjects to eat all they want of the food reward. There are also various possible procedures for generating control data. All of the rats are kept away from the lever on the second day. Finally, on a third day, the rats are again given the opportunity to press the lever, but this test is conducted *in extinction*, meaning that the rats do not receive a reward for lever-presses. Adams and Dickinson and subsequent investigators have found that the rats for whom the outcome was devalued press the lever significantly less than the controls in the test phase on the third day.

In these studies, the two groups of rats perform differently in the test phase, and this difference must be explained by some difference in the rats' experiences during the experiment. The only such difference comes on the second day, when one group experiences the pairing of the food reward with illness, and the other does not. It is hard to resist the conclusion that the rats behave differently in the test phase because they value the food differently, and if this is accepted, it must also be accepted that the rats anticipate receiving this food when they press the lever. In other words, the rats' behaviour is controlled by states apparently representing the values of outcomes, and further states that seem to represent relationships between actions and outcomes. Rats apparently possess a basic form of belief-desire psychology, and I will later refer to these two kinds of states as desires and instrumental beliefs, respectively.[3]

However, rats do not always show sensitivity to outcome devaluation. After some training regimes, they will continue to perform actions even when the outcomes presented in training have been devalued. In particular, *overtraining* leads to loss of sensitivity to devaluation (Adams 1981). Since insensitivity to devaluation is what would be predicted by traditional accounts on which S-R associations are learnt by operant conditioning, it is widely accepted that this shows that rats use two systems, one goal-directed and one habitual, to control actions. It is thought that the systems work in parallel, but that the habit system learns more slowly than the goal-directed system, which would explain why extensive training leads to insensitivity to devaluation, since it causes a transition to habitual control. Further evidence for the view that habits as identified by outcome devaluation studies are indeed S-R associations is provided by the observation that these behaviours are sensitive to changes in the environment in which the action is performed (Killcross & Coutureau 2003).

In similar studies, humans have been found to perform in similar ways. Instead of devaluing foods by inducing illness, in human studies devaluation is produced by inducing specific satiety. In one study (Tricomi et al. 2009), participants were trained to press two buttons to receive small quantities of two different foods. After a limited amount of training, one food was devalued, and the participants reduced their performance on the button associated with that food, in extinction, to a greater extent than the other button. However, after overtraining, devaluation did not have this effect. These results are similar to those found with rats in very similar paradigms (involving two actions, and using specific satiety as the devaluation

[3] Why 'basic'? One reason is because outcome devaluation does not show that rats can learn about action-outcome relationships in any way other than by performing actions themselves, or about the values of outcomes other than by experiencing them themselves. Another is that it does not show that they can perform chains of instrumental reasoning.

mechanism; Balleine & Dickinson 1998). A recent experiment gives a vivid illustration of the two systems at work in humans, and shows why S-R associations are thought of as habits (Neal et al. 2011). Participants were in a cinema, watching a film, and were given popcorn that was either fresh and delicious or stale and unpleasant. Those who had frequently eaten popcorn in cinemas before ate the same amount of popcorn, even if it was stale, whereas those were not regular popcorn-eaters ate far more of the fresh popcorn. That is, those who were not 'in the habit' of eating popcorn were sensitive to whether eating the popcorn produced a pleasant outcome, but those that were 'in the habit' ate on regardless. Psychologists would describe their behaviour as controlled by the stimulus, rather than by the outcome.

Outcome devaluation has now become a central experimental paradigm in this area of psychology and neuroscience. When researchers want to study the effects of different manipulations of the learning environment specifically on either goal-directed or habitual behaviour, they train those behaviours using schedules known to produce either sensitivity or insensitivity to devaluation. When they want to know how a particular brain area is involved in goal-directed or habitual control, or whether a given area is necessary for one of these two processes, they use outcome devaluation as a means to test what systems animals are using. The fruitfulness of these experiments, some of which are discussed below, is further evidence both that outcome devaluation is a robust effect and that it tells us something substantial about cognitive architecture. However, it is worth bearing in mind that outcome devaluation is not without idiosyncracies as a means of testing for goal-directed or habitual control. For instance, although it is an advantage of outcome devaluation that the same effect is found whether induced illness or specific satiety are used, it is noteworthy that these both affect the value of an action in the same direction. Also, outcome devaluation seems to test directly for the involvement of a representation of an outcome in action selection, and only indirectly, via the idea that stimulus-control or outcome-control exhaust the possibilities, for the involvement of an S-R association. Issues like these may be important when considering the implications of specific observations, but can only be assessed in the context of specific claims about such observations.

Another behavioural technique which is used to distinguish goal-directed from habitual responding is *contingency degradation*. Goal-directed control relies on pairs of states; agents produce goal-directed actions only when they place high values on outcomes, and expect their actions to lead to those outcomes. Contingency degradation is the counterpart to outcome devaluation, but with agents' expectations of outcomes, rather than the values they place on those outcomes, being manipulated. It is important for contingency degradation experiments that the theory of S-R learning requires only *contiguity*, not *contingency*, in the relationship between responses and the delivery of reinforcers. What this means is that S-R associations are strengthened whenever reinforcers are offered, even if they are also frequently delivered at times when the relevant action has not been performed. For the reinforcers to be contingent on the response, they would have to be delivered only when the response occurred.

Two different paradigms are collectively referred to as contingency degradation. First, outcomes can be delivered without the relevant action's being performed, degrading the contingency of outcomes on actions. In theory, this manipulation should reduce goal-directed action, but not habitual action, because actions continue to be reinforced. Hammond (1980) and more recent researchers (see Balleine & O'Doherty 2010) have found that this procedure does reduce goal-directed responding, confirming that action-outcome representations require contingency rather than merely contiguity, and this manipulation is now used when testing for goal-directed control (e.g. Yin et al. 2005). Second, Dickinson and colleagues (1998) found that undertrained rats, but not overtrained ones, could adapt their behaviour in response to *omission schedules*. These are manipulations in which an outcome that has previously always followed a particular action is now delivered only when that action is withheld. This again shows the sensitivity of rats to action-outcome contingencies, using evidence from rewarding outcomes that occur without action as well as those that follow action, and it can also be used to test for goal-directed or habitual control.

I will discuss further evidence for the existence of these two distinct systems in section 2.5; before we turn to that evidence, it will be useful to review the functional anatomy of the basal ganglia.

2.3 The Basal Ganglia

The basal ganglia are a group of nuclei contained within the cerebral hemispheres which are of particular importance for understanding motivation and action.

Understanding their basic functional anatomy is useful both for understanding the goal-directed and habitual systems specifically, and for beginning to see the how the brain achieves adaptive action-selection more generally. The basal ganglia are connected to many parts of the cortex by loops that pass from the cortex, through the basal ganglia and thalamus, and back to the same areas of cortex. As I will describe in this section, they are thought to provide a mechanism for selecting actions by disinhibiting cortical activity (Redgrave et al. 1999). This should not be misinterpreted; it has been argued that the selection and control of action is the ultimate purpose of the brain as a whole, and there are several other areas that make relatively direct contributions. The suggestion is rather that the basal ganglia choose among possible actions, which have been identified and evaluated by partially distinct systems, and a detailed account has been developed of how they do this. This account provides useful background for understanding how the two systems could be implemented. In addition to this, however, the basal ganglia have been implicated in action learning and selection in a number of further roles, related to the wide range of cortical areas with which they are closely connected (Balleine & O'Doherty 2010). These more varied and complex roles are more directly relevant to understanding goal-directed and habitual control, and I discuss them in elsewhere in this chapter, and in subsequent chapters.

The basal ganglia are connected to the cortex by parallel, partially segregated loops (Alexander et al. 1986). Each small area of cortex sends projections to the basal ganglia, which in turn project to the thalamus, which projects back to the cortex. Projections from the cortex to the basal ganglia and from the thalamus to the cortex are primarily excitatory, using the neurotransmitter glutamate, but those from the basal ganglia to the thalamus are inhibitory, using the neurotransmitter GABA. So in general, the effect of loops through the basal ganglia is to suppress activity in the cortex. However, when signals from the cortex to the basal ganglia are relatively strong, this causes the inhibitory signal to be weakened, so the area of cortex that produced the strong signal will be disinhibited. The effect of the basal ganglia is therefore to inhibit weak signals, and disinhibit strong signals. We can already see how a mechanism like this could select actions; the action associated with the strongest signal would be selectively disinhibited, while the others are suppressed. Very roughly, then, the idea is that candidate actions are associated with instances of activity in the cortex, and when things are going well the strength of this activity will be proportional to the value of the action; these signals compete with one another, some getting stronger and others weaker, and actions that get ‘strong’ enough are performed. The contribution of the basal ganglia is in facilitating (and as we will see later, influencing) the competition. Philosophers may find this reminiscent of the idea that the mind somehow ‘weighs’ desires, causing the agent to act on the strongest.

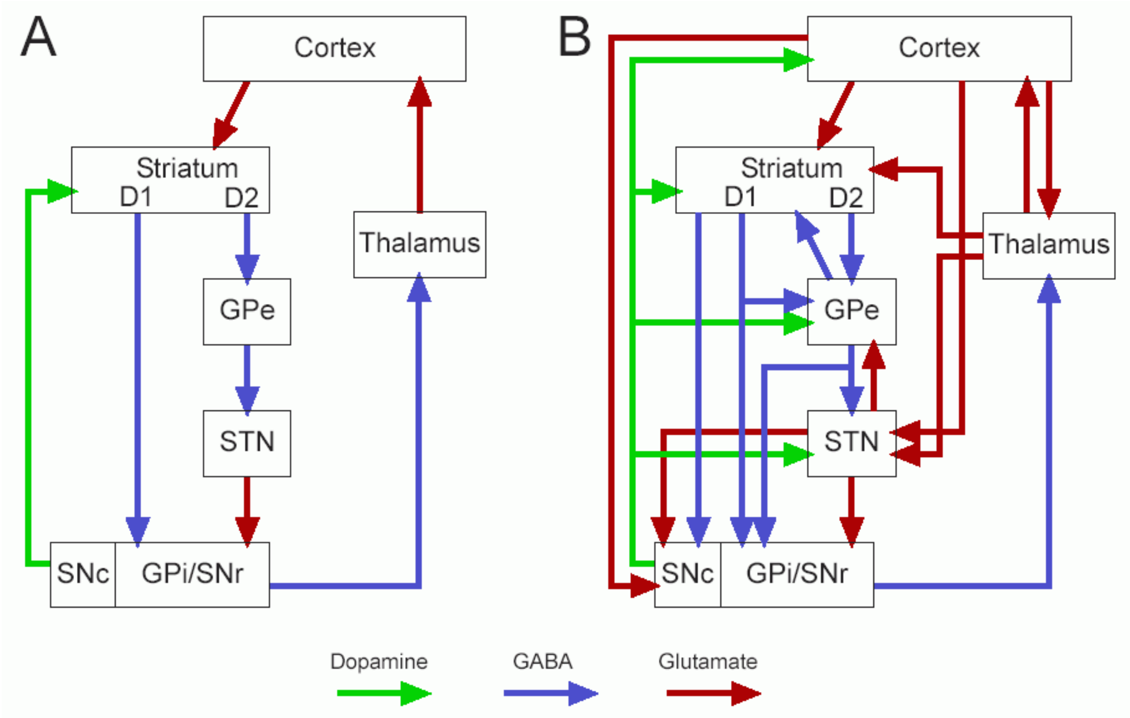

Fig. 1: Basal ganglia connectivity showing direct and indirect pathways (A), and the full extent of connectivity (B). Redgrave (2007).

A number of mechanisms contribute to the function of disinhibiting strong signals, and inhibiting weak ones. The first of these is the direct and indirect pathways, shown on figure 1A. The direct pathway is the GABAergic connection from the striatum, the input nucleus of the basal ganglia, direct to the GPi/SNr (globus pallidus internal segment and substantia nigra pars reticulata), the output nuclei, while the indirect pathway is the other route shown in 1A, that goes through the GPe (globus pallidus external segment) and STN (subthalamic nucleus). The direct pathway inhibits the output nuclei, meaning that the overall effect of this pathway is disinhibitory, because the inhibition of the thalamus from the output nuclei is itself inhibited. This disinhibition will be greater for stronger inputs. The way in which the direct pathway works relies on the fact that loops through the basal ganglia are largely segregated, meaning that the effect of the basal ganglia on a particular area of cortex is primarily determined by the signal that very area sends to the striatum. So the effect of the direct pathway alone amplifies strong signals, and suppresses weak ones. In addition, though, the connections making up the direct pathway are focused on specific areas of the GPi/SNr, while those making up the indirect pathway are diffuse. The indirect pathway excites the output nuclei, and consequently has an inhibitory effect on activity in the cortex. This means that when a strong signal is received by the striatum, the direct pathway will strongly inhibit a focused area of the GPi/SNr, while the indirect pathway will excite a diffuse area. So the indirect pathway has the overall effect of inhibiting activity in loops close to those that are most strongly active.

This brings us to another important point. The motor cortex is arranged topographically, meaning that areas associated with nearby parts of the body are close to one another. Consequently, it is particularly important that one signal is selected to the exclusion of others in nearby corticostriatal loops, because each part of the body can only perform one action at a time. So as well as inhibiting weak signals and disinhibiting strong ones, the basal ganglia have mechanisms that generate competition between proximal loops. The indirect pathway is one such mechanism. There are also inhibitory interneurons in the striatum, and inhibitory collaterals on striatal cells, that have a similar effect. When cells in the striatum are strongly activated, these connections inhibit activity in other striatal cells. Finally, the medium spiny neurons, which are the cells in the striatum with which the direct and indirect pathways originate, can be in 'up' or 'down' states. They only fire in up states, which they only enter when they are strongly activated.

These mechanisms allow the basal ganglia to select actions, but they are not in themselves action-specific; in principle, they could make other choices too. It has been suggested that the basal ganglia could select subjects for thought,[4] and it is very likely that they select actions at multiple successive levels of description. When we are playing squash, for instance, our brains need to select where to direct the ball, a choice at a 'higher' level, and how to move the legs, arms and fingers to produce

[4] Redgrave et al. 2010 mentions 'poverty of thought' as a symptom of lesions to certain areas of the basal ganglia, comparable to the bradykinesia (impaired movement) seen in Parkinson's Disease, which is a consequence of damage to other basal ganglia regions.

such a shot, which are choices at a ‘lower’ level. Both of these processes could involve resolving competitions between various alternatives. These suggestions are also anatomically plausible, since the basal ganglia are connected in loops to many areas of the cortex, not just the primary motor cortex, and all of these loops pass through the direct and indirect pathways. For more information and references on the basal ganglia, see Redgrave (2007).

2.4 Dopamine and the Habit System

The basal ganglia are also the major site for dopamine activity. Dopamine is a neurotransmitter released by the ventral tegmental area and substantia nigra pars compacta (VTA/SNc), collectively known as the midbrain dopamine neurons, which is often thought to be closely associated with desire. For instance, Schroeder’s (2004) theory of desire, as it applies to humans, amounts to roughly the claim that to desire an outcome is to be disposed to produce a dopamine signal when that outcome occurs. Dopamine seems to have multiple functions, and there is considerable uncertainty about at least some of these, which I discuss in section 3.5. This uncertainty is currently a major obstacle to a fuller understanding of both goal-directed and habitual control. So dopamine is an important subject, but for now I will only give a preliminary account of two of its functions.

One proposed function of dopamine is to facilitate and motivate action. The direct pathway originates with medium spiny neurons (MSNs) that have D1 dopamine receptors, meaning that they are excited by dopamine, while the indirect pathway originates with MSNs that have D2 receptors, meaning that they are inhibited by dopamine. This means that when dopamine levels are relatively high, the direct pathway is predominant, and action representations in the cortex are readily excited by the basal ganglia. When dopamine levels are low, the indirect pathway is predominant, and action representations are more suppressed. So overall dopamine levels determine overall readiness to act, and correspondingly, Parkinson’s Disease is caused by underproduction of dopamine (Redgrave et al. 2010). This function of dopamine is thought to be related to *tonic* rather than *phasic* dopamine release (Niv et al. 2007). Dopamine neurons in the VTA/SNc fire continually at a tonic rate, with sudden bursts of firing or pauses in firing in response to specific events. So according to this proposal the tonic rate, which changes relatively gradually, determines readiness to act, while the phasic events – the sudden bursts and pauses – have other functions.

This account of the function of tonic dopamine is relatively uncontroversial, but the functions of the phasic dopamine signal are a matter of ongoing debate (Berridge 2007, Redgrave et al. 2008, Horvitz 2009, Bromberg-Martin et al. 2010). One hypothesis that is particularly well-known and relevant for present purposes is that phasic dopamine constitutes a reinforcement signal for habit learning. According to one computational model of habit learning, the temporal difference model (Sutton & Barto 1998), habit learning can be facilitated by a *reward prediction error* (RPE) signal.[5] The RPE signal indicates the difference between the level of reward currently being received by the organism and the level that was predicted. In the context of a habit system, actions that are performed shortly before positive RPE signals are reinforced, while those followed by negative RPE signals are inhibited. When things go exactly as it expects, the agent does not change its behaviour. It makes sense that a system like this would work, at least if we assume that predicted levels of reward are determined by current habit strengths, because this means that positive RPE signals would constitute evidence that the response is rewarding enough to make a stronger habit appropriate. Reward prediction error signals effectively allow agents to keep track of the running average reward that performing each habit brings them, to which the strengths of those habits should plausibly be calibrated. So computational considerations give us a reason to expect to find RPE signals somewhere in the habit system.

In this context, neurophysiological results of three kinds support the hypothesis that phasic dopamine indicates RPEs. First, a classic study by Olds and Milner (1954) found that rats would electrically self-stimulate the VTA/SNc to the exclusion of other activities, so dopamine has long been thought to signal reward or act as a reinforcer. Second, dopamine is known not to simply signal reward, because Schultz (1998) found that phasic dopamine responses diminish as rewards become predictable, and transfer to conditioned stimuli that predict primary rewards. That is, dopamine bursts are produced when rewards are surprising, but not when they are predicted. Dopamine also appears to signal negative reward prediction errors – if an

[5] Shea (2014) discusses the RPE signal from a philosophical perspective.

expected reward is not delivered, dopamine firing pauses briefly. Third, phasic dopamine causes long-term potentiation of cortico-striatal synapses (Reynolds & Wickens 2002, Arbuthnott & Wickens 2007), which is a mechanism by which dopamine could reinforce recent actions.

A fairly simple account of the neural basis of the habit system thus appears to be emerging. Sensory stimuli are represented in the cortex, and other cortical areas represent and are capable of causing actions, while cortical areas of both kinds are connected to the basal ganglia by loops of the sort I have described. Dopamine signals RPEs, and is capable of affecting the strength of cortico-striatal synapses. So habits are connections in the striatum between representations of stimuli and responses, which are modified according to the extent to which they produce rewards. This account seems to be supported by evidence from lesion studies in rats (Yin et al. 2004, 2006) and imaging studies in humans (Tricomi et al. 2009), identifying the dorsolateral striatum (DLS), a part of the striatum connected to the sensorimotor cortex, as crucial for learning and performing habitual behaviours.

However, there are a number of complications to this picture. One is that it involves excitatory connections in the striatum between projections from distinct cortical areas, which are not part of the anatomical description given above. Still, recent studies suggest that cortico-striatal loops can interact (Haber & Knutson 2010). Excitatory interneurons connect medium spiny neurons, and cortical areas have diffuse as well as focused projections to the striatum. Another complication is that temporal difference learning is often thought to involve separate elements, called the ‘actor’ and the ‘critic’, that are responsible for driving action-selection via S-R associations, and representing and updating the values of the actions performed, respectively (Sutton & Barto 1998, Landreth 2009, Shea 2014). It has been suggested that the dorsal striatum and associated cortical areas form the actor, while the ventral striatum, along with parts of the cortex, plays the role of the critic (O’Doherty et al. 2004). Third, the simple account does not explain how the RPE signal is generated, and while the actor-critic view does account for reward predictions (producing these being the main function of the critic), it does not explain how received reward is measured. In addition to all of these, of course, there is the point that the function of phasic dopamine signals is a matter of controversy. But the idea that these signals are RPEs for updating the habit signal is now a common starting-point for that debate, and it certainly gives us further insight into how the habit system *could* work.

2.5 Could there be just one system?

So far I have presented the basic behavioural evidence for the existence of the two systems, described some of the functional anatomy of the basal ganglia, and suggested a partial account of the implementation of the habit system. I have not discussed how the goal-directed system might work, because that is the topic of chapter 3. Before moving on to it, I will discuss whether there is a viable alternative to the two-systems account of action selection. I will consider two possible alternatives. It should be noted that the goal-directed and habitual systems cannot possibly be *wholly* separate, since they both have the function of controlling the behaviour of a single body – at some point there must be an action-determining mechanism that is influenced by both systems.

One reason to consider this question is that it may be easier to explain how a more sophisticated version of the habit system could have evolved, than how a new, goal-directed system could have been added.[6] However, this idea is to some extent undermined by the fact that the main elements of the goal-directed system seem to be independently useful. First, states representing the values of outcomes are useful not only for deciding which outcomes to pursue, but also for assessing the level of value provided by ongoing states of affairs. Even creatures that lack goal-directed systems need the capacity to perform this kind of assessment, if they are to learn new habits. This does not show that any creature with a habit system must also have something akin to desires, because an important feature of goal-directed control is that outcome values themselves change according to the individual's experiences, and the assessment of value for the purposes of habit-learning could rely just on innate states. But nonetheless, desire-like states are valuable even before they can play a direct role in action selection. Second, the ability to learn about action-outcome contingencies is just a special case of the ability to learn about contingencies between events in general, which is very plausibly of independent value for several purposes. So given that it is made up of independently-useful

[6] David Papineau has repeatedly pressed me in conversation to explain exactly why studies in this area show there to be two systems, and has emphasised this point.

components, perhaps we should not be surprised by the evolution of a largely separate goal-directed system.

The first possible alternative to the existence of two separate systems is that occurrent outcome values act as internal stimuli to which the habit system determines responses. There is reason to believe that *basic drives* such as hunger and thirst influence the habit system in this way (see section 3.3). Assuming it is possible for outcome values to do so too, an alternative account of outcome devaluation seems to be available. When outcome devaluation has taken place, the agent will no longer generate positive occurrent outcome values for the reward used, so their habitual behaviour will no longer be triggered. This account faces several difficulties, however. It does not explain why some actions are sensitive to outcome devaluation, and others are not; it is worth noting that the nature of the training regime, as well as its extent, affects this sensitivity (Dickinson & Nicholas 1983). Neither does it explain differences in the results of contingency degradation studies. It is also unclear whether a outcome value could become part of the stimulus in a typical outcome devaluation experiment, since these involve agents experiencing new outcomes for the first time. This account asks us to believe that a outcome value for sucrose solution can be salient to a rat as he takes an action that leads to his first exposure to this type of food.

However, a second alternative to the two-systems account is harder to refute. On what I will call the *sensitive habits* view, states representing action-outcome contingencies only cause actions by their effects on S-R associations. If a particular rewarding outcome is found to be contingent on a particular action, an excitatory connection is formed between a representation of the outcome and the S-R association that controls that action. This would mean that habits are boosted when they are expected to lead to valuable outcomes. Just as on the two-systems account, action-outcome contingencies and S-R associations are learnt in parallel, but at different rates depending on the exact nature of the training regime. When the S-R association is relatively weak, performance will be reduced by outcome devaluation or contingency degradation as these experiences will lead to reduced excitation by the action-outcome association, but after extensive training the S-R association will be strong enough to consistently produce action independently. The sensitive-habits view does require a very different form of action learning from that involved in pure

S-R learning, because it involves both sensitivity to contingency rather than contiguity and the association with actions of subsequent events. But this view is still simpler than the two-systems view, because it does not posit two separate routes to action, but only two systems of action learning.

The best evidence we have in favour of the two-systems view, as against the sensitive habits view, comes from a series of lesion studies on rats performed by Yin and colleagues. This group found that lesions of the posterior dorsomedial striatum (DMS) produced either before or after training that would otherwise produce goal-directed behaviour left rats insensitive to both outcome devaluation and contingency degradation (Yin et al. 2005). In two other studies (Yin et al. 2004, 2006), they also found that lesions to the posterior dorsolateral striatum (DLS) made rats more sensitive to these tests, when trained in ways that usually produce habitual behaviour. These studies are taken to support the view that S-R associations are formed in the DLS, that action-outcome contingencies are represented in the DMS, that these states are learnt in parallel, and that they are independently capable of causing action. However, only one of these studies produced a result contradicting what would be predicted by the sensitive habits view.

First, we can consider the experiments showing that DMS lesions cause insensitivity to outcome devaluation and contingency degradation (Yin et al. 2005). These experiments show that there is some part of the overall system for action selection such that without it, rats behave as though they are only capable of purely habitual control – as if their behaviour is determined by S-R associations alone. This does not distinguish between the hypotheses, as both predict this result: on both hypotheses, if action-outcome representations are destroyed, only S-R associations will remain. So to distinguish the hypotheses, we need to consider the studies that produced goal-directed behaviour. However, in one set of experiments (Yin et al. 2004), the DLS lesions took place prior to any training. This leaves open the possibility that the DLS is necessary for the development of S-R associations that are strong enough to produce devaluation-insensitive responses, but that sensitive habits themselves can exist despite such lesions. Admittedly, this account does involve rejecting the common view that the same mechanism is responsible both for the formation and the gradual strengthening of S-R associations, so this result does put pressure on the sensitive habits hypothesis. But the crucial study for establishing the two-systems view comes from the work by Yin and colleagues from 2006.

In this experiment, intact rats were trained to press a lever for sucrose solution using an interval schedule, a training regime known to produce habitual behaviour. After this training, half of the rats had muscimol, a GABA-A agonist which prevents normal function, injected into the DLS. Half of each group of rats were then given training on an omission schedule before being tested for lever-pressing in extinction, while the other half were given similar training without the omission schedule before extinction testing. Rats that had not received muscimol injections performed similarly in extinction, regardless of whether they had experienced the omission schedule, showing insensitivity to contingency degradation. However, those that did not have a normally-functioning DLS were sensitive to contingency degradation; those of this group that had experienced the omission schedule responded less in extinction.

The behaviour of the non-lesioned group in this study showed that the rats developed strong, stimulus-controlled habits in their initial training. So the sensitive-habits view should predict that no lesion would be possible which would make the action sensitive to contingency degradation without destroying it entirely, because on the sensitive-habits view the S-R association would be critical to the action's performance. On the other hand, the two-systems view predicts that in addition to the S-R association, an action-outcome representation should also exist and be independently capable of controlling the action, so a lesion destroying the S-R association should render the action sensitive to contingency degradation. In this experiment the lesioned group which did not experience the omission schedule continued to perform in extinction, but the lesioned group which experienced the omission schedule reduced performance. So the experiment confirms the prediction of the two-systems view, and contradicts that of the sensitive-habits view.

In this chapter, I have introduced the goal-directed and habitual systems, and described evidence that they exist in animals including rats and humans, and are largely separate from one another. In the light of this evidence, I will now begin to use the terms 'desire' and 'instrumental belief' to refer to the two types of inputs to the goal-directed system. I do not take the case for identifying representations of outcome values of this type as desires to be complete, but it will be convenient to be able to use this term. My use of the term 'instrumental belief', meanwhile, should be regarded as wholly stipulative – I do not intend to make any substantive claims about the nature of belief. I have also given initial accounts of the structure and function of the basal ganglia, and of two putative functions of dopamine. The evidence described in this chapter helps to establish at least my first two premises, but in order to further develop the case for premise I, and to establish premise III, we need to turn to the neuroscience of desire.

Chapter 3: The Neuroscience of Desire

3.1 Anatomy of the Goal-Directed System

In this chapter, I will address three main topics: first the anatomy of the goal-directed system; then the distinction between occurrent and standing desires, and the roles they play in action selection; and finally the processes by which desires are formed and updated. These three topics will each contribute to a thorough account of the neuroscience of desire, and I will close the chapter with a diagram showing the main components and processes involved in the habitual and goal-directed system. My discussion of the latter two topics will provide direct support for premise III:

> III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.

The first topic of this chapter, the anatomy of the goal-directed system, is important because successful anatomical studies of proposed psychological systems help to show that they are real.

As we have seen, the cortex and striatum are connected by partially-segregated loops which also pass through the thalamus. Not surprisingly, then, the functions associated with different parts of the cortex are also associated with the most closely connected parts of the striatum, and vice versa. Many neuroscientists take the view that we can productively think of corticostriatal loops as integrated units for performing particular functions – for example, habitual control seems to be the joint responsibility of the DLS and the sensorimotor cortex. Similarly, the two key elements of the goal-directed system are also thought of as located in corticostriatal loops: representations of action-outcome contingencies in the DMS and medial prefrontal cortex (mPFC), and desires in the ventral striatum and associated prefrontal areas. Because the prefrontal cortex is one of the parts of the brain that differs most between rats and humans, the areas associated with desire in these two groups are anatomically different: in rats it is the insular cortex, and in primates the orbitofrontal cortex (Wise 2008, Schoenbaum et al. 2009, Padoa-Schioppa 2011).

Evidence placing action-outcome contingencies in the DMS and mPFC in rats comes primarily from lesion studies. As described above, lesions to the DMS cause insensitivity to both outcome devaluation and contingency degradation, whether they are inflicted before or after training (Yin et al. 2005). This is what would be predicted if the DMS was necessary for the formation and use of action-outcome contingencies, because without these only habitual control is possible. Lesions to the prelimbic cortex, which is part of the mPFC, cause insensitivity to both outcome devaluation and contingency degradation, when inflicted before training (Corbit & Balleine 2003); post-training lesions do not reduce sensitivity to outcome devaluation (Ostlund & Balleine 2005). This suggests that prelimbic cortex, or perhaps mPFC more generally, is necessary for the acquisition of action-outcome representations, but not for their later use. It would be worthwhile to test the effect of post-training mPFC lesions on sensitivity to contingency degradation, since it is natural to think that this area would be necessary for later modification of action-outcome states, as well as their initial acquisition. Lesions to the medial dorsal nucleus of the thalamus, through which DMS-mPFC loops pass, also produce insensitivity to both tests. Meanwhile in humans, fMRI studies have found activity in the ventromedial prefrontal cortex (vmPFC) and the anterior caudate (a part of the striatum to which it projects) to be consistent with these regions encoding action-outcome contingencies. One study (Tanaka et al. 2008) found that activity in these areas was higher when subjects were performing tasks with high reward contingencies than tasks with lower contingencies. The medial PFC also showed activity related to reward contingencies.

Of greater interest to us, however, is the neuroanatomy of desire. Balleine and O’Doherty (2010) argue that the ventral striatum, which includes the nucleus accumbens, and the basolateral amydala are centrally involved in desire in rats. The basolateral amygdala is thought to be a crucial site for the integration of sensory and emotional information, and has been implicated in representing outcome values (Balleine et al. 2003); it is also of interest because it is connected to the DMS and mPFC, offering a glimpse of how desires and action-outcome contingencies might be combined by the goal-directed system. Incidentally, the interaction between desires and action-outcome contingencies may also be mediated by spiralling feed-forward connections which project through the striatum from ventral to dorsal areas (Haber & Knutson 2010). The core of the nucleus accumbens has been found to be necessary for sensitivity to outcome devaluation in lesion studies (Corbit et al. 2001), but not to be necessary for contingency degradation, a result which is consistent with a role in desire formation, and the nucleus accumbens has also been associated with desire in debates over the role of dopamine (e.g. Berridge 2007).

In humans, the ventral striatum and amygdala are almost certainly still important for desire, but the orbitofrontal cortex (OFC), which is connected by a loop with the ventral striatum (Haber & Knutson 2010), is also of particular interest. Evidence linking the OFC with desires, and in particular *offer values*, which seem to be occurrent desires (see section 3.2), has been discussed in several recent reviews (e.g. Rangel & Hare 2010, Kennerley & Walton 2011, Padoa-Schioppa 2011). There is evidence for this view from a variety of sources. Studies of brain-damaged patients indicate that the OFC and vmPFC are necessary for normal decision-making (Damasio 1994). Single-cell recording studies in primates have found that many cells in the OFC encode both the identity and the value of stimuli, and have led researchers to conclude that the OFC is the first site at which representation of stimuli is modulated according to their values (Rolls & Grabenhorst 2008). Notably, fMRI studies on humans have found that the magnitude of OFC activity is correlated with the amount that participants are willing to pay for available goods (Plassman et al. 2007). Several other imaging studies have also found OFC activity in response to a wide range of rewarding stimuli, including attractive, smiling faces (O’Doherty et al. 2003), aesthetically pleasing paintings and musical sequences (Kirk et al. 2009), and monetary gains and erotic stimuli (Sescousse et al. 2010). This wide range of different rewarding stimuli is particularly noteworthy, since philosophers typically take our intrinsic desires to have a wide range of objects, from foodstuffs to career objectives and outcomes valued for their aesthetic properties.

3.2 Occurrent and Standing Desires

Philosophers also often distinguish between *occurrent* and *standing* desires. Drawing this distinction helps us with several potential challenges to the simple account of the goal-directed system given so far. These include: explaining certain features of individuals' behaviour patterns; fitting the goal-directed system into more general accounts of how the brain works; explaining how the goal-directed system takes into account physiological needs; and explaining how the goal-directed system sets up tractable decision problems on individual occasions. Occurrent desires are often thought of as conscious desires, but I prefer to avoid relying on claims about the phenomenology of desire, because phenomenological claims in general are hard to verify. Also, it is unclear whether our conscious experiences of desire should be expected to reflect causally-significant categories. So I will draw the distinction in functional and neurological terms.

As I introduced it at the start of chapter 2, the goal-directed system involves states that keep track of the apparent values of outcomes, based on the agent's experiences. These states, which I have called 'desires', do not arise spontaneously, motivate actions, and then dissolve once they are satisfied. This means that the goal-directed system as it has been described so far is well-suited to accounting for some ways in which desires seem to influence our actions, but not others. For instance, consider my behaviour with respect to ice-cream. In my life so far, I have sampled many flavours of ice-cream on a large number of occasions, and this has led me to have relatively settled preferences; at a well-stocked gelateria, I am now likely to order passion fruit, and unlikely to choose rum and raisin. My father also has relatively settled preferences, which are different from mine. The goal-directed system as it has been described so far can do a reasonable job of explaining this: my father and I have each learnt about how rewarding different flavours of ice-cream are over time, and because the desires we have formed in this way change only slowly, our behaviour is somewhat predictable even though ice-cream eating opportunities occur only rarely. The only mystery is why we end up with different preferences, and I discuss this point later in the chapter. On the other hand, we also need to explain why I occasionally spend my time eating ice-cream, but usually do not do so, or work to bring it about that I am doing so. Several factors are capable of contributing to such an explanation, but it is particularly noteworthy that I often do not seek ice-cream even when I know it is immediately available, and I am not engaged in any other time-constrained task. So there is reason to believe, even without considering the phenomenology of desire, that my desire to eat ice-cream is sometimes active and capable of motivating me to act, and at other times dormant.

Why this should be the case is a further question, but reflecting on that question gives us more reasons to think that desires can be either standing or occurrent. We need standing desires, which change only when we receive new information about the values of (types of) outcomes, in order to allow us to modify our behaviour appropriately to experiences which may have taken place over wide spans of time. Having occurrent desires as well permits short-term variation in desire-strength independently of this, and one reason such variation is valuable is that some outcomes are particularly good or bad under specific circumstances. For example, it is possible to learn that ice-cream is particularly good when the environment is hot, but not good when one has already overeaten. More generally, it is useful for desires to change in strength in response to changes in our physiological needs and other basic drives (Padoa-Schioppa 2011). But it is crucial that this happens without changing the strengths of our standing desires: for example, when an animal is very salt-deprived, it may be vital for its survival that its desires for salty food are very strong at that time; but if these desires for salty food are made permanently overwhelming to cope with this situation, the animal will soon suffer from excessive salt consumption. Another advantage of short-term variability might be to adapt behaviour to variation in the quality of token outcomes such as individual samples of food (Holton & Berridge 2014). Here the idea is that when one comes across a particularly good sample of its type, one should be very highly motivated to consume it, but one's standing desire should not be so strongly strengthened. Holton and Berridge suggest that dopamine is involved in producing this kind of occurrent boosting of desire, and that this could contribute to explaining drug binges.

The distinction between occurrent and standing desires also helps to explain how the goal-directed system fits into a popular general account of how the brain works. According to this general picture, the brain resolves uncertainty through competitions. For instance, if the information from the senses that the brain receives at a given time does not unambiguously tell it how things are, then instances of activity representing different possibilities will compete with one another, and what is perceived will be the situation represented by the winning pattern of activation (Clark 2013). These competitions are to some extent facilitated by the basal ganglia, as explained in section 2.3. Given that the structure of the cortex is almost the same across the whole brain, we should expect action-selection to involve computationally similar processes, and Cisek (2007) suggests that it amounts to a competition between action representations prompted by the perceived environment. In the context of desire, what this suggests is that at any one time a number of instances of activity in the OFC and ventral striatum will be occurring, each representing the value of some available outcome. This idea would make sense of the results of fMRI studies of the OFC mentioned above. These instances of activity will tend to strengthen action representations in proportion to their own strength and the probabilities associated with occurrent instrumental beliefs. In addition to this, the brain learns and stores information by changing and maintaining structural features, which determine future patterns of activity. So we should also assume that structural features of the OFC and ventral striatum, such as synapse weights, constitute standing desires.

A final respect in which the distinction between occurrent and standing desires is significant is that it helps to explain how the goal-directed system sets up tractable decisions for itself. It is relatively easy to see how the habit system responds to present circumstances; it generates action representations for those responses that are most strongly associated with present stimuli, which then compete with one another. In the goal-directed system, however, there is no obvious privileged means by which perceptual information can initiate the process of choice. Also, if too many possible actions and outcomes are considered in the course of one decision, the calculations required will quickly become unmanageable. The solution to the latter problem seems to be that only occurrent desires are taken into account, and a partial solution to the former one is that when the environment directs an individual's attention to a particular object or outcome, this causes related desires to become occurrent, and that the strength of the occurrent desire is affected by the degree of attention (Hare et al. 2011). The phenomenon of Pavlovian-instrumental transfer, in which reward-related cues prompt increased performance of goal-directed behaviours, may be a manifestation of this process. Interestingly, it is likely that this process involves a positive feedback loop, because the strength of associated occurrent desires will also tend to increase attention to features of the environment. Such positive feedback loops may help to resolve competition between neural coalitions. In addition, goal-directed choice may be initiated by the recognition of the environment's *affordances* – actions that are associated with particular stimuli not by habits, but because the stimuli show that the actions are possible. If these actions are believed to be likely to lead to certain outcomes given the circumstances, then desires for those outcomes may become occurrent.

Standing and occurrent desires therefore play different roles in action-selection, and may be expected to have different representational content. Standing desires store information about value, and intuitively represent the average levels of reward that outcomes have provided in the past, which is a reasonable estimate of the level that will be provided on arbitrary future occasions. Several factors influence the strengths of occurrent desires, including most importantly the strengths of standing desires, and they intuitively represent the levels of reward that outcomes are expected to provide on the occasions on which the desires occur.

3.3 Basic Drives

We have now made considerable progress in showing that the goal-directed system is real, and in understanding how desires causally influence action. However, there is an important part of the picture still missing. We have not yet seen how desires are formed and updated, which is of particular importance since they seem to have the function of tracking the reward values of outcomes. A central element of this process is the role of basic drives, and the nature of drives and their influence on action-selection is also an important topic in its own right. So we can now turn to basic drives, before moving on to the role of dopamine in desire-formation in the next section.

The following claim is intended as a stipulative definition of the term ‘basic drive’: an animal has a basic drive for some outcome if and only if it is innately disposed to treat that outcome as rewarding, in virtue of the successful functioning of some bodily or psychological process, and not in virtue of having a basic drive for some other outcome.[7] The reason for the second clause is that humans are innately disposed to find most, if not all, addictive drugs rewarding, but this is because those drugs ‘hijack’ reward systems, which are then not functioning successfully, rather than because we have basic drives for them. The reason for the third clause in the definition is that humans are innately disposed to treat doughnuts as rewarding, but this does not show that they have a basic drive for doughnuts,

[7] By ‘innately’, here and throughout, I simply mean ‘not as a consequence of learning’.

since it is explained by their basic drives for food and for sugar in particular (see Foddy & Savulescu 2010 for evidence relating to the basic drive for sugar). Given this definition, it is likely that humans have basic drives for food, water, sex (when mature) and positive social interactions such as smiles, among others. We also have basic aversions, for instance to pain, but I will leave these aside – to simplify matters, I attend only to the positive side of motivation and action-selection throughout this thesis.

Basic drives interact with the habitual and goal-directed systems in two main ways. First, basic drives are the primary means by which we detect reward. Humans and other animals cannot simply see, feel or taste that things are good for them, but they need information about the levels of reward that new situations provide in order to acquire adaptive habits and desires. Part of the way this problem is solved is through basic drives, the objects of which are features of the environment (and their own bodies) that animals can perceive relatively directly, and which have tended in the evolutionary past to promote survival and reproduction. In many cases, the reason why the objects of basic drives *promote* survival and reproduction is that they are *necessary* for one or both of these. So if animals have basic drives, we can account more fully for how their habits are formed and updated. I described in section 2.4 how habit-updating could be performed by reward prediction error signals, and to generate these the animal needs predictions of reward, and the ability to measure occurrent reward. Habits themselves can be used to predict reward, and we have now seen that basic drives can be used to measure it. Reward could be measured by measuring aggregate drive-satisfaction. This idea is relevant to the goal-directed system as well, because we need a way to measure reward in order to form and update desires; I discuss this issue in more detail in the following two sections.

The second way in which basic drives influence the habitual and goal-directed systems is by influencing action-selection on particular occasions. In both systems, and for almost all basic drives, this influence is dependent on learning; animals need to learn which habitual behaviours and desired outcomes are good for which basic drives. Before moving on to the evidence for this, however, we should note that this kind of influence requires something like a distinction between occurrent and standing drives. The degree to which we are motivated by the need for food must change dramatically over relatively short periods of time, because there are some times at which getting food quickly is absolutely necessary for survival. However, for the purposes of desire- and habit-learning, it may on the whole be counter-productive for there to be significant fluctuations in how rewarding we find the satisfaction of basic drives at various times. When we consume a specific foodstuff when hungry, the appropriate point to learn is not that this foodstuff is particularly good, but that it is the kind of good that is particularly worth pursuing in that state of need. So it is plausibly beneficial for us to have stable standing drives that contribute to reward learning. But it is also certain that the brain must respond to short-term changes in physiological needs (and perhaps other comparable states, like being in season for animals that have few opportunities to mate) in a way capable of influencing action at that time. So I propose that we have both standing drives, which are simply dispositions to find outcomes rewarding, and occurrent drives, which are instances of activity representing internal states of particular motivational significance.

Occurrent drives influence action in similar ways in the two systems. Regarding the habit system, Dickinson et al. (1995) showed that rats that had been extensively trained when hungry to perform an action for food would reduce responding when they were sated. This result could be explained in any of three ways: it could be that the rats' overall level of arousal was reduced; or that the reduced responding was caused by a mechanism that identified outcomes as good for particular drives; or that the hunger the rats experienced when they were trained formed an internal component of the stimulus to which they learned to respond (Niv et al. 2006). Experiments by Niv and colleagues (2006) showed that rats that are trained when sated to lever-press for sugar solution do not increase responding when thirsty, and that rats that are trained when hungry to lever-press for sugar solution reduce responding when thirsty. Sugar solution is good for both hunger and thirst, so these results seem to contradict both of the first two hypotheses. The first hypothesis, that basic drives influence habits only by influencing overall arousal, would presumably predict that thirsty rats would respond more than sated rats in the first experiment. If thirst does increase arousal, then this effect must have been cancelled out in that experiment, and one explanation of how this could happen is that the rats' representation of situation when they were thirsty was different from that when they were trained. The second hypothesis would predict again that thirst would increase responding for sugar solution, and also that thirst and hunger would both motivate the rats to perform habits leading to sugar solution, so it is undermined by both experiments. The best available hypothesis at present, then, is that states such as hunger and thirst can act as internal stimuli in stimulus-response learning – that the representations of their environments that animals come to associate with actions in habit learning include details about some internal states.

Somewhat similarly, in the goal-directed system, several studies have shown that behaviour is only modulated by occurrent drives after animals have undergone 'incentive learning'; that is, after they have experienced the outcomes involved when in the relevant drive-state. Perhaps the most famous study showing this effect, by Dickinson and Dawson (1988), also provides support independent of outcome devaluation experiments for the claim that rats represent the outcomes of their actions. In this experiment, hungry but not thirsty rats were trained to perform two different actions for different food rewards. They pressed a lever to receive food pellets, and pulled a chain to receive sucrose solution. The rats were then given the opportunity to perform these actions when thirsty. Rats that had previously consumed sucrose solution when thirsty preferentially pulled the chain in extinction, but those that had not had this experience performed the two actions equally. The fact that the thirsty, experienced rats pulled the chain more than they pressed the lever shows that they represented the outcome of this action. But for present purposes the point is that only the experienced rats showed this effect. More strikingly, similar results are found in simpler tests. Rats that have been trained to press a lever for food will not increase their performance when hungry, compared to controls, unless they have previously consumed that food when hungry (Balleine 1992; for further examples and references see Niv et al. 2006).

These results about the two systems apparently show that rats might sometimes fail to perform actions that have led to food becoming available to them in the past, even when they are hungry and no other food is available. They also suggest that rats, and presumably other mammals, might continue to desire food even when they are not hungry. This would make sense given the apparently indirect connection between hunger and desires for food, and certainly speaks to the modern human experience. These characteristics would not be seriously maladaptive in environments in which hunger was common – so that it was unlikely that any given food opportunity would be discovered only when sated – and in which overconsumption was not typically dangerous. In fact, in environments in which the latter is the case, it is likely to be adaptive for desires for food not to be dependent on hunger, since this will motivate the discovery of food resources which may be useful in leaner times. Strikingly, salt appetite seems to work in a quite different way from basic drives more generally: rats that have experienced actions leading to the delivery of salt into their mouths will perform these actions when deprived of salt, even though salt delivery has previously always been an aversive experience for them (Tindell et al. 2009). This suggests the existence of a special-purpose mechanism for tracking the presence of salt and responding to salt appetites, which would make some sense given that in contrast to other foods, salt appetite is a rare condition and overconsumption of salt is dangerous even in the short term.

3.4 Dopamine and Desire-Formation, Part 1

Part of what it means for a system for action-control to be goal-directed, as that term is used in current neuroscience, is for the inputs to the system representing the values of outcomes to change in response to evidence. So it is not necessary to examine how desires are formed and modified in order to test my claim that this process is to some extent responsive to evidence for the reward values of outcomes (part of premise III). Instead, this claim will be shown to be true if we can establish that humans have a goal-directed control system with inputs that can reasonably be thought of as desires. Nonetheless, this section and the following one will be dedicated to the question of how desires are formed and modified. I need to address this question in order to fill a gap that currently exists in my account of the goal-directed system, and because it is relevant to showing that what I am calling ‘desires’ deserve that label. More importantly, though, the issue is fascinating in its own right. Our desires do not emerge spontaneously, but are products of the ways in which years of experiences, in combination with congenital individual differences, guide us along intricate pathways in accordance with – presumably – some relatively simple basic principles. In this section, I will discuss a principle which may be central to desire-formation and desire-updating. Then in section 3.5 I will outline some significant challenges to giving a more detailed account of this principle, and to saying what role dopamine plays in its implementation.

The claim that I will explore and defend in this section is the following:

> *Core Claim about Desire Modification*: Desire modification is the result of associations between representations of outcomes and reward signals, which are generated using basic drives and desires themselves.

Very roughly, then, my desire for a given outcome will get stronger when I represent that outcome as occurring at the same time as, or shortly before, a positive reward signal occurs in my brain, and weaker when my representing that outcome is accompanied by a negative reward signal. There are several respects in which we can make this account less rough, and several reasons to think that something of the kind must be right. I will start with the idea that desires are updated by reward signals.

By *reward signals*, I mean signals in the brain the strengths of which correspond, according to the functions of the signals, with the level of reward detected in the environment at the time, which are attuned in this way to reward in general, rather than to rewards of particular kinds such as the satisfaction of particular basic drives. The sense of 'correspond' I have in mind here is fairly weak, so that if phasic dopamine signals are reward prediction error signals, as suggested in section 2.4, then they would count as reward signals. So the kinds of signals that would count as reward signals include RPE signals of various sorts (there is not just one kind of RPE signal; see section 3.5), as well as signals that simply represent the level of reward detected at the time. It is very likely that, regardless of whether dopamine is a reward signal, there are such signals in the brain, which contribute to habit learning. Habit learning requires assessing how much reward the current situation provides. So a first reason for believing that desires are updated by reward signals is that we can be confident that such signals are present in the brain. Given that such signals are present, it would be very surprising if they were not used. Another reason to think that reward signals are crucial to updating desires is that incentive learning is necessary for rats to adapt goal-directed behaviour to their physiological needs, as discussed in the previous section. This suggests that desire-formation is a process of associating outcomes with reward in general, rather than particular sources of reward, such as the satisfaction of occurrent drives.

A second point to note about this claim is that what matters for desire-modification is not what is actually happening, but what the agent represents as happening. Misrepresentation of either the outcome for which the desire is modified, or of other features of the situation, can affect the ways in which desires are changed. For example, someone who, walking in a garden at dusk, notices the pleasant scent of what looks to them like a rose will come to have a slightly stronger desire to be around roses – perhaps to have them in their own garden – even if the flower they noticed was in fact a peony. In this case, misrepresentation causes the wrong desire to be updated. But to take another example, someone who tries playing pool for the first time, has a weak opponent, and is congratulated on his performance may falsely believe that he has performed particularly well, and may consequently generate a reward signal which will boost his desire to play pool. In this case, the problem is that a misrepresentation has contributed to the generation of the reward signal, with the result that the right desire is updated in the wrong way. However, it is also important that we can represent the same situation in different ways without any misrepresentation, if our attention is drawn to different features or if we apply different representational resources. For example, if an adult and a child both sample a delicious handmade blackberry ice-cream, the adult's desire for *handmade blackberry* ice-cream may be boosted, whereas for the child the main change may be to their desire for *purple* ice-cream.

This point, that it is representation rather than reality that matters, also raises the possibility of offline desire-updating. In principle, it may be possible for humans to generate reward signals by imagining rewarding situations, associate them with imagined outcomes, and thus form and update desires for outcomes without experiencing those outcomes. It is certainly a familiar thought that we often try to work out what we want by imagining what it would be like to be in different possible situations, and this could explain the potentially puzzling datum that we often have desires for outcomes that we have never ourselves experienced (such as to become a surgeon, climb Chimborazo, or win an Olympic gold). There is also some empirical evidence suggesting that this may be possible. In a recent study, subjects were found to be able to cause increased activity in their own midbrain dopamine neurons by imagining pleasurable scenarios (Sulzer et al. 2013). Also, one study has found that artificially stimulating dopamine using the drug L-DOPA while participants imagined possible future events led to them predicting greater pleasure from those events (Sharot et al. 2009), and another that the OFC is activated by both real and imagined rewards (Bray et al. 2010).

It is also important that reward signals are generated using both basic drives and desires themselves. RPE signals can only be generated if some means is available for measuring the level of reward currently being received, and in the previous section I described how basic drives contribute to this process. In addition to this, desires themselves may also be used. If desires represent the reward values of outcomes, it makes sense for them to be used when the organism needs to know the value of outcomes that are occurring at the time. I have already suggested (section 2.5) that precursors of desires could usefully play this role in habit-learning regardless of whether a goal-directed system was present. So on this view, desires would have two roles in action-selection, one direct and one indirect, and would relate to reward signals in two different ways, being both updated by them, and involved in generating them. The phenomenon of secondary reinforcement (noticed by early behaviourists including Skinner 1938 and Hull 1943) provides evidence that desire does contribute in this way; in secondary reinforcement, actions are reinforced when they lead to outcomes that animals have been trained to find rewarding, such as lights and tones. Tim Schroeder (2004) also finds evidence that desires play this role, and argues that it is disposing us to find outcomes rewarding, rather than disposing us to pursue them, that most centrally characterises desire (for more on Schroeder's view, see chapter 5).

The proposal at hand is therefore that desires tend to be strengthened when the outcomes that are their objects occur together with outcomes for which we have either basic drives or existing desires. This proposal may raise a concern about the possibility that desires could cause themselves to be strengthened, and I will turn to this topic in the next section. But it is also noteworthy in that it may be adaptive for animals such as rats to be capable of secondary reinforcement, because it would provide one way in which they could learn to take actions that would lead to basic-drive satisfaction, but only through relatively lengthy processes; developing intrinsic desires for means is one alternative to developing the ability to perform multi-step means-end reasoning. A related point is that on this proposal, we can better explain how humans come to have desires for outcomes that are apparently far-removed from the satisfaction of basic drives.

The core claim about desire modification is attractive for three further reasons. First, this account of how desires are formed and updated seems to get the level of responsiveness to reasons in these processes about right. It does not suggest that we desire at random, but instead explains our desires as products of a somewhat crude system, shared with many other animals, for getting us to desire things that tend to contribute to the satisfaction of our basic drives. If at least this level of responsiveness to reasons in desire-formation was not achieved, it would not be plausible that having desires at all would be more adaptive than lacking them. On the other hand, it also avoids the mistake of suggesting that our desires will always track our conscious, explicit judgments about what is good for us, biologically or otherwise. This is clearly not the case; for instance, someone who believes their doctor when they are told that eating cheese is bad for their health will not immediately lose their desire to eat cheese – they may never lose it. This point shows that desires are only responsive to evidence when it is made available in the right form. Getting sick after eating cheese may be less good evidence that it is unhealthy than the doctor's testimony, but more effective in changing our desires. Also, it can often be hard to tell why we have the desires we do, and why our desires change (although not difficult to come up with plausible candidate explanations), and this shows that desires do not change as a result of conscious, explicit reasoning, but instead by some process of which we are only partially aware. On a related point, the origins of our desires often seem to reach back deep into our pasts, as the present account would predict, since its recursive form implies that desires may often be the products of complex and gradual processes of development. For example, I like to hike in the mountains, and although I can give the 'desirability characterisation' (Anscombe 1957) of this activity that it feels adventurous, I can only speculate about why adventure is attractive to me.

Second, if dopamine is the reward signal that updates desires, then the account has the advantage of offering an attractive explanation of drug addiction. By various mechanisms, alcohol, nicotine, cocaine, opiates and amphetamines all boost the strength of dopamine signals. This means that if the present hypothesis about dopamine is right, then drug addiction is apparently the result of 'hijacking' of the system for desire-formation (Hyman 2005, Holton & Berridge 2014). On this view, because dopamine signals are reliably caused by addictive drugs, regardless of whether they are more rewarding than expected, the strength of standing desires for addictive drugs grows with every hit. In the presence of drug cues, therefore, addicts will experience extremely strong occurrent desires to take drugs. This explanation of addiction has some independent advantages over its rivals, such as that drug-seeking behaviour is flexible and sophisticated, and consequently harder to explain in terms of distorted habits than distorted desires (Berridge & Robinson 2011, Holton & Berridge 2014). It therefore provides indirect evidence for the idea that dopamine updates desires, which would fit the present picture. Unfortunately, as we will see in the next section, there is considerable uncertainty about dopamine's role.

Third, the account can explain how we come to have the wide range of intrinsic desires that we typically attribute to one another. It is perhaps surprising that a mechanism that we share with rats and mice could explain our subtle aesthetic preferences and often abstract, specific ambitions, but the present account puts no limits on what can be desired except the agent's representational capacities. Humans may be expected to have many, varied and sophisticated desires, given the following several points: we have sophisticated representational capacities; we can imagine outcomes in detail without having experienced them before, and we often exercise this ability; it is plausible that we have basic drives for social status and evidence of approval by those around us; and we live in remarkably rich and complex cultures. It is also worth noting that capitalist culture includes practices designed to exploit weaknesses of our desire-formation systems, to motivate us to work for things we may or may not otherwise desire. The point is that when we add these features of human psychology to the simple general-purpose system the present account describes, it is likely to result in our having many and varied desires. This is important because it supports the claim that what I am calling 'desires' really are desires, if they are likely to take the same sorts of objects.

I endorse the core claim in its present, rather vague form. Its main deficiency is leaving some important and basic questions about desire-modification unanswered, as I will describe in the next section.

3.5 Dopamine and Desire-Formation, Part 2

The core claim about desire-modification outlined and defended in the previous section leaves two kinds of questions unanswered. First, it says nothing about the implementation of the desire-modification system in the brain – for instance, about whether the reward signal used to update desires is dopamine or something else. Second, the account it gives of how, in computational terms, the reward signal is generated and used to update desires is really too vague to be satisfactory. Unfortunately, however, the current state of neuroscientific research on these topics does not permit the development of a more detailed account, largely because of uncertainty about the function of phasic dopamine signals. I will first discuss the debate about the function of phasic dopamine, then turn to computational issues.

So far, we have seen that phasic dopamine signals may constitute RPE signals, and be used for habit learning. The evidence supporting this view includes Schultz' classic studies showing that phasic dopamine signals transfer from occurring after rewards are delivered, to occurring after reward cues, and computational, anatomical and physiological evidence that if dopamine signals were RPEs, they would be suitable for updating habits (see section 2.4). We have also seen that the effects of drugs of abuse on dopamine levels are apparently responsible for their addictive nature, a point which indirectly supports the idea that dopamine signals update desires. However, the function of phasic dopamine signals has been a topic of intense research interest among neuroscientists for many years, generating a very large body of literature, and this literature remains inconclusive. Among the experimental results produced are some that appear to give fairly strong support to the idea that dopamine updates desires, and others that seem to contradict it.

For example, studies show that the effect of blocking dopamine transmission on instrumental behaviour is similar to the effect of removing the reward (sometimes called 'extinction mimicry'; see Wise 2004, Berridge 2007). If a rat is trained to press a lever to receive a food reward, and then allowed to press the lever without receiving the reward, then the rate at which it presses the lever will gradually fall. If it is given the same training, then allowed to press the lever after having been given a drug that blocks dopamine (called a 'neuroleptic'), but with food still available, the rate of lever presses will also fall gradually. So the absence of dopamine, which according to the theory would be interpreted as a negative RPE signal, seems to teach rats to abandon learnt actions; this result could be explained if dopamine was used as a reward signal in either habit-learning or desire-updating. In another experiment, blocking dopamine receptors specifically in the nucleus accumbens prevented rats from developing approach behaviour towards initially neutral cues that were paired with rewards (Parkinson et al. 2002), which seems to favour the view that desires are updated by dopamine signals. Also, the phenomenon of transreinforcer blocking (Burke et al. 2007) supports the view that desires are updated by prediction error signals. In the first phase of this procedure, a rat is trained to associate a neutral stimulus, such as a tone, with a rewarding outcome such as delivery of a food pellet. This training is sufficient to make the tone a conditioned reinforcer; afterwards the rat will perform an action such as lever-pressing in order to hear the tone. Then in a subsequent phase, the tone and another neutral stimulus, such as a light, are presented simultaneously prior to the delivery of a food pellet. Although this combination of events is repeated, the light does not become a conditioned reinforcer for the rat; learning about the value of the light is 'blocked', because the light only ever leads to a reward that is predicted by the tone.

On the other hand, Wassum and colleagues (2011) found that allowing rats to consume sucrose after a 23-hour period of food deprivation led them to pursue that food more vigorously on a future occasion, when they were not so food-deprived. The apparent explanation is that the rats came to desire sucrose more, having experienced it in circumstances in which they were likely to find it particularly good. But the significant result for our purposes is that the experimenters conducted this test both on controls, and on rats that were treated with a dopamine-blocking drug at the time at which they consumed the sucrose when food-deprived, and found that the drug had no effect. The rats that had lacked dopamine when they had the opportunity to update their desire for sucrose had apparently strengthened it just as much as the untreated rats, suggesting that dopamine is not necessary for desire-updating.

As philosophers, we are not in a particularly good position to assess the significance of these and the many other relevant studies. But this brings us to a more serious problem for assessing the claim that phasic dopamine signals update desires, which is that this claim is not one of the main alternatives under discussion in the debate about what these signals are for. This means that we cannot benefit from the expertise and cumulative efforts of the neuroscientific community in evaluating it, which strongly suggests that we should suspend judgment on this claim.

Instead, the two most prominent views in the debate about the function of dopamine are that it is an RPE signal for updating habits, and that it is for promoting action at the time of release, in response to cues that predict the availability of reward. The latter view is advocated by Kent Berridge and his colleagues (e.g.

Berridge 2007), who use the slogan that dopamine is for ‘wanting’ rather than ‘learning’ or ‘liking’ (they themselves use the scare quotes). Berridge’s view is that the classic experiments by Schultz described in section 2.4, which show that dopamine signals transfer from occurring subsequent to rewards, to occurring subsequent to reward cues, do not distinguish between the hypotheses that dopamine is an RPE signal for learning and a signal with the function of producing immediate motivation to exploit rewarding situations. He favours the latter view primarily because experiments on mutant mice suggest that the absence of dopamine does not wholly prevent habit learning (Cannon & Palmiter 2003, Hnasko et al. 2005), and that higher than normal levels of dopamine do not seem to boost habit-learning (Cagniard et al. 2006, Yin et al. 2006). Berridge does endorse the idea that dopamine updates desires in his paper on addiction co-authored with Richard Holton (Holton & Berridge 2014), but there is little discussion of this view in his papers in scientific journals. Meanwhile, neuroscientists who are positive about the suggestion that dopamine updates habits describe the idea that it also updates desires as anatomically attractive but computationally puzzling (Balleine et al. 2008); I will turn to some such computational puzzles next.

There are three main respects in which the algorithm used for updating desires is underspecified by the core claim of the last section. First, the notion of a reward signal is deliberately vague. It is not specified whether this is a signal representing perceived reward or an RPE, and this issue is compounded because RPEs of different kinds are apparently suitable for updating habits and desires, respectively.[8] Second, there is a problem about how basic drives and desires should combine in measuring received reward. Third, given that we experience more than one outcome at any one time – and indeed that this may be a necessary condition for desire learning – it is not obvious exactly how reward signals should influence desires for outcomes represented at the time, and this appears to depend on the nature of the reward signal. I will give some more detail about these problems and about potential solutions, but my aim is not to defend any particular account. My view is that these are challenging but not insoluble problems, and that to solve them – that is, to give a more satisfactory account of desire modification – will require the full range of

[8] This point was brought to my attention by Nick Shea.

techniques of modern neuroscience, including computer modelling and simulation as well as behavioural, anatomical and physiological studies. Because the problems interact with one another, I will describe all three before discussing potential solutions.

On the nature of the reward signal, matters would be relatively simple if the same kind of signal was suitable for updating both habits and desires. In that case, the only obstacle to concluding that phasic dopamine signals perform both of these two roles would be some awkward experimental results. However, the evidence suggests that phasic dopamine is some form of RPE signal, and RPE signals of different kinds are, at least on the face of it, required for the two roles. To update habits correctly the signal required is one representing the difference between the level of reward currently being experienced and the average level of reward that *the action just performed* has led to in the past, whereas to update desires correctly the signal required is apparently one representing the difference between the level of reward currently being experienced and the level of reward that *the current outcomes* have provided on average in the past. Because of this difference, an agent using one signal for both purposes would be subject to certain systematic failings. In particular, an agent which used an RPE signal adapted for habit-learning to update desires might dramatically overestimate the reward value of outcomes that are often delivered unpredictably.

For example, imagine a creature for which strawberries are rewarding, which often discovers them when foraging, and suppose further that it calculates RPEs by subtracting the reward level predicted by its actions from the level received. On typical occasions on which it discovers strawberries, this creature will produce a positive RPE signal, because discovering strawberries is a better-than-average outcome from any given moment's foraging, and we are assuming that it is difficult to predict. Also, because the creature's desire for strawberries will itself be used in working out how much reward it is receiving at any given time, it will rarely produce negative RPEs when discovering strawberries, even if it expected to find them, and hence predicted a very high level of reward (such negative RPEs would only occur when discovering strawberries was concurrent with some aversive outcome). So the strength of the creature's standing desire will enter a runaway feedback loop, spiralling upwards just in virtue of the fact that strawberries are somewhat satisfying to basic drives and that they are often discovered unexpectedly.

It is worth noting that this is not the only form the problem could take; if the same creature expected to receive strawberries as the result of some outcome, but actually received a similarly-rewarding novel foodstuff – say melon – it would not produce a strong positive RPE signal, and would therefore miss out on the opportunity to learn that melon is a worthy object of desire. It is possible that we do tend to overvalue rewards that are hard to predict – this might explain why we are so attracted to gambling and following sports – but that is only one of a number of possible responses to this challenge.

Moving on to the next issue, it is in any case unclear how desire-updating avoids runaway feedback loops. Consider the desire to eat strawberries, and assume that strawberry-eating satisfies some basic drives. Assuming that there is a good way to measure the level of reward currently being received, it would make sense for the strength of this desire to be updated by a signal representing the difference between this level, and the average level of reward that eating strawberries has provided in the past. This is because the strength of the standing desire for strawberries should be proportional to the average level of reward the agent has received in the past from strawberries, and a running average may be maintained using the following formula:

$$v_t = v_{t-1} + \frac{1}{n}(r - v_{t-1})$$

where v_t is the new average, $v_{t\text{-}1}$ the old average, n the number of experiences of eating strawberries including the present one, and r the level of reward measured at t. So it looks like all we need to keep this desire at the right level is for it to get stronger or weaker according to the value of a signal corresponding to ($r – v_{t\text{-}1}$), and to change less in response to this signal on each subsequent occasion (Holton & Berridge 2014).

However, combined with the idea that reward is measured using both basic drives and existing desires, this proposal runs into trouble. For suppose that strawberries provide about the same amount of basic drive satisfaction on each occasion – call this value a – and that r is found by adding this level of drive satisfaction to the level of reward that the agent represents eating strawberries as providing. Then on each occasion when the agent eats strawberries we will get:

$$r = a + v_{t-1}$$

because the level of reward r that they perceive will be the sum of the reward they from drive-satisfaction and that which they get from having their desire to eat strawberries satisfied. Substituting in, we get the result that:

$$v_t = v_{t-1} + \frac{1}{n}a$$

and hence that the strength of the agent's desire for strawberries after n occasions will be given by:

$$\frac{a}{1} + \frac{a}{2} + \frac{a}{3} + \cdots + \frac{a}{n}$$

assuming that the initial strength of the desire is 0. But this is a problem, because this series – called the Harmonic Series – is divergent. So if we assume that strawberries will always satisfy the agent's basic drives to the same degree, the straightforward view described here implies that the desire for strawberries will get stronger every time they are consumed, with no limit to its potential strength.

Turning now to the third problem, we are assuming that some outcomes are desired not because they satisfy the agents' basic drives, but because they are associated with other desired outcomes. This suggests that agents must represent outcomes as occurring together, as we would expect – it is normal for more than one thing that one desires to be salient at a time. But this raises a range of questions about how reward signals should influence the strengths of desires for multiple simultaneous outcomes. For instance, if an outcome previously taken to be mildly aversive occurs at the same time as one that is desired, it is plausible that the reward signal should affect the valuations of each of these two outcomes in opposite ways – the former should become more liked, and the latter less liked. This does not fit the idea, which has so far been assumed, that one reward signal is produced at a time, and positive reward signals strengthen desires and negative ones weaken them. It is this kind of case in particular that I will from now on refer to as the 'third problem'; it is the problem of how to use one reward signal to update multiple desires of different current strengths. But there are also other difficulties in the area: what if it is obvious that one of the outcomes being represented, and not the others, are responsible for some basic drive's being satisfied – will the way desires are updated reflect this? And another complication is that the desirabilities of some outcomes seem to be dependent on one another, while others are independent – for instance, for most people the value of having butter depends a lot on whether one has bread, but the value of having chocolate is not dependent in this way on other outcomes.

A final complication which I will not attempt to address is that experiencing pleasure is plausibly both a consequence of desire- and drive-satisfaction, and itself an object of desire; experiencing pleasure is something we can represent, which is likely to be associated with reward.

I now turn to some potential solutions to these problems. First, there is a simple solution to the second problem, which is for all desires to gradually lose strength over time, to compensate for the inflationary effect of using desires themselves to measure received reward. This weakening would only need to be gradual to make up for the effect described, because although the harmonic series is divergent, it grows slowly. The weakening would apply equally to all desires, and would have the potentially maladaptive effect of weakening desires for objects that were rarely experienced, and hence not subject to problematic inflation, as well as those for desires that were experienced frequently. So in principle it could cause standing desires to be lost, even when their objects had only ever been found to be positively rewarding. Our everyday experience arguably shows that this is possible.

The first and third problems require somewhat more complicated solutions. I will describe two ways in which desire-updating could go, each of which seems to deal with (although not necessarily *solve*) both problems. These two ways that desire-updating could go are distinguished by whether reward signals update desires for outcomes that are represented *at the time of the signals*, or for outcomes that are represented *shortly before the signals*.

If the reward signal updates desires for outcomes that are represented at the time of the signal, and this reward signal is an RPE signal of the kind suited to updating habits, then the system will face the first problem: the agent will come to have disproportionately strong desires for hard-to-predict outcomes, and may also make some other 'mistakes' in updating their desires. Also, there will sometimes be situations in which two outcomes are represented as occurring at a time, which ought to be updated in opposite directions, but which will be modified by the same reward signal. These two problems could both be solved, however, if the reward signal did not represent a reward prediction error, but instead the level of reward received at the time. This signal could then be used to update each desire individually, with desires being strengthened if the signal represented that a greater level of reward was currently being received than the average for that outcome in the past, and weakened if the signal represented a lower level of reward. The same signal would cause some desires to become stronger, and others weaker. This is in contrast to the way that RPE signals are thought to update habits, because under that hypothesis any positive RPE signal will strengthen the habit, regardless of its current strength.

This approach would solve both problems for the same reason: in effect, desires would be updated according to the difference between the level of reward being received at the time, and the level that the outcome concerned had been associated with in the past. The most obvious apparent weakness of the approach is that if several rewarding outcomes were experienced at the same time, desires for them all might be strengthened considerably, even though none were better than they had been in the past (this assumes that the reward signal is calculated by adding the levels of reward that each feature of the situation is taken to provide). But this is simply another version of the same inflationary effect that I described as the second problem, and can be solved in the same way. It is important to bear in mind, too, that if desires are to change at all, they should get stronger when the desired objects are associated in experience with other rewards. It is a mistake to think that desires should not change when outcomes are *in themselves* the same as on past occasions – what matters for desire-updating is associations between outcomes. A less obvious potential weakness is that this kind of system may require a relatively sophisticated neural implementation.

Finally, the second potential solution is that desires might not track the levels of reward that occur concurrently with outcomes, but instead the levels of reward that those outcomes predict. This would mean that they could be updated directly by the same kind of RPE signal that is suitable for updating habits. To learn how much reward an outcome predicts, the necessary signal is one representing the difference between the level of reward that occurs on occasions subsequent to the outcome's occurrence, and the average level of reward that has followed it in the past. If we think of actions as a special kind of outcome, this is a generalisation of the RPE signal that works for habits (see Barto 2007 for a more detailed formal treatment). This approach would solve the first problem, because while unpredicted rewards would boost the strength of desires for the features of the states of affairs that preceded them, these would be weakened again when the rewards failed to occur on subsequent occasions. It would not solve the third problem, because the desires for all of the outcomes that were salient prior to a given reward signal would be updated in the same way. But evidence for transreinforcer blocking, described above, seems to show that the desire-updating mechanism has this very flaw. In transreinforcer blocking, one outcome which is desired and another which is neutral simultaneously occur prior to the same reward, and neither is strengthened, even though it would plausibly be adaptive for the desire for the neutral outcome to be strengthened while the other is left alone.

An interesting feature of this approach is that it would mean that desires are, or are similar to, *Pavlovian values*. Pavlovian values are psychological states which are usually thought of as guiding attention and unlearnt behaviours, such as approach behaviours. Consider a rat which learns to approach a certain corner of its cage when a light there turns on, because this is often followed by the delivery of a food pellet. This behaviour could be explained by either the habit system, or the goal-directed system – the rat might have learnt either that approaching the light tends to be rewarding; or that it tends to lead to getting food, which is rewarding. But it could also be explained in the following way: that rats have an unlearnt tendency to approach stimuli which predict reward (e.g. things that look or smell like food), and the rat has learnt that the light predicts reward. This would be an instance of Pavlovian learning (as in the case of Pavlov's dogs, the animal now performs an unlearnt action in response to a new stimulus) and would involve the acquisition of a new Pavlovian value. Pavlovian values may also have various important parts to play in the habitual and goal-directed systems (Balleine, Daw & O'Doherty 2008). Apparently, Pavlovian values are like desires in that they are supposed to track something about the reward values associated with features of states of affairs, but unlike them in that they do this for a different purpose; desires are primarily for assessing possible outcomes in the goal-directed system, whereas Pavlovian values are for guiding responses to perceived stimuli (Balleine & O'Doherty 2010).

The most obvious apparent weakness of this approach is that it would mean that outcomes that were consistently represented concurrently with very strong reward signals would not come to be desired. But this effect would be alleviated if the outcomes persisted through time, or were represented in anticipation by the agent. More generally, it is somewhat counterintuitive that we would work for outcomes that predict reward, rather than those that occur concurrently with reward, and therefore might be thought of as constituting or providing rewards. But it is important to bear in mind that on this view satisfying existing desires would still cause reward signals to be produced – desired outcomes would still be rewarding in themselves. It is also hard to think of a class of outcomes that we desire because they occur together with, rather than predicting, the satisfaction of our basic drives.

In my view, the arguments and evidence of this and the previous section show that although we are in a position to give a rough account of how desires are formed and updated, we are not yet able to fill in the details of that account. However, even the rough account helps to support the idea that the inputs to the goal-directed system are appropriately thought of as desires, since they behave in the kinds of ways, and may take the kinds of objects, that we usually associate with desires.

3.6 Diagram Showing Findings So Far

In chapters 2 and 3 my aim has been to present a compelling overall account of the habitual and goal-directed systems, while focusing on those details that will be most important later. In this section, which concludes chapter 3, I present some of this information in a different way: figure 2 shows the architecture of the goal-directed and habitual systems, as described in chapters 2 and 3.

There are several points to notice. It is important to note that the diagram is based on the assumption that desires are like Pavlovian values, as suggested towards the end of the last section, and are updated by RPEs in the form of phasic dopamine signals. However, this should not be taken as an endorsement of this view (see section 3.5). Also, the green arrow from ‘Reward Prediction Error’ to ‘Action-outcome contingencies’ reflects the view that dopamine signals are necessary for action-outcome learning. This view is not universal, but is defended by Horvitz

(2009), and I include it for the sake of showing a possible mechanism for the acquisition of these states.

The red and blue arrows show the routes by which action is produced by the habitual and goal-directed systems, respectively. The habitual system is activated by environmental and internal states – such as occurrent basic drives – which are associated via habits with particular responses, and tends to cause the production of these responses. Or to put it a different way, the habitual system uses information about current environmental and internal states to directly assess the present values of a range of actions, and these value-representations are then consumed by the action selector. In the diagram, the 'Action Selector' box is a place-holder for the mechanism by which the outputs of the two systems are combined for the ultimate determination of action; see chapter 4 for more discussion of this mechanism. The action selector sends information about what action is to be performed to the system which generates RPEs.

Meanwhile, the goal-directed process (blue arrows) can start in any of three ways: either internal states or perception of the environment can trigger occurrent desires, or the environment can cause the recognition of affordances. Occurrent desires and instrumental beliefs (representations of action-outcome contingencies) then interact to establish how to satisfy occurrent desires and what the other consequences of these actions would be, and/or to assess the consequences of the actions afforded by the environment. New desires become occurrent, and existing occurrent desires change in strength, in the course of these processes. These processes ultimately result in values being assigned to a range of actions, and again these representations of action-value are consumed by the action selector.

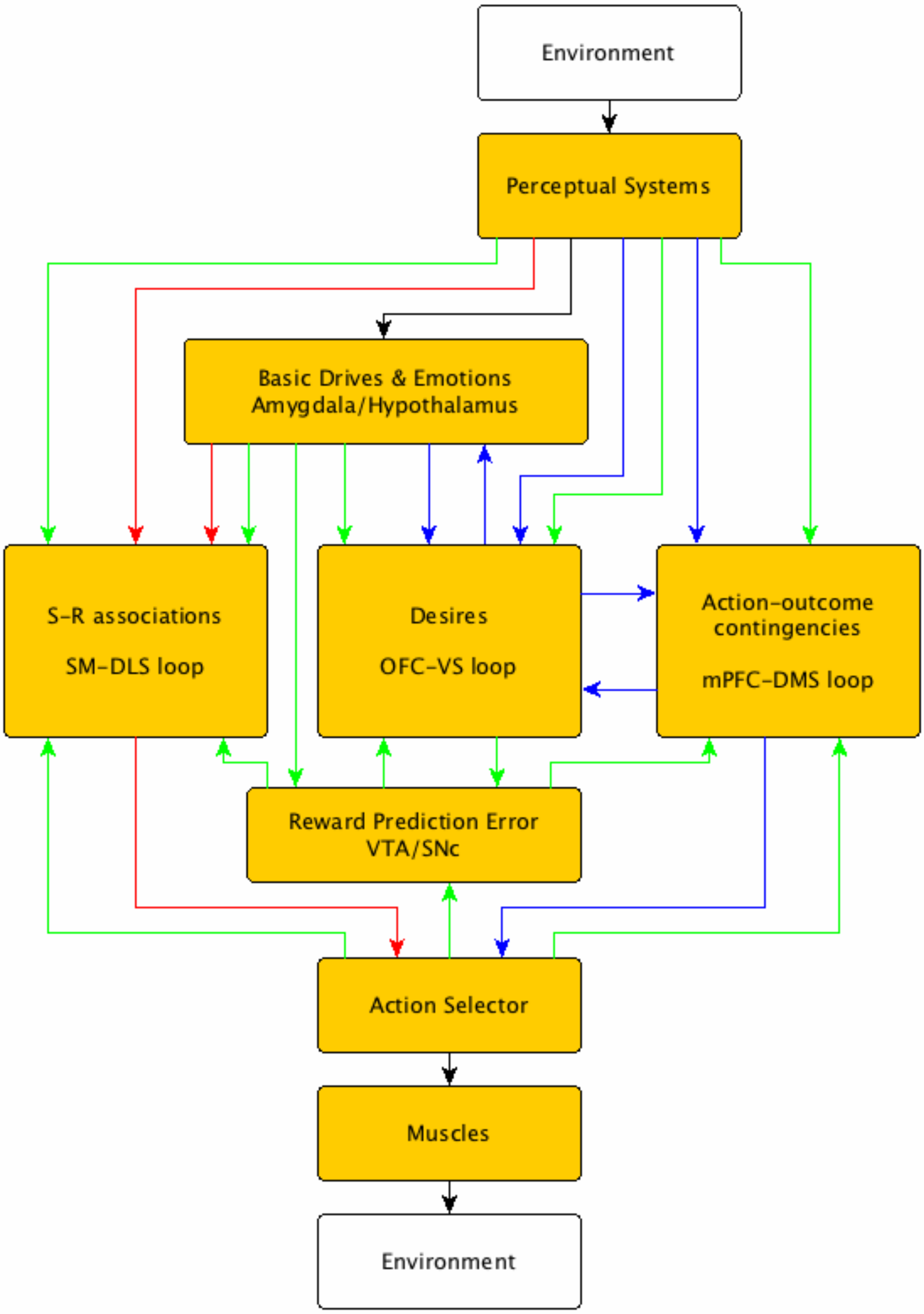

Fig. 2. A cognitive architecture for action learning and selection. SM = sensorimotor cortex; DLS = dorsolateral striatum; OFC = orbitofrontal cortex; VS = ventral striatum; mPFC = medial prefrontal cortex; DMS = dorsomedial striatum; VTA/SNc = ventral tegmental area/substantia nigra pars compacta. Different coloured arrows represent projections that form parts of distinct systems. Red arrows are for the route by which the habitual system drives action; blue arrows are for the goal-directed control of action; and green arrows are for connections that are not directly involved in action-selection, but are crucial for learning. Black arrows are for connections that are essential to more than one of these three processes, or are not integral to any of them.

In this chapter my topic has been the neuroscience of desire and of goal-directed control. Among the most important points that have been established are:

- that there is good evidence that human desires are located in the orbitofrontal cortex and ventral striatum, including some which is independent of my claim that desires are the value-representing inputs to the goal-directed system;
- that for several reasons, it is important to distinguish standing and occurrent desires, and that it is occurrent desires, which are modulated in strength by many occurrent factors, as well as by their corresponding standing desires, that directly influence actions;
- that basic drives, including physiological needs, influence action in the moment by acting as internal stimuli for habits, and causing the formation and affecting the strength of occurrent desires;
- that basic drives and existing desires both also affect future action by contributing to the formation of reward signals;
- and that desire-formation and -updating takes place when representations of currently-occurring outcomes are associated with reward signals.

Together with the work of chapter 2, the results of this chapter make a great deal of progress in support of the three empirical premises that I am aiming to establish. However, I will not give my final word on these premises until the end of chapter 4, because we are yet to consider how the habitual and goal-directed systems interact with one another, and crucially how they fit into an overall account of human action-selection. We cannot conclude that the inputs to the goal-directed system are desires without examining from a wider perspective what roles they do and do not play in determining human behaviour. That is the topic of the next chapter.

Chapter 4: Sources of Motivation and Action

4.1 Action-Selection by the Two Systems

In this chapter, I will address two questions that take a wider perspective on the systems that determine how we act. First, a missing piece of the picture outlined in chapters 2 and 3 is an account of what determines how animals with two systems for action-selection choose what to do, given that these two systems will not always recommend the same action. So I will consider various proposals that neuroscientists have made on this point. Second, I will discuss whether there are any other systems that are capable of driving action and motivation in humans and other animals, and if so, how they are related to the goal-directed system. This will involve considering a number of questions that have been discussed extensively by philosophers, including whether emotions or evaluative beliefs are capable of motivating action and how desires and intentions are related, but I will aim to give empirically-motivated responses. My approach will be to try to sketch answers to these questions that fit well into the overall picture of action selection that I am in the process of developing. To my knowledge, no scientific consensus exists about either of the two issues I will discuss in this chapter, although there are many interesting theories and results that bear on them. One point that bears on both questions is that neuroscientists typically assume that both systems work stochastically, meaning that the action predicted to be the most rewarding is not always taken – instead, it has the highest probability of being selected.

Since this is the last of my three chapters on the empirical evidence relevant to my overall project, I will close by explaining how the evidence, theory and arguments I have presented support the three premises that I am aiming to establish in part I. Chapter 5, the last in part I, will be dedicated to showing why my account of the nature of desire is preferable to that developed by Tim Schroeder (2004, Arpaly & Schroeder 2014).

Regarding the first question, it is not known whether there is some process that determines which of the two systems controls action, or whether the two systems merely assign values to actions, which are used by a downstream mechanism to determine what is done. The difference between these two options is illustrated by cases in which three options are salient, with the following properties: one is strongly favoured by the goal-directed system but strongly disfavoured by the habitual system; the second is equally strongly favoured by the habitual system and disfavoured by the goal-directed system; and a third is moderately favoured by both systems. If the process of action selection was delegated to one or other system – the former option – then one of the first two 'controversial' actions would be most likely to be selected. But if a mechanism existed that could amalgamate the results of the two systems – the latter option – then a compromise might be most likely.

One prominent account of how actions are selected, of the latter type, is the *affordance competition hypothesis* (Cisek 2007). Although Cisek does not frame his view in terms of the goal-directed and habitual systems, it is distinguished by two features that are relevant to the present discussion. These are that distinct 'stages' of action selection actually take place in parallel, and that action selection is achieved by a competition in which all represented aspects of a possible action are involved. As we interact with our environments, new affordances are continually being identified (primarily by the dorsal visual stream). The frontal and parietal cortices then go to work on these action possibilities, calculating both what their results would be, and how to perform them. Multiple sources of information are simultaneously recruited: the mPFC and OFC might model the actions' consequences and the levels of reward they would provide, while the anterior cingulate cortex (ACC) models the costs of performance, and habits represented in the DLS bias activity in the sensorimotor cortex. Actions that are favoured, or even merely salient, become associated with strong activity across these networks, while less favoured actions are inhibited. When some threshold level of activity is reached, the action that has reached it will be performed. A significant attraction of this picture is its coherence with the account of action selection by the basal ganglia given in section 2.3. It is also a flexible hypothesis; for instance, we could drop the claim that only affordances get the process of action-selection started in favour of the view that occurrent desires prompted by the enviroment can also play this role, without compromising the central claim that action-selection is achieved by parallel processes generating competing neural coalitions.

The affordance competition hypothesis might be thought to make the unintuitive prediction that we cannot choose our goals prior to considering how to act to achieve them. The intuition that this is possible is supported by a study by Wunderlich and colleagues (2010), who found neural evidence of choice between possible outcomes, before it had been revealed what actions could produce those outcomes. From one point of view, this is utterly unsurprising; it seems obvious that we can choose whether we would prefer an apple or an ice-cream without considering what it would take to get them. Some opponents of the affordance competition hypothesis (e.g. Padoa-Schioppa 2011) take the view that at least a significant proportion of action selection is achieved by selecting the most rewarding outcome, then considering what action to perform to obtain it. This process differs from that proposed by Cisek in that it divides action selection into stages, to be accomplished in series. It would also involve the selection of only one action by the goal-directed system prior to any competition with the habitual system, assuming that this competition does not take place before consideration of specific actions. However, we should take care to distinguish the capacity to run some processes relevant to action selection off-line, which is what we need in order to choose in principle between an apple and an ice-cream, from the separate capacity to use prior deliberation to constrain future action-selection processes. The latter would present a more serious challenge to the hypothesis, but is less obviously a real human trait (although it certainly seems that prior deliberation can *bias* future decision-making).

A further prediction of the affordance competition hypothesis is that there is no dedicated neural system for arbitrating between the goal-directed and habitual systems. This is in contrast to a proposal by Daw, Niv and Dayan (2005), who argued that whichever system was more likely to be accurate at a given time would control action. They suggest that a particular brain area (the authors mention the ACC and the infralimbic cortex) would be responsible for tracking the uncertainty associated with predictions made by each system, and delegating control to the system exhibiting the lower uncertainty. Typically, the habit system would be uncertain regarding relatively novel circumstances and actions, since it learns slowly, while the goal-directed system would be uncertain regarding distant outcomes that require the construction of the most extensive models. As well as being an account of how the two systems might interact, the hypothesis put forward by Daw and his colleagues also says something about their relative advantages. Another example of a proposal of this sort is made by Keramati, Dezfouli and Peray (2011), who suggest that the habit system has an advantage in producing fast responses, but that the goal-directed system is better at making accurate choices in changing environments (cf. Sterelny 2003).

A further advantage of the approach taken by the affordance competition hypothesis is that it makes no difference to the theory if there are more than two systems involved in controlling action in humans or other animals. On that approach, the goal-directed and habitual systems are used in parallel to evaluate salient possible actions, with the evaluations provided by each of these systems contributing to the 'scoring' of a competition between those actions. It is entirely compatible with this approach that the 'scoring' of the competition could also be affected by other factors, and I will argue in the next section that several other factors do indeed influence action. There are also some other respects in which the ideas I will put forward there fit well with this approach, so I will assume that it is broadly correct, and that there is no higher-level system responsible for delegating control of action-selection to one or other of the two systems.

4.2 Further Sources of Motivation and Action

In this section, I will discuss three possible further sources of motivation and action. Humans and other animals may also be motivated by a drive to explore; by their emotions; or by evaluative beliefs, plans and intentions. In each case, it is plausible that there are processes or systems that are capable of causing action independently of the habitual and goal-directed systems, or at least of having a significant influence on which actions those systems select. As well as these three, there are also reflexes, which I will not discuss. In pointing out these possible sources of motivation and action, one of my aims is to show the variety of resources that an empirically-informed approach can draw on to explain human action. Philosophers take a particular interest in potentially puzzling aspects of human behaviour such as weakness of will, 'desiring the bad' (Stocker 1979), and playful or apparently purposeless action, and an important challenge for accounts of action-selection is to show how these can be explained, given that the mind has largely been shaped by natural selection. My main aim in this section, however, is to further clarify my view of desire by contrasting desires with other psychological states that influence action.

In order to maximise the amount of reward that they receive from their environments, agents typically need to engage in both *exploration* and *exploitation*. For many animals, exploration might involve going to new places, approaching, smelling, tasting or manipulating new objects, and behaving in new ways in interacting with conspecifics. The value of this sort of behaviour is obvious: it allows animals to learn about their environments, the consequences of their behaviour, and the reward values of outcomes. However, it is not obvious how such behaviour is generated. Both the habitual and goal-directed systems seem to be suited to producing exploitative behaviour; they cause the actions with the greatest expected reward, of those that the agent already has some experience of. But it would be wrong to conclude that these two systems are incapable of causing exploratory behaviour, for two reasons. First, if it is correct that both systems operate stochastically, it is possible for them to cause actions that are not represented as being particularly rewarding, and this would also allow them to cause actions even in novel environments, as long as possible actions can be identified. Second, exploratory behaviours are only worth performing if their probable reward value, over the long term, is higher than alternatives. For instance, the choice to abandon the exploitation of one food source in order to search for another should only be taken if (simplifying somewhat) the value of the new source, multiplied by the chance of finding it, is greater than the value of the known source. So the distinction between exploratory and exploitative behaviours is not as clear as it might at first appear, and an ideal goal-directed system would be capable of recognising the value of exploration. In practice, it may be that the property of *being novel* is associated by the goal-directed or habitual systems with some value, which could change as a result of learning, and which allows exploratory actions to be chosen above exploitative ones.

However, there is also another way in which exploratory behaviour could be generated, which relies on the plausible idea that for an action to be selected, the activity-strength of a representation of that action just needs to pass some absolute or relative threshold. If this is correct, and it is also the case that the attentional salience of actions, outcomes and stimuli consists in the level of neural activity that is dedicated to them (Ruff 2011), then in principle actions could be selected just in virtue of capturing our attention to a sufficient extent. One might say that we perform them because they are fascinating, rather than because we are in the habit of doing so, or because we believe they will lead to outcomes we desire. A process of this kind could lead us to perform actions with very little knowledge of their likely outcomes, because either the actions themselves or the objects we might act upon, or places we might go, attracted our attention. It might also motivate us to act without employing any representations of reward values associated with our actions. Such a process may contributing to explaining a wide range of phenomena, from many of the behaviours performed by human infants to the disconcerting motivation we sometimes feel to explore dangerous objects and places – to jump on to the railway tracks at stations, for instance.

I suggested in section 3.2 that occurrent desires are influenced by attention, and it is very likely that we are innately disposed to attend to what is novel. So whether or not the goal-directed system learns about the value of pursuing the new – and there is no obvious reason why it shouldn't – it is likely that the goal-directed system has an in-built bias to promote exploration. In my view, then, both the goal-directed system and the more basic process described in the last paragraph may be capable of causing exploratory behaviour. However, often exploratory actions are not driven by desires. An adult or child who picks up an unfamiliar object, inspects it, manipulates it, or puts it in their mouth may have some desire that they believe will be served by this action, but they may equally be motivated by a quite different, non-goal-directed process. If their action was caused by a simple process by which attention drives the manipulation of unfamiliar objects and other exploratory actions then it would be wrong to think of it as driven by desire.

The second possible source of motivation to be discussed is the emotions. Philosophers have argued that some actions are correctly explained as expressions or other manifestations of emotion, rather than as the product of desires or habits (Döring 2011). A famous example is the action of a woman who scratches out the eyes of a photograph of a rival (Hursthouse 1991), which is notable for being particularly difficult to explain in other ways, since it does not seem to promote anything the woman would be likely to value. But actions need not have the same symbolic quality as this one to be plausibly driven by emotion. For instance, fear seems to drive us to take quite practical actions, such as retreating from the situations that scare us; we are motivated to express love in ways which are communicative as well as expressive (e.g. by kissing people we love); and we sometimes communicate anger in person, verbally or even by physically assaulting those who offend us. There are many possible processes by which these actions could be generated, but it is plausible that they could be controlled by an emotional system independent of the habitual and goal-directed systems. Some expressions of emotion, such as facial expressions, exist in the same form across all human cultures (Ekman 1992), which suggests the existence of a specific mechanism by which emotions cause action. If this system could interact with some relevant general-purpose mechanisms, it could be responsible for motivating many actions that might be thought of as expressions of emotion.

For instance, the system may include an unlearnt drive to perform escape behaviours when we experience fear, and this could explain the way a rock climber behaves to get down from a route when they lose their nerve, provided it could interact appropriately with the climber's relevant beliefs and skills. Similarly, a skilled fighter who attacked someone who offended her may be motivated to attack by their emotions, and led to attack in the specific way she did by her training in combat. In Hursthouse's example, the woman may be motivated to hurt her rival, and scratch out the eyes of the photograph because this is an object with a strong learnt association with the rival. A person who kisses someone because they feel love for them at that time does so because their culture has taught them that kissing is an appropriate way to express love, perhaps in combination with an unlearnt drive to express this emotion. So the present hypothesis is that emotions may be capable of motivating us independently of desires, but in combination with our beliefs. And if an account like this is necessary in order to explain some actions which are hard to construe as caused by desires, since they do not seem likely to bring about any outcomes the agent would value – such as in Hursthouse's example – then it may be the best way to explain other actions too. There is no doubt that rock climbers sometimes back off from routes because they judge quite coldly that they are unacceptably dangerous, and they want to avoid being hurt, but this does not mean that on other occasions when they retreat they may not be motivated wholly or partially by fear.

Empirical studies of anger and aggression lend some support to this view. A picture is emerging according to which impulsive aggression is caused by activity in the amygdala, which can be controlled in healthy adults by serotonin signals from the prefrontal cortex, including the OFC (Davidson et al. 2000, Nelson & Trainor

2007). This view is supported by a study of patients with Intermittent Explosive Disorder (IED), which is specifically characterised by frequent, disproportionate displays of violence and aggression (Coccaro et al. 2007). Using fMRI, this study found hyperactivity in the amygdala in IED patients in response to emotionally salient faces, and diminished activity in the OFC. Second, it has recently been argued that amygdala activity causes rejections in the Ultimatum Game (Gospic et al. 2011). This is a game in which one player proposes dividing a pot of money with another; the proposer can suggest any way of dividing the pot that they like, but neither player receives any money unless the proposal is accepted. So in one-shot games, the second player will benefit by accepting any proposal except a 100%-0% split, but in studies of this game rejections of 'unfair' proposals are common and may be interpreted as expressions of anger. So there seems to be a quite different mechanism causing angry responses, as compared to normal goal-directed behaviour, since in angry behaviour response strength is related positively to amygdala activity and negatively to OFC activity, rather than positively to activity in the OFC and other prefrontal areas. Like exploratory actions, then, expressions of emotion may on some occasions be caused in ways that do not involve desires.

The third topic to be discussed is how conscious reasoning and explicit evaluative beliefs can influence action. In addition to the habitual and goal-directed systems, humans seem to possess a *planning system* (Papineau & Butlin forthcoming). The habitual and goal-directed systems are both for deciding what to do *now*, but we also have the ability to think about what to do in the future, and to form plans, intentions and resolutions.[9] In the process of doing so, and at other times, we reflect on our own desires and other psychological states, sometimes endorsing them and sometimes judging that they are harmful to us or bad in other ways. We also form evaluative beliefs which are to some degree independent of our desires, and which may affect what we are motivated to do either directly, or by influencing our

[9] A distinction is sometimes drawn between *intentions for the future* and *the intentions with which we act*. Intentions for the future, discussed by Bratman (1987) and Holton (2009) are akin to plans. These intentions are psychological states which can compete with desires and other sources of motivation in cases of temptation, and we often act without them. The intentions with which we act, discussed by Anscombe (1957) are thought to be present whenever we act out of choice. Our intentions in this sense are, roughly, what we take ourselves to be doing or trying to do. It is a matter of controversy how these intentions relate to the psychological states that cause actions. My concern here is only with intentions for the future.

intentions for the future. It is important in the present context to distinguish *beliefs* from *states with (only) the mind-to-world direction of fit*. My view is that desires have only the mind-to-world direction of fit, but now the question at hand is how states other than desires are involved in motivation.

Although intentions and evaluative beliefs are usually thought of as states of different kinds, they are sufficiently closely related that I will address them together. One question is to what extent desires are involved in forming intentions and evaluative beliefs, and here the issues are intertwined in part because evaluative beliefs (or judgments) seem to play a major role in the formation of intentions. If this is correct, then if desires are involved in the formation of evaluative beliefs, they must also be involved in forming intentions. On the other hand, it is implausible that evaluative beliefs but not intentions could be formed independently of desires, because at least some evaluative beliefs seem sufficient alone to give rise to intentions (such as the belief that a particular course of action is *by far the best* available to one). So we can focus on evaluative beliefs in addressing this question. A second question is to what extent intentions and evaluative beliefs must interact with the goal-directed system in order to cause action, and it would be strange if evaluative beliefs but not intentions could cause action independently. This would mean that there was a separate system for motivation and action which could be controlled by conscious reasoning, but in which intentions played little or no part. Given that evaluative beliefs can give rise to intentions, this means that the question amounts to whether intentions can cause action independently, or only through interaction with the goal-directed system. So on this issue, I will focus on intentions.

Both of these two questions are challenging and ultimately empirical. On the question of the role of desires in the formation of evaluative beliefs, I shall simply note some plausible relevant hypotheses (which are not in competition with one another). First, Velleman (1992) and Hawkins (2008) have suggested that desires are centrally involved in the acquisition of evaluative concepts. Even if this is correct, however, it leaves open the possibility that we can form new evaluative beliefs in ways that are wholly independent of desire. And indeed this does seem to be the case; for example, we can form new evaluative beliefs simply by accepting testimony. If somebody was told that *sati* is a very bad thing, without being told what it is, by someone they trusted, they would be likely to come to believe this (*sati* is the now-rare practice of Hindu wives self-immolating on their husbands' funeral pyres). More significantly, we often seem to form new evaluative beliefs from existing ones in ways that do not engage our desires. If I believe that being slim is good and eating swiss chard is an effective way to become slim, then I may well come to believe that eating swiss chard is a good thing to do. This process *might* engage my desires – for instance, I might see a picture of someone with an enviable physique munching chard, which would help to direct my attention to the link – but it is hard to see why this would be necessary, since presumably we draw comparable inferences concerning non-evaluative matters frequently without our desires being involved.

However, even if *this* is correct, there are still good reasons to think that our settled evaluative beliefs are primarily the products of conscious reflection on our desires and how they relate to one another. One is that we cannot rely on our perceptual faculties to tell us what is good in the same way as we can when forming many other beliefs. Besides desire, this leaves testimony and *a priori* reasoning as possible sources of belief about what is good; and although we may sometimes accept others' testimony about evaluative matters, we will more typically expect them to justify their evaluative claims by explaining how they relate to pre-existing shared evaluative beliefs. So unless other inputs are available, it seems that desires must play an important foundational role in our evaluative thought. Another possible input is moral intuitions; but even if we have a stock of moral intuitions which are distinct from desires with moral content, our desires would still be needed to play this role when we think about what is good *for us*.

Turning now to how evaluative beliefs and intentions lead to motivation and action, we can assume that humans possess a planning system, which enables us to consciously reflect on our desires and form plans, intentions and resolutions. Some neuroscientific evidence exists suggesting that this system relies on the dorsolateral prefrontal cortex, which has been implicated in both self-control (Hare et al. 2009) and planning (Kaller et al. 2011). It is clear that this system sometimes affects how we behave, but less clear how it does so; one possibility is that this system constitutes a further route to action, in addition to the habitual and goal-directed systems, and the systems for exploratory and emotional behaviour just discussed, but another is that plans and intentions affect our actions by influencing the operation of the goal-directed system.

To see this contrast, it is helpful to consider cases of temptation, in which our intentions conflict with strong occurrent desires at or just before the time of action, and specifically those in which we exhibit *strength of will*, acting on our intentions despite these strong occurrent desires (Holton 2009). If the planning system affects motivation and action independently of the goal-directed system, cases in which we resist temptation will be ones in which our actions are contrary to our strongest desires. But an alternative hypothesis is that in cases like this while our occurrent desires may initially motivate us more strongly to succumb to temptation, the effect of our intentions is to change the strengths of our occurrent desires, with the result that we end up acting in accordance with our strongest desires.

Richard Holton (2009) has argued in favour of the former view, drawing on studies by Baumeister and colleagues (e.g. Baumeister et al. 1994, 1998). These studies suggest that we possess a domain-general, limited capacity for self-control, by showing that a certain class of tasks that intuitively involve concentration and resolve are effortful and that our performance on them diminishes over time, and when we are depressed, anxious or tired. Among the tasks that require self-control is resisting temptation – that is, sticking to a resolution in the face of desires to do otherwise. So Holton uses Baumeister's work to support his claim that resisting temptation involves a struggle, and argues that this struggle must be a contest between the agent's desires, on one side, and their resolutions, on the other. This argument is not wholly persuasive, however, because it is plausible that processes taking place within the goal-directed system could be extended over time, effortful and require self-control. A key reason to believe this is that we often make efforts to control what we attend to, and attention is an important factor in determining the strength of occurrent desires. So it may be that when we are tempted our intentions cause us to try to shift our attention away from whatever is tempting us, and towards the potential benefits of sticking to our plans. It is easy to imagine that such a process could sometimes have the character of a struggle.

I therefore leave it open to what extent evaluative beliefs and intentions represent a further route to action, separate from the goal-directed system. But these points do show that on my account of desires as outcome values, desires may relate to evaluative beliefs and intentions in very much the ways that philosophers have suggested. Desires sometimes come into conflict with intentions, and these internal conflicts may be won by either side; desires may contribute to the formation of intentions, but evaluative beliefs, which are distinct from desires, may also do so; and desires contribute to the formation of evaluative beliefs, and especially beliefs about what is good for us, but it can easily happen that our desires and evaluative beliefs do not coincide.

Overall, then, while desires are of crucial importance for determining how we act, this importance may be overstated. Exploratory drives, emotional responses and the human abilities to employ evaluative concepts and plan for the future all affect how we act, resulting in a complex overall picture. The picture is made more complex still by the apparent interactions between various systems, such as the tendencies of desires to affect our emotions and evaluative judgments.

4.3 Conclusions of Chapters 2-4

In order to show that desires have only the mind-to-world direction of fit, I am arguing for the following five premises:

I. Desires are outcome values.
II. The goal-directed control system works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.
III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.
IV. Biological representations with consumers that have discretion have only the mind-to-world direction of fit.
V. It follows from I-III that desires are biological representations with consumers that have discretion.

Of these, premises I-III are the targets of part I, while premises IV and V will be defended in part II. Since the positive work towards establishing premises I-III is now complete (chapter 5 will focus on a possible objection), in this section I will explain how the account developed in chapters 2-4 supports each of them. Since the case for premise I relies substantially on premises II and III, I shall discuss the three premises in reverse order.

Premise III is made up of three claims: that outcome values are inputs to the goal-directed system; that they are produced and modified by a system that is to some extent responsive to evidence; and that is normal for more than one outcome value to act as an input to the system at any one time. The first of these follows immediately from the definition of a goal-directed system, as a system for behavioural control that relies on two kinds of states, that track the values of outcomes and contingencies between actions and outcomes, respectively. The second also follows from this definition, but I further discussed the mechanism which forms and updates outcome values in sections 3.4 and 3.5. Although we don't yet have a clear picture of how this mechanism works, we can be confident that there is such a mechanism, which uses basic drives, existing desires, and a relatively unsophisticated associative learning procedure to update outcome values. So for these first two parts of premise III, the key question is whether the goal-directed system is real.

The third claim in premise III also coheres well with the account of the goal-directed system I developed in chapter 3. In section 3.2 I proposed that at any one time several outcome values are occurrent, and that the goal-directed system's task is to use these outcome values to calculate expected reward values for possible actions. Apparent advantages of this picture include allowing the system to readily take into account the fact that any given action is likely to affect the likelihoods of many rewarding outcomes, and making sense of the plausible idea that occurrent desires sometimes change in strength in the course of a single process of action-selection. However, I will also now give a more direct argument for this claim.

Suppose that only one outcome value acted as an input to the goal-directed system at any one time. There are two ways that the goal-directed system could work, if this were the case. First, it could be that some other system would work out which outcome value the goal-directed system should try to satisfy, and then the goal-directed system would cause whichever action seemed likeliest to satisfy it. This way of dividing up the task of selecting an appropriate action is not only of dubious merit, but very different from the way in which goal-directed systems work, as a matter of definition. If things worked this way then the goal-directed system itself would not take the value of the outcome into account. Second, it could be that the goal-directed system evaluates available actions and outcomes one at a time, taking only one outcome value as input when considering each action. It would continue doing this until it comes across an action that passes some threshold, then perform that action. We have seen little evidence so far that bears directly on this possibility, but it is very doubtful that animals with goal-directed systems never either directly compare two possible outcomes when considering how to act, or take into account more than one outcome to which an action is likely to lead when considering whether to perform it. For instance, in the Iowa Gambling Task (Bechara et al. 1997) humans learn to act in a way that takes into account the relative likelihoods of two different outcomes (winning money and losing money) when selecting cards from different piles, and a similar task has been developed for rodents (van den Bos, Koot and de Visser 2014). In this task rats learn to pull one lever that provides greater expected reward than another, when both levers have some chance of providing either a positive outcome (sugar pellets) or a negative one (quinine-soaked sugar pellets), even though the lever with the lower expected reward provides greater positive outcomes. So even rodents seem to use multiple outcome values as inputs to at least some goal-directed control processes.

The details of premise II also follow from the definition of a goal-directed system, so again here the empirical evidence is important primarily for the support it gives to the claim that humans possess such a system. In the last three chapters, I have presented both direct and indirect evidence for this claim. The direct evidence includes the results of outcome devaluation and contingency degradation experiments, discussed in section 2.2, and the lesion studies on rats which show that the different patterns of behaviour revealed by these experiments are facilitated by distinct neural systems, working in parallel (sections 2.5 and 3.1). The neuroscientific evidence showing that the values of outcomes are represented in the OFC and employed in action selection, presented in section 3.1, also supports this conclusion. More generally, the work of all three chapters helps to show that the goal-directed system is real, by showing how such a system could work, and how it could cause, in combination with others, real patterns of human behaviour. Finally, an important point which I have not been able to demonstrate in detail here is that there is a substantial body of recent literature in cognitive science which relies on the existence of the habitual and goal-directed systems. We should believe in these systems because they are among the central theoretical posits of a successful scientific research programme.

This leaves premise I: the claim that desires are outcome values. I have assumed (chapter 1) that if there is some natural kind of psychological state that does enough of what desires are normally thought to do, then what it is to be a desire is to be a member of that natural kind. So the question at hand is whether outcome values form such a natural kind. The best evidence we can have that some class of psychological states form a natural kind is that they are treated as such by modern science, and this is true of the states I have been calling ‘desires’. When the brain is functioning normally, they play a clearly-defined role in the operation of the goal-directed system, which is widely recognised as a distinct psychological unit. To show that these states do not form a natural kind it would be necessary to show either that some other psychological states also belong to the same, wider kind, or that the class is best thought of as made up of more than one distinct kind.[10] The former is implausible, given that only desires are both inputs to the goal-directed system and used in assessing the reward value of present states of affairs for the purpose of generating reward signals (section 3.4), but I will discuss this issue further in chapter 5. On the latter, by far the most significant distinction between different kinds of desire that we have seen in the last three chapters is that between occurrent and standing desires (section 3.2), and it is fairly plausible that these form distinct natural kinds; they perform different roles, and belong to different ontological categories. But this point in no way undermines the case for identifying the states concerned as desires, since this distinction is also widely recognised by philosophers. Arguably, the fact that the inputs to the goal-directed system come in distinct standing and occurrent variants is an instance of striking similarity between the class of states that I have described, and desires as they are thought of by many philosophers.

Finally, then, premise I turns on whether outcome values have enough of the properties usually associated with desire. The most important of these properties is combining with instrumental beliefs to cause a high proportion of human actions. Further to this, desires are usually thought to be both among the inputs to the formation of intentions for the future, and to be capable of interfering with the

[10] But not straightforwardly sufficient, in either case, since there are thought to be hierarchies of natural kinds.

execution of these intentions by tempting us to abandon or otherwise act contrary to them, and these points also seem to be true of the inputs to the goal-directed system. Because the strength of our occurrent desires is highly dependent on the details of the situation, it is not surprising that they can lead us to plan for one course of action in advance, then be strongly motivated to perform another when the time comes. Another very important point is that we usually take desires to have a wide range of objects, and I have explained (in section 3.4) why outcome values are likely to have a similar range of objects. Also, as I have said, there are occurrent and standing variants of outcome values, just as there are usually thought to be with desires, and the occurrent versions are sensitive to our physiological needs, our environment and to what we attend to, just as we would expect to be true of desires (sections 3.2 and 3.3). It also reflects common sense and widespread philosophical views that in addition to the actions caused by the goal-directed system and hence by outcome values, there are also habitual actions and expressions of emotion which are not caused by this system. And although it is not certain exactly how the inputs to the goal-directed system are formed and updated, they do seem to be responsive to evidence about reward to about the same extent, and in the same ways, as we would normally think of desires as being.

This does leave out some properties typically associated with desire, which I have not discussed. Perhaps the most notable is the putative relationship between desire and pleasure: that the satisfaction of desires tends to cause pleasure. Although I have not discussed pleasure, my account of the behaviour of outcome values does cohere well with this idea, because I have argued that they are used in assessing the levels of reward provided by our current circumstances. It is also important to note that, given my account of the sources of motivation and action, it is hard to see any natural kind of psychological state with a better claim to being desire than outcome values. Taking all these points together, and pending an objection to be discussed in the next chapter, I conclude that desires are outcome values.

Chapter 5: Schroeder’s Theory of Desire

5.1 Outline of Schroeder’s Theory

In this, the final chapter of part I, I address a possible objection to my view. Partly in collaboration with Nomy Arpaly, Tim Schroeder has developed a theory of desire that draws on a similar body of empirical evidence to mine (Schroeder 2004, Arpaly & Schroeder 2014). Schroeder’s work shares my aim of giving an account of desire as a natural kind, and he also shares my view that this is best achieved by using the results of psychology and neuroscience. These common features are no coincidence; Schroeder’s work has been a major influence on how I think about desire, and how I approach philosophy of mind more generally. But Schroeder’s theory differs from mine in some substantial ways, which raises the question: why should someone accept my view about what current neuroscience tells us about desire, rather than his? So in this chapter, I explain why my account is to be preferred. In this section, I outline Schroeder’s theory, and explain the principal ways in which it differs from mine.

Schroeder states his theory as follows:

> To have an intrinsic (positive) desire that p is to use the capacity to perceptually or cognitively represent that p to constitute p as a reward. (Schroeder 2004, p. 131)

> To have an intrinsic appetitive desire that p is to constitute p as a reward. (Arpaly & Schroeder 2014, p. 128)

In his 2004 work, he also gives an explicit statement of what he takes ‘constituting p as a reward’ to mean:

> For an event to be a reward for an organism is for representations of that event to tend to contribute to the production of a reinforcement signal in the organism, in

> the sense made clear by computational theories of what is called ‘reinforcement learning’. (Schroeder 2004, p. 66)

Schroeder’s view, then, is that to desire an outcome is to have a certain kind of psychological disposition towards that outcome, which amounts to treating it as a reward. He characterises what it is to treat an outcome as a reward by saying that this means being disposed to generate a more positive reinforcement signal when one represents that outcome than when one does not, other things being equal. And his view is that what it is for something to be a desire is for it to be the categorical basis of such a disposition. This means that the states that I take to be standing desires would also count as desires on Schroeder’s view, since they are used for measuring present levels of reward for the purpose of generating reward signals. Schroeder takes phasic dopamine signals – which he says represent reward prediction errors – to be the only reinforcement signals in humans. So his theory implies that human desires are the categorical bases of the disposition to produce (positive) phasic dopamine signals as a consequence of representing outcomes.

This theory differs from mine both in how it characterises desires, and in its extension. One difference in how desires are characterised is that in my view, what it takes for a psychological state to be a desire is for it to be a member of a particular natural kind. In contrast, Schroeder gives a single characteristic functional property which he claims is possessed by all and only desires. However, this difference will not be central to my critique of Schroeder’s theory.

Another difference in how desires are characterised is that Schroeder’s theory focuses on the relationship between desires and reinforcement, whereas mine focuses on the relationship between desires and goal-directed behaviour. Part of the explanation for this difference may be that Schroeder takes desires to contribute to motivation and action by causing the production of dopamine signals, and hence by a similar mechanism to that by which they contribute to reinforcement learning. So for Schroeder, the roles of desire in reinforcement and in motivation and action are relatively closely related. In section 5.2, I argue against Schroeder’s account of how desires contribute to motivation and action, and show that his view has implausible consequences if this flaw is corrected. Schroeder also places much less weight than I do on the distinction between occurrent and standing desires, because his proposed mechanism allows standing desires to contribute to motivation and action directly.

Regarding the extensions of our two theories, a striking difference is that on Schroeder's view, basic drives count as desires. This is because basic drives contribute to reinforcement learning in broadly the same way as desires: they dispose us to generate reward signals when we represent outcomes of specific kinds. I will argue in section 5.3 that this consequence counts against Schroeder's theory, because desires and basic drives contribute to goal-directed control in different ways. Then in section 5.4 I will argue that it also counts against his theory in another way, which is that standing basic drives are fundamentally the wrong kind of psychological feature to be desires. A further difference in extension is that Schroeder's theory entails that any creature capable of reinforcement learning must have at least some desires, even though some such creatures lack the capacity for goal-directed behaviour.[11]

In the following three sections I will argue that these differences amount to relative advantages of my theory, and disadvantages of Schroeder's. I will focus on the relationship between reinforcement learning and goal-directed control in section 5.2, and on basic drives in sections 5.3 and 5.4.

5.2 Reinforcement Learning and Goal-Directed Control

One significant disadvantage of Schroeder's theory is that he characterises desire with reference to its role in reinforcement learning, rather than its role in goal-directed control. This generates two problems. First, Schroeder's theory entails that creatures which are just too different from humans can have desires. Second, it seems to give the wrong account of what happens when creatures which are capable of both reinforcement learning and goal-directed control lose the former ability, but not the latter. This is a real possibility, because the mechanisms underlying these two abilities are less closely connected than Schroeder thinks; so in this section I will argue against Schroeder's account of how desires contribute to goal-directed motivation and action, as well as his view about what desires are.

The first problem with Schroeder's theory is simply that because reinforcement learning is necessary for habitual control, Schroeder's theory entails that creatures

[11] A third difference is that Schroeder's theory allows that robots and aliens may have desires, whereas mine denies this – assuming that we cannot share psychological natural kinds with robots and aliens.

which possess habit systems but not goal-directed systems nevertheless have desires. In such creatures, these desires will typically be basic drives. In this case, it would be wrong to say that the creatures *act on* their desires, because they will not be sensitive to the outcomes their actions bring about. For instance, suppose some such creature acquired the habit of performing action A whenever it was hungry (treating this internal state in just the same way as a stimulus in the environment, as discussed in section 3.3). If a scientist arranged that whenever this creature was hungry and performed action A, an outcome would follow which satisfied another of its basic drives, but did not alleviate its hunger, then the creature would continue to perform the action when hungry. So it would be wrong to describe this animal as acting on a desire for food, because its tendency to act in this way would be wholly unresponsive to whether it received food as a result. Furthermore, there is nothing in Schroeder's theory, or in the nature of reinforcement learning or habitual control, which implies that 'purely habitual' creatures must be capable of sensing their own physiological states or otherwise forming occurrent basic drives. So on Schroeder's view, there could be creatures which had desires, but never initiated actions or changed their behaviour in response to anything other than environmental stimuli.

It is also possible that there could be purely habitual creatures with *acquired* psychological states that would count as desires on Schroeder's view. These would be standing states which are updated by reward learning and used in measuring the values of outcomes, but which do not contribute to action-selection more directly (see section 2.5). If anything, it is clearer still in this case that the states concerned are not desires. This is because actions learnt because they led to outcomes represented as valuable by such states would be insensitive to outcome devaluation. So if we were to say of these creatures that they sometimes acted on their desires, we would have to also accept that they could sometimes lose those very desires, and yet continue to behave in the same way. Schroeder's theory seems to imply that there are creatures with desires on which they are constitutionally incapable of acting.[12]

[12] Schroeder accepts that his theory has this consequence, and appeals to Strawson's (1994) *Weather Watchers* in suggesting that it is intuitive that there could be creatures with desires but no capacity for action. The Weather Watchers have beliefs about the weather, and feel pleasure or displeasure depending on changes in the weather, but cannot act. However, it is likely that the intuition that Weather Watchers have desires comes from imagining them as having conscious mental lives much like ours (or perhaps like parts of ours), and it is doubtful whether this is possible in creatures that cannot act.

In order to bring out the second problem with his theory, I need to explain why it is implausible that, as Schroeder claims, reinforcement signals in the form of phasic dopamine are a necessary part of the process that leads to goal-directed action. Schroeder's account of goal-directed action is as follows: when we think of highly desired outcomes, this causes dopamine to be released, which in turn makes us more likely to immediately perform the actions we expect to lead to those outcomes. I will start by explaining why the evidence Schroeder offers in favour of this account is not convincing. He cites three forms of evidence which are directly relevant to this claim: studies suggesting that dopaminergic projections to the motor prefrontal cortex are necessary for maintaining motor intentions over time; the point that dopamine boosts action-selection by its effects on D1 and D2 receptors on medium spiny neurons in the striatum (see section 2.4); and the fact that loss of dopaminergic activity causes impaired motion in Parkinson's disease (Schroeder 2004, pp. 116-118). The problem with all of this evidence is that while it does show that dopamine is necessary for motivation and action, this is far from sufficient to show that phasic dopamine signals are the means by which desires affect how we behave. More specifically, a widespread view is that while phasic dopamine signals represent RPEs and are used for reinforcement, motivation and the ability to control movements (which is lost in Parkinson's) are products of *tonic* dopamine levels – that is, the levels of dopamine release that obtain between phasic bursts (Niv et al. 2007, Schultz 2007). So it is possible to explain the evidence that Schroeder describes just on the grounds that dopamine is necessary for the normal functioning of action-selection circuits, without making the much more specific claim that dopamine signals are the means by which desires influence goal-directed control. Also, as I have mentioned, there is an ongoing debate about the function of phasic dopamine signals, and neither of the two most prominent positions in this debate sits happily with Schroeder's account. On one hand, the neuroscientist Kent Berridge has defended the view that phasic dopamine signals have the function of generating motivation to bring about desired outcomes, but he sees this as an incompatible alternative to the view that phasic dopamine signals are for reinforcement (Berridge 2007). On the other hand, the orthodox view is that phasic dopamine signals are reinforcement signals, but do not play the role in motivation that Berridge proposes (Wise 2004, Balleine et al. 2008, Glimcher 2011).

What’s more, it is hard to see how the mechanism Schroeder describes could work. The problem is that dopamine release is not targeted at particular groups of cells, and carries little information about how it is caused, so there is nothing except timing to distinguish one dopamine signal of a given strength from another. This means that the only way in which desires could be connected, via dopamine signals, to the correct actions is if the goal-directed system worked by considering actions in turn. In order to choose between a number of available actions, the goal-directed system would have to represent them one at a time, use dopamine signals generated by desires at that time to associate these actions with reward values, and (presumably) store these reward values for later comparison. But this would be entirely at odds with the basal ganglia’s mechanism for action selection, and with the point that not only action selection, but many other cognitive processes, seem to be facilitated by competition between simultaneous coalitions of cortical activity (see sections 2.3, 3.2 and 4.1). We have excellent reasons to think that action-selection involves the representation of possible actions simultaneously, rather than in turn.

This argument shows that the process by which desires contribute to reinforcement learning is likely to be substantially distinct from that by which they contribute (most directly) to goal-directed control. It turn, this makes it very likely to be possible for humans to sustain brain damage which would prevent some or all of their desires from continuing to contribute to learning, while leaving their capacity for goal-directed control intact. This damage would mean that the objects of the relevant desires were no longer ‘constituted as rewards’ by the agent, as Schroeder defines that property, and were therefore no longer desired. However, it is not plausible that this damage would affect which natural kinds the agent’s desires belonged to, so they would continue to count as desires on my account. This is an advantage of my view, because someone who had suffered in this way would continue to perform actions that, prior to sustaining the brain damage, would undeniably have been motivated by desires, and these actions would still be caused in the same way. We should not deny that they still had the desires they did before, because they could still act on those desires.

5.3 Desires and Basic Drives (I)

As we have seen, Schroeder's theory classifies basic drives as desires. One reason why this is problematic, as I have suggested, is that it is possible for creatures to have basic drives without having goal-directed control systems. In this section I will argue that even in rats and humans, desires and basic drives contribute to goal-directed control in different ways. And then in section 5.4, I will describe a further important difference between desires and basic drives.

In section 3.3, we saw that the process of incentive learning is necessary for occurrent basic drives to influence goal-directed behaviour (Balleine 1992, Niv et al. 2006). For instance, rats that have been trained when hungry to press a lever in order to receive a specific foodstuff will continue to press the lever at the same rate when sated, unless they have previously experienced that food when sated. Conversely, rats that have been trained to perform an action like this when only slightly hungry will not increase responding when they are much hungrier, unless they have experienced the foodstuff concerned in the hungrier state. What these results seem to show is that levels of motivation to perform goal-directed actions are determined directly by the strengths of occurrent desires for the specific outcomes concerned, not by the strengths of occurrent basic drives. The role of occurrent basic drives is to modulate the strengths of occurrent desires, in a way determined by the individual's experience of the relationships between the strengths of occurrent drives, and the values of outcomes. As in the habit system, occurrent basic drives play a similar role to environmental stimuli (which may also be associated with, and boost the strength of, particular desires). So occurrent basic drives and occurrent desires play different roles in goal-directed control: desires make a direct contribution to action-selection, and drives make only an indirect one.

A particularly clear demonstration of this point comes from the contrast between outcome devaluation studies that devalue the outcome using specific satiety (e.g. Balleine & Dickinson 1998) and experiments on incentive learning that involve a transition from hunger to satiety (e.g. Balleine 1992). In studies of both kinds, rats are given the opportunity to press a lever to receive a novel foodstuff, and learn to perform this action, then taken away from the lever and fed until they are sated, before being given the opportunity to press the lever again ‘in extinction’[13]. The only difference is that in the outcome devaluation experiment, the same food is used that the rat has learnt to get by pressing the lever. Surprisingly, rats that undergo these procedures will press the lever less when they have been fed to satiety on the same food, and will not reduce responding when they have been fed to satiety on a different food – unless they have undergone the necessary incentive learning. If the drive for food was directly affecting the rats’ behaviour, they would have reduced responding in both paradigms. These points show that although basic drives and desires contribute to reinforcement learning in the same way, they contribute to goal-directed control in different ways; so it is an advantage of my theory over Schroeder’s that I distinguish between these two importantly different kinds of psychological states.

5.4 Desires and Basic Drives (II)

The second difference between basic drives and desires is more fundamental. Standing basic drives are what I will call *mental rules*, whereas desires – both standing and occurrent – are *mental states*. Furthermore, it is a plausible desideratum on theories of desire that only mental states should count as desires. The distinction between mental rules and mental states needs to be drawn carefully, so I shall start by setting it out without reference to the present debate, before showing that standing basic drives are mental rules.

Intuitively, if we assume that mental states such as beliefs, desires and perceptual states are representations that control our behaviour by interacting causally in ways that respect their semantic properties (Fodor 1975); or even if we make the weaker assumption that cognition relies on internal representations *in some way*; then there must be mental rules that govern how these representations interact. An argument closely analogous to that made by Lewis Carroll in ‘What the tortoise said to Achilles’ (1895) supports this point. Imagine we want to explain why someone has inferred a proposition Z from two others, A and B, which they believe. If we only appeal to mental states in this explanation, it seems we must attribute to them the further belief that Z follows from A and B. But then Carroll’s tortoise would ask us

[13] i.e. without any reward being delivered.

why the agent would infer Z from A, B, and the proposition that Z follows from A and B, and our only recourse – again assuming that we could only cite mental states – would be to cite a further belief, which is that Z follows from A, B, and the proposition that Z follows from A and B. There is something seriously wrong with this approach, and the mistake seems to be the practice of only appealing to mental states, which here manifests as an assumption that whenever someone acts as though they accept some proposition, this can only be explained by attributing to them the belief that that proposition holds. A much better approach would be to look at some point for a mental rule that the agent follows, perhaps of the form *draw salient modus ponens inferences*. In general, then, at the bottom of explanations of behaviour that appeal to mental representations, there must be explicit or implicit appeals to rules. It would be too quick to conclude that these rules cannot themselves be representations, but they must in some sense be *built in* to the system concerned – otherwise we would need to cite further rules to explain how they were acquired.

It is not obvious how to make the intuitive distinction between representations and rules more concrete, and it is likely that different ways of drawing the distinction will be useful for different explanatory purposes. Here I will revise one version of a different, but closely related distinction, that between *implicit* and *explicit* representations.[14] This distinction is closely related, because mental states are explicit representations, whereas if they are representations at all, mental rules are implicit representations. For example, we might think of a system with a 'built-in' tendency to draw *modus ponens* inferences as implicitly representing the conditional claim that *if p and (p→q), then q*. The version of the implicit/explicit distinction I will employ is articulated by Shea (2015), who writes that implicit representations, 'can have no impact on subsequent processing except via the representations which they connect.' In contrast, explicit representations can act as inputs to many further processes (pp. 79-80). Shea's idea is that the implicit/explicit distinction can be understood in terms of the degree of promiscuity exhibited by representations. Those that can enter into only one kind of process, in one way, are implicit, while those that can enter into more than one kind of process are explicit.

[14] I use the terminology of states and rules, rather than of implicit and explicit representations, primarily because the implicit/explicit distinction has previously been drawn in a number of different ways for different purposes. See e.g. Dennett 1983, Davies 2015.

My distinction between mental states and mental rules follows Shea's in focusing on the range of ways in which features of the mind enter into psychological processes (their 'promiscuity'), but removes the restriction to *subsequent* processing, and the idea that what matters is the way that features can act as *inputs*. Mental states can also be promiscuous in virtue of being the outputs of psychological processes – that is, being formed or modified by such processes. So as I think of them, mental rules are necessarily innate; that is, acquired in development rather than through learning. Mental states may be innate, but for this to be the case they must count as explicit representations in Shea's sense – they must be capable of acting as inputs to psychological processes in more than one way. Thus:

> *Mental rules* are innate features of the mind, each of which causes exactly one kind of transition between mental states.

> *Mental states* are features of the mind which enter into psychological processes in two or more ways, one of which may be by being formed or modified by a psychological process.

Before turning to the status of desires and basic drives, I will give two brief arguments for the interest and relevance of this distinction.

First, one possible shortcoming of Shea's distinction is that it seems that whenever a system is disposed to make a transition from one state to another, this transition could in principle be caused either by a 'built-in' rule, or by a further state which the system accesses and applies to the present situation. Shea's criterion only distinguishes these two cases if the system is also capable of using the state that it accesses for purposes other than making transitions of this kind, but in principle it seems that access for other purposes could be prevented by the system's architecture. By distinguishing between innate and learnt features, my distinction can account for this difference.

Second, my distinction implies that mental rules will typically be close to universal among agents of the same species, while mental states will more often be idiosyncratic. Much of the time, the appropriate kind of explanation to give of the presence of a mental state will be psychological, whereas the best explanation of a mental rule will be broadly biological. In psychological explanations based on causal generalisations, mental rules will typically provide the generalisations, while appeals to mental states will explain why, granted those generalisations, matters turned out the particular way they did on the occasion in question. These points are important because part of what makes desires interesting is the particular role that they play in explanations of behaviour of various kinds, both causal and rational, a subject which I take up in part III. So for my present purposes it is appropriate to draw a distinction between mental states and mental rules that divides two classes of psychological features which are apt to play different kinds of explanatory roles.

It now remains only to show that standing basic drives are mental rules. By my criterion, for this to be the case they must cause exactly one kind of transition between mental states. Standing basic drives certainly dispose us to produce reward signals when we perceive certain kinds of states of affairs, such as our eating nourishing food, having sex, and being in the presence of smiling friends. It may be the case that the strength of these reward signals depends on the extent to which the object of the drive concerned is needed at the time; for instance, eating is plausibly more rewarding when we are hungrier (although desires will also contribute to this effect), and getting warmer is certainly more rewarding when we are cold. But this does not suggest that more than one transition is mediated by standing basic drives. It merely shows that they employ more than one input to generate outputs. The key question is whether standing basic drives do anything other than contribute to generating reward signals, and on this point there is a noteworthy disanalogy between basic drives and desires. Standing desires contribute to psychological processes both by generating reward signals, and by generating occurrent desires, but standing basic drives are not needed to generate occurrent basic drives. Instead, we simply perceive our own physiological states, and the representations of these states which are thus formed act as occurrent basic drives – they contribute to habitual and goal-directed control in the same way as perceptual representations of environmental stimuli, as I have explained (section 3.3). A further disanalogy between standing basic drives and standing desires is, of course, that standing desires are learnt – they are formed and modified by psychological processes – whereas standing basic drives are innate. I therefore conclude that Schroeder's theory is unsatisfactory because it counts some features of the mind as desires which are not mental states, but are instead mental rules.

In this chapter, I have described Tim Schroeder’s empirically-informed theory of desire, and argued that my own theory is preferable. Schroeder’s theory is weaker because it links desire to reinforcement learning rather than to goal-directed control, and because it implies that basic drives are desires. This completes part I of this thesis, in which I have focused on the nature of desire. Relying on the assumption that desires form a natural kind, I have drawn on psychology and neuroscience to develop a detailed account of the role that desires play in our mental lives. Next I will turn to the other main topic of this thesis, which is representation, and in particular direction of fit.

Part II: Direction of Fit

Chapter 6: The Case for a New Theory of Direction of Fit

6.1 Introduction to Part II

In this and the following three chapters, I will argue for a new theory of direction of fit, and show that in combination with the account of desire developed in part I, it implies that desires have only the mind-to-world direction of fit. So here in part II I argue for premises IV and V of my overall argument:

> IV. Biological representations with consumers that have discretion have only the mind-to-world direction of fit.
> V. It follows from I-III that desires are biological representations with consumers that have discretion.

My theory of direction of fit will draw heavily on the framework of ideas and terminology provided by teleosemantic theories of representation, so I will also present extensive exposition and defence of this framework. This approach is in accord with my overall strategy for establishing the direction of fit of desire, which is to use a naturalistic theory of direction of fit that allows us to draw almost direct inferences from empirically-discoverable properties of desire, to conclusions about direction of fit. So once I have presented my argument for my theory, it will take little further work to reach my overall conclusion. It will, however, be a fairly arduous process to get that far, because of the need to explain the complex and controversial teleosemantic framework. This chapter and chapter 7 contain preliminary work for chapter 8; in chapter 8 I present and argue for my theory of direction of fit, which entails premise IV; and in chapter 9 I argue for premise V. By the end of chapter 9 my positive argument for the conclusion of this thesis will be complete, and in part III I will take on only the relatively minor task of describing how my views relate to the Humean Theory of Motivation.

In my view, direction of fit is very closely connected to representation. As I will explain, Ruth Millikan's teleosemantics (1984, 2004) claims that what makes an entity a representation, determines its content, and fixes its direction of fit is the way in which it mediates the interaction of *producer* and *consumer* systems, according to its function. Producers and consumers also have functions with respect to representations – producers are supposed to produce representations under certain circumstances, and consumers are supposed to behave in different ways, depending on the representations they consume – and these functions are crucial to direction of fit. According to my theory, what I will call *biological representations* have the mind-to-world direction of fit if and only if their producers have the function of producing them whenever some specific substantive condition holds, and have the world-to-mind direction of fit if and only if their consumers have the function of behaving in some specific way whenever they, the representations, occur. This means that desires lack the world-to-mind direction of fit, because what the system that consumes desires should do when any given desire occurs depends on what other desires the agent has at the time, and on their instrumental beliefs. Of course, I will develop these claims in much more detail later.

In the remainder of this chapter, I will give a more detailed introduction to the topic of direction of fit, outline the most familiar approach to the topic, and explain why it is not suitable for my purposes. This familiar approach lacks the advantages of the teleosemantic approach to direction of fit, which I will also describe. I will introduce some alternative terms for talking about direction of fit, and I will also explain some assumptions about how the notions of *mental states*, *representations*, *content*, *attitudes*, and *direction of fit* relate to one another.

In chapter 7, I will describe a version of teleosemantics, and defend it against some well-known objections. The topic of this chapter will be representation in general, rather than specifically direction of fit, but my account of direction of fit is too dependent on teleosemantic ideas, and those ideas are too subtle, for it to be possible to avoid this preparatory work. The objections to teleosemantics that I will respond to concern functional indeterminacy, which has been widely discussed since this problem was first raised by Fodor (1990); the ‘swampman’ thought-experiment introduced by Davidson (1987); and the liberality of some versions of teleosemantics, which has been the subject of more recent work (e.g. Burge 2010). These objections are often thought to be devastating to teleosemantics, and if they succeeded they would be very damaging to my theory of direction of fit. The topic of indeterminacy is particularly important, and it comes up again in the following chapter.

In chapter 8, I state and argue for my theory of direction of fit. After stating the theory, which is inspired by a proposal by David Lewis (1969), I first set out the scope of my claims, by describing what I mean by ‘biological representation’. I then argue for the theory, and discuss challenges related to the indeterminacy problem. In essence, my argument is that my theory does a better job than existing accounts of direction of fit of identifying a deep difference between two kinds of representation. As I will argue, the two directions of fit correspond to two different jobs that representations can do – saying how things are, and saying what to do – or to put it another way, two ways in which representations can contribute to the functioning of wider systems. My account accurately categorises representations in this respect, while orthodox views fail to do so.

Finally, in chapter 9 I return to the topic of desire. My main task in this chapter is to show that desires satisfy my criterion for the mind-to-world direction of fit, but fail to satisfy the criterion for the world-to-mind direction of fit. In the light of my argument from chapter 8, this reveals that desires contribute to action-selection in a fundamentally similar way to beliefs. In this chapter, I also discuss the nature of reward, because one possible objection to my view is that there are no suitably mind-independent facts about the reward levels of outcomes for desires to represent. So I show that it is possible to give a coherent account of reward which avoids this objection. In closing chapter 9, and with it part II, I recap my argument for my overall conclusion, that desires have only the mind-to-world direction of fit.

6.2 Introduction to Direction of Fit

Direction of fit is most often thought of as a property of mental states, which distinguishes beliefs and desires. However, the idea is applicable and important much more widely than this; many kinds of representation have directions of fit.

In fact, the underlying phenomenon of direction of fit is applicable more widely still. Whenever two things are supposed to fit one another there are various possibilities about how this fit should be achieved. For example, we often try to find shoes to fit our feet, but in the *Cinderella* story the Prince tries to find a foot to fit a shoe. These are two different ways of getting shoes and feet to fit: by starting with a foot and looking for a shoe that fits it, and vice versa. Sometimes fitting between pairs of entities matters, but neither of these approaches is taken, since neither entity is privileged in the right way – for instance, electrical plugs and sockets need to be designed to fit together, but are presumably usually developed simultaneously. So the situation is that when two entities fit one another in some way, there can only be a 'direction' to this fit if the entities are *supposed*, in some sense, to fit one another in this way. Also, it is not enough for them merely to be supposed to fit one another, but – as it were – the *responsibility* for achieving the fit must lie with one of the entities, and not the other. It must be the case that it is in some way correct or appropriate for one of the entities, and not the other, to be changed to achieve a fit.

As many philosophers have noted, beliefs clearly satisfy these criteria for having a direction of fit. Beliefs fit the world when the states of affairs which are their contents are actual states of affairs (see section 6.3 for more on my assumptions about the structure of mental states such as beliefs). Furthermore, they are supposed to fit the world – there is a sense in which beliefs succeed when they are true and fail when they are false – and the 'responsibility' for achieving this fit lies with the belief, not with the world. Beliefs that fail to fit the world should be abandoned and replaced with ones which do; it seems to get the point of believing wrong to try to change the world so that it fits one's beliefs. With this idea in place, philosophers have theorised that desires have the opposite direction of fit – that they too are supposed to fit the world in virtue of having actual states of affairs as their contents, but that this fit is to be achieved by changing the world, not by changing desires.

For example, consider my belief that Alpamayo is a mountain in Peru, and my desire to see it. Both of these mental states succeed in important ways if they fit the world, according to philosophical orthodoxy: my belief succeeds if Alpamayo is a mountain in Peru, and my desire succeeds if I see it. But the orthodox view claims that different kinds of action would be required if these conditions were not satisfied, and not just because of the practicalities. If my belief about Alpamayo was false, then I should abandon it, and it would be irrational for me to try to make it true by changing the world, just for the sake of having a true belief. That this is a matter of the nature of belief, and not of the practicalities of the case, can be illustrated by considering another: if I believe I will fail tomorrow's exam, then I can easily make this belief true, but I have no reason (not even a defeasible one) to do so. On the other hand, if my desire to see Alpamayo was unsatisfied, then I should not abandon it, but should travel to the Cordillera Blanca, other things being equal. So the thought is that beliefs are subject to, and perhaps defined by, norms concerning the circumstances in which they should be held, while desires are subject to, and perhaps defined by, norms concerning what those who have them should do. Among the best-known works discussing direction of fit in this tradition are Anscombe (1957), Searle (1983), Platts (1979), Smith (1987), and Humberstone (1992).

The direction of fit traditionally associated with beliefs is called the *mind-to-world* direction of fit, because in this case the 'mind' – in the form of some mental state – is supposed to be changed, if necessary, to fit the 'world'. The direction of fit traditionally associated with desires is called the *world-to-mind* direction of fit, because the 'world' is supposed to be changed to fit the 'mind'.

However, while philosophers have been particularly interested in direction of fit in the context of belief and desire, it also seems to be a crucially important feature of many representations. For example, consider the descriptive sentence, 'The door is shut,' and the imperative, 'Shut the door!'. These two sentences have different meanings, which are distinguished not by the states of affairs that we most naturally identify as their semantic contents, but by their directions of fit. Very plausibly, both of these sentences have the state of affairs of *the door's being shut* as their content, but they differ in meaning because of the different complex relations in which they stand to this state of affairs. We might make a first pass at distinguishing these relations by saying that the description is supposed to occur when the state of affairs holds, and the imperative is supposed to occur when some action ought to be taken to bring about the state of affairs (in both cases, presumably, the sentence should only be produced if the speaker reasonably judges that it is worthwhile to convey this to the listener). In particular, this way of thinking seems to be necessary to distinguish imperatives like, 'Shut the door!' from normative declarative sentences like, 'You ought to shut the door.'

Furthermore, direction of fit apparently figures in the same way in many simple representations. Animal signs, simple signals within many organisms, and human conventional signals all frequently seem either to say that some state of affairs obtains, or to tell the systems that consume them to behave in particular ways or bring about particular states of affairs, or to do both of these things. So the central representational properties of these entities can be accounted for by identifying their directions of fit and the states of affairs which are their contents. For example, honeybee dances may be representations telling fellow worker-bees that nectar is available at a given angle and distance, and also telling them to go there. This can be analysed by saying that the dances have, simultaneously, the mind-to-world direction of fit towards states of affairs like *there being nectar at angle A and distance x from the hive*, and the world-to-mind direction of fit towards states of affairs like *the other worker bees going to the point at angle A and distance x from the hive* (Millikan 1995). Turning to human conventional signals, red traffic lights have the world-to-mind direction of fit towards the state of affairs that *traffic on the road stops at the white line*, and doorbells have the mind-to-world direction of fit towards the state of affairs that *someone is at the door* (either of these signals might have the other direction of fit as well; to find out we need a theory of direction of fit). And regarding signals within organisms, there is an ongoing philosophical debate about the direction of fit of pains *qua* representations. Some writers have recently defended the view that they have the world-to-mind direction of fit with respect to states of affairs like *the person's keeping their weight off their ankle*, in opposition to the more immediately intuitive view that they have the mind-to-world direction of fit with respect to propositions like *the left ankle's being damaged* (e.g. Klein 2007). Incidentally, these examples show the limitations of the terminology of 'mind-to-world' and 'world-to-mind', and I will shortly revert to an alternative.

As well as potentially distinguishing beliefs from desires, then, direction of fit seems much more generally to distinguish different kinds of representation – roughly speaking, representations which say how things are, and others which say what to do. Those which say how things are succeed when they fit the world, while those that say what to do seem to succeed if they prompt their consumers to change the world to fit them. The examples also illustrate that representations can have both directions of fit, with respect to different states of affairs; some representations both say how things are, and what to do.

Not all representations are like this, however. Some representations lack direction of fit, because they are not supposed to fit the world at all. The clearest examples are individual words and concepts; these are contentful, but their contents do not correspond to states of affairs, so they cannot fit the world in the present sense. Imaginings and suppositions may also lack direction of fit, and it is hard to know what to say about the directions of fit of questions and fictions. But it is noteworthy that all of these representations enter into or are formed by relatively complex systems. Plausibly, the most basic forms of representation, requiring the lowest levels of sophistication, represent whole states of affairs in non-compositional fashion, and have one or both directions of fit.

6.3 Representations, Mental States, and Direction of Fit

In this chapter, my main aim is to argue that given my overall objectives, and given some points about the role of direction of fit in representation, teleosemantics offers the most promising way to think about direction of fit. However, before proceeding with this, I need to set out explicitly a number of assumptions about the relationships between representations, mental states, and direction of fit. Some of these assumptions just amount to terminological stipulation, but others are more substantive.

A terminological point on the use of ‘representation’ is that this and cognate terms are sometimes used to denote only what I would call representations with the mind-to-world direction of fit. An attraction of this way of using the term is that it follows from the thought that representations are entities that represent things as being certain ways; that is, that purport to say how things are. In contrast, I use the term ‘representation’ more broadly, as I have already suggested, to cover entities including imperative sentences and questions (which do not ‘represent things as being certain ways’), similar entities outside human language, words and concepts, and other things besides. Very, very roughly, representations are things that are *about* other things.

A more substantive issue concerns the structure of those mental states, such as beliefs and desires, that are often called ‘propositional attitudes’. It is very widely agreed that for an agent to believe that grass is green is for that agent to stand in some relation to the state of affairs of *grass being green*. We can call this state of affairs the *content* of the belief. Some readers may already be tempted to object that we believe propositions, not states of affairs, so believing that grass is green is actually a matter of being related to the proposition *grass is green*. But since I have not said that the relation that believers stand in to states of affairs is the believing-relation (the relation denoted by ‘believes’), and since whatever is related in some way to the state of affairs of *grass being green* is thus also related in an only slightly different way to the proposition *grass is green*, this objection is no threat to the position just stated.

Now, some philosophers question whether the belief that grass is green requires the presence of a representation with the content *grass is green* (e.g. Schwitzgebel 2002). I assume the opposite, that believing does require such representations. However, even among those who share this view, at least two substantially different accounts have been proposed of how representations are involved in relating agents to the states of affairs that they believe. I will first outline some more of my assumptions, then contrast them with a prominent alternative way of thinking about the same issues.

I assume that beliefs and other similar mental states are representations with directions of fit. What makes these representations beliefs is the roles they play in agents' minds. In humans, these representations are, at least typically, structures and instances of activity in our brains.[15] So the relation that those who believe that grass is green stand in to the state of affairs of *grass being green* is the following: they have representations playing the appropriate role in their minds, which have the mind-to-world direction of fit with respect to the state of affairs of *grass being green*. The state of affairs that is the content of the belief, is also the content of the representation – the representation *is* the belief. Also, crucially, the direction of fit of the belief is the direction of fit of the representation – and what it is for a mental state to have a direction of fit is for it to be a representation with a direction of fit.

These assumptions amount to one way of construing the idea that propositional attitudes such as beliefs are relations to propositions, via mental representations (Fodor 1975, Field 1978). However, another way of construing it does not identify mental states with representations, and therefore distinguishes the directions of fit of mental states from those of the mental representations that make them up. On this alternative view, the same representations may be involved in mental states with opposite directions of fit. For example, assuming for now the usual view that desires have the world-to-mind direction of fit, a mental representation with the state of affairs *I am eating an ice-cream* as its content could be part of a desire to eat ice-cream, or of a belief that the agent is eating ice-cream, depending on the 'attitude'

[15] If the extended mind hypothesis (Clark & Chalmers 1998) is correct then non-neural representations may also be beliefs.

taken to that representation – as it is sometimes put, whether the representation was in the mind's 'desire box' or its 'belief box'. So the same representation could be part of a desire, with the world-to-mind direction of fit, or of a belief, with the mind-to-world direction of fit. On this view the representation itself would presumably lack any direction of fit – it would merely *stand for* the state of affairs *I am eating an ice-cream*, rather than saying that I am eating an ice-cream.[16]

This way of thinking of things may be a consequence of the thought that what it is to believe a proposition is to stand in a certain relation to a representation with that proposition as its content. However, this thought is also true on my assumptions; what it is to believe a proposition is to have a belief, which is a representation with that proposition as its content. Also, it is sometimes said that the content of a belief is a representation, and this way of putting things coheres relatively well with the present view. To say this, we would have to distinguish between the content of the belief, a representation, and the content of the representation, which would be a state of affairs or a proposition. But this last point is not crucial to the approach I am describing; what is crucial is that it distinguishes the directions of fit of mental states from those of mental representations.

These assumptions matter because part of my overall strategy is to develop a theory of direction of fit for representations of a certain kind, then apply this theory to desires. This approach would fail if desires were not representations, or if direction of fit for mental states such as desires was in some way a distinct phenomenon from direction of fit for representations. But I do not assume that beliefs and desires are representations with directions of fit just for the sake of expediency; in my view, this picture is much more attractive than the alternative. On any plausible version of the alternative approach, what it would take for a representation to be in the 'belief box' would be for that representation to play a certain role in the mind, and playing this role would be, on any plausible theory of direction of fit for representations, sufficient to entail that the belief was a representation with the mind-to-world direction of fit. So it complicates matters

[16] Compare what the string 'grass is green' does in the sentences, 'Grass is green' and 'If grass is green, then snow is white'. The former asserts that grass is green. The latter uses the same string to stand for the state of affairs of *grass being green* in the service of asserting a conditional. Both sentences have the mind-to-world direction of fit, but the string 'grass is green' in the latter sentence does not have a direction of fit.

unnecessarily to think of the belief as anything but the representation, and to distinguish the direction of fit of one from that of the other.

6.4 Two Approaches to Direction of Fit

In this section, I will argue for the use of a teleosemantic framework in theorising about direction of fit, by comparing that approach to another, more familiar non-normative approach.

So far, I have described the directions of fit in fairly imprecise, normative terms. For instance, in section 6.2 I suggested that representations with the mind-to-world direction of fit are those that are supposed to fit the world – meaning that their contents are actual states of affairs – with the 'responsibility' for achieving the fit lying with the representation, or perhaps more literally with whatever produces it. But I hope to provide a much more precise account of direction of fit than this, and also one that distinguishes the directions of fit without using normative terms. Such a theory would allow us to infer a conclusion about the direction of fit of desire relatively directly from the facts about desire described in part I. A non-normative theory is also an attractive goal because such a theory of direction of fit could help to show how apparently normative features of mental states and other representations (such as their being capable of forms of success and failure) are grounded in their non-normative properties. Both of the two approaches that I will describe have been taken by philosophers with aims of this kind: the teleosemantic approach as part of a reductive theory of representation, and the more familiar approach in the context of moral psychology.

The familiar approach to direction of fit is exemplified by the work of Michael Smith (1987, 1994) and I. L. Humberstone (1992). Smith uses his account in defending a Humean theory of motivating reasons, and I will discuss this topic in much more detail in part III. According to Smith, the direction of fit of a mental state with content *p* depends on how it is affected when the agent has a perceptual experience with the content *not-p*:

> *Smith's Theory of Direction of Fit*: A mental state has the mind-to-world direction of fit with respect to *p* if and only if it tends to go out of existence when *not-p* is perceived.

> A mental state has the world-to-mind direction of fit with respect to *p* if and only if it tends to endure when *not-p* is perceived, and dispose the agent to bring about *p*.

Meanwhile, Humberstone proposed the following theory, partly in response to Smith:

> *Humberstone's Theory of Direction of Fit*: A mental state has the mind-to-world direction of fit with respect to *p* if and only if it is regulated by a conditional intention to have it only when *p* is the case.
>
> A mental state has the world-to-mind direction of fit with respect to *p* if and only if it is regulated by a conditional intention to bring about *p* when one has this mental state.

These two theories have been criticised on various grounds (e.g. Sobel & Copp 2001), but I will not discuss the specifics of either. Instead, I will argue that they are both instances of a general approach to direction of fit which is less promising than the teleosemantic approach. The two theories are both intended to distinguish the directions of fit of mental states – as Smith and Humberstone thought of it, to identify the fundamental difference between beliefs and desires, broadly conceived – rather than to apply to representations more generally. Both theories appeal to psychological states – perceptions and intentions, respectively, and Smith's also appeals explicitly to representational properties, since it involves the notion of perceptual experiences with specific contents.

The teleosemantic approach to direction of fit is very different. Teleosemantics is a group of reductive, naturalistic theories of representation, which have in common the idea that biological functions are vitally important for understanding representation. Several authors have proposed teleosemantic theories of representation, including Millikan (1984, 2004), Papineau (1984, 1993), Dretske (1988), Neander (1995, 2013), Price (2001) and Shea (2007), and many others have suggested revisions to these proposals. The teleosemantic project starts from something like the thought that what makes beliefs representations, which can succeed or fail by being true or false, is not the causal properties that they actually exhibit or are disposed to exhibit, but their biological functions; that is, what they are *for*, from a biological point of view. This insight is thought to account for the apparent normativity of representation, while also facilitating the development of a non-normative, reductive theory, because teleosemantic theorists also adopt non-normative theories of biological function. Most commonly, teleosemantic theorists argue that an entity's biological function, if it has one, is to do what similar entities did in the past, that contributed to the survival and reproduction of wider systems in which those entities are embedded. Teleosemantic theorists do not claim that all representations should be understood in terms of specifically *biological* functions, but do suggest that this reductive account of biological functions can be generalised to functions of other kinds, and that a theory of simpler biological representations can be a valuable step towards a more complete theory of representation.

Direction of fit takes a central role in this project, because representations with different directions of fit seem to be *for* different things; they seem to have functions of different kinds. Those with the mind-to-world direction of fit are for saying how things are – for making information about the world available to some co-operating system, perhaps – while those with the world-to-mind direction of fit are for saying what to do – for controlling the behaviour of co-operating systems. This would explain why truth seems to be the standard of success which is characteristic of representations with the mind-to-world direction of fit, and satisfaction seems to be the characteristic standard of success of those with the world-to-mind direction of fit. It also follows that, if representations are distinguished from other entities by characteristic functions, then the two directions of fit must be two ways of being representations. According to the teleosemantic approach to representation, one way for an entity to be a representation with the content *p* is for it and its producer(s) and consumer(s) to have the functions characteristic of the mind-to-world direction of fit with respect to the state of affairs that *p*. Another way is for the entity, its producer and consumer to have the functions characteristic of the world-to-mind direction of fit with respect to *p*. So an entity's status as a representation, its direction of fit, and its content, are all determined by the kind of function it has with respect to some state of affairs.

Teleosemantic theories of representation therefore typically take a disjunctive form. They say that representations are things that have one or other of two functions; representations with one direction of fit have one kind of function, and those with the other direction of fit have the other. These functions are not typically incompatible, so it is possible for representations to have both directions of fit on most teleosemantic theories. Teleosemantic theorists do not typically present themselves as giving theories of direction of fit, but instead of giving theories of representation in general – which for them implies an analysis of the directions of fit, as an indispensible part of the wider theory.

The teleosemantic approach has three key advantages over the more familiar approach of Smith and Humberstone. First, the thought that what it is to have one or other of the directions of fit is to have a function or purpose of a certain kind is highly attractive, and the teleosemantic approach focuses very specifically on this idea. Humberstone's theory shares some of this advantage, since it concerns the agent's intentions for their own mental states, and thus, we might think, the functions or purposes to which agents put those states. But Humberstone's theory cannot capture the plausible idea that sub-personal representations have functions or purposes, and that these are crucial to their properties as representations. For example, we do not *intend* to feel pain only when our bodies suffer damage. Humberstone's theory also seems to entail that animals can only have beliefs or desires if they are capable of conceiving of, and forming intentions about, their own mental states. So the advantage of teleosemantics here is that it promises to analyse the directions of fit in a way which captures the idea that they are tightly connected to functions or purposes, but does not rely on agents' attitudes towards representations in doing so.

Second, a teleosemantic analysis of direction of fit has the potential to make a significant contribution to theories of content and of the status of entities as representations, as a result of the central place that teleosemantic theorists envisage for direction of fit in the theory of representation. In particular, the apparent normativity of representation may be explained by representations' having characteristic functions or purposes, and theories of direction of fit should aim to illuminate these functions. One reason why the familiar approach cannot offer this advantage is that, at least in the theories proposed by Smith and Humberstone, it relies on claims about the contents of representations (perceptions and representations) in analysing the directions of fit.

Third, the teleosemantic approach offers the chance to develop a more widely-applicable theory of direction of fit. As I have suggested, direction of fit does not seem to be a property only of mental states, but of representations of many kinds.

The teleosemantic approach may allow us to give a unified theory of direction of fit for a relatively wide range of representations. This theory is unlikely to cover representations of *all* kinds, as I will explain in section 8.2, because the variety of kinds of representation is too great. But it will be possible to give a theory of direction of fit that is applicable both to desires, and to many representations which are not mental states, expressed in terms of their functions and those of their producers and consumers. This greater breadth, in connection with the previous two advantages, suggests that the teleosemantic approach will offer a deeper account of the nature of direction of fit than would be possible on the familiar approach.

In this chapter, I have prepared the ground for the detailed examination of direction of fit to come in following two chapters. In introducing the topic, I have suggested that direction of fit is common to many kinds of representations, not just beliefs and desires. I have also explained my assumption that the directions of fit of mental states such as beliefs and desires are those of the representations with which they are identical. I have introduced the teleosemantic approach to direction of fit, and argued that it is more attractive than the more familiar approach – that it offers exciting prospects that the familiar approach does not. However, like my choice to investigate desire empirically, as a natural kind, this choice should be judged by its results. I hope to show in the remainder of this thesis the value of the insights that a teleosemantic approach can deliver in the study of direction of fit, and of desire.

As I noted in section 6.2, 'mind-to-world' and 'world-to-mind' are not entirely satisfactory names for the directions of fit, because they are easily mixed up, and because direction of fit is not only a feature of mental representations. So from now on, I will turn to an alternative. Several other pairs of expressions are sometimes used for this purpose, of which 'indicative' and 'imperative' is most common in teleosemantics. I will adopt this terminology; representations with the mind-to-world direction of fit will be called *indicatives* or said to have *indicative content*, while those with the world-to-mind direction of fit will be called *imperatives* or said to have *imperative content*. Representations that have only indicative content will be called *pure indicatives*, those that have only imperative content will be called *pure imperatives*, and representations with both directions of fit will sometimes be called *bi-directional*. An important point to bear in mind about this terminology is that it should not be taken to imply any particular claims about sentences in the indicative or imperative moods; I am open to the possibility that, for instance, some sentences in the indicative mood could be pure imperatives as far as direction of fit goes. A further important point is that the expressions ‘indicative content’ and ‘imperative content’ should not be taken to imply that there are two different kinds of content. I am assuming that representations have both directions of fit with respect to states of affairs. Instead, these expressions are appropriate because the directions of fit are two ways of having content, and representations which have both directions of fit have both indicative and imperative contents, which need to be distinguished.

Chapter 7: Teleosemantics Described and Defended

7.1 A Version of Teleosemantics

In this chapter, I present a fairly orthodox version of teleosemantics, and defend it against three well-known objections. At the end of the chapter I will also briefly explain why other notable objections to teleosemantics are not relevant to my project. Among the various teleosemantic theories that have been proposed, the one I will present is most similar to Millikan's original theory, from her *Language, Thought and Other Biological Categories* (1984).

My objective in this chapter is not to provide a complete naturalistic theory of representation. Instead, I aim to introduce the technical terminology of teleosemantics, explain how its core ideas fit together, and show that the teleosemantic project as a whole is not defeated by any of the three objections I discuss. So at times in this chapter I will appear to advocate claims that I do not accept, or at least not without qualification. For instance, it will be convenient for me to discuss how the content of representations is determined, even though my concern is with direction of fit, not content, and even though the theory I will present would, in my view, have to be made far richer to amount to a plausible overall theory of content. I will hardly discuss concepts, let alone issues like sense and reference, vagueness or context-sensitivity. Something similar is true of representational status – that is, of what makes entities representations; my view is that the teleosemantic framework I will present here offers important insights on this issue, but needs to be qualified and supplemented with more detail to be plausible as a complete theory. Also, I will describe Millikan's criteria for the directions of fit in this chapter, even though I will go on to argue in chapter 8 that her criterion for imperative content is not correct.

The first component of teleosemantics is a naturalistic account of functions. As I have already suggested, theories of representation need to have some way of capturing the idea that representational status, direction of fit and content are all tightly connected to apparently normative facts about what representations are supposed to do, and teleosemantics uses its account of functions to do this job.

According to Millikan (1984), an entity has a *direct proper function* if and only if it is one of a set of similar entities that are reproduced from one another, and some explanation of the existence of current members of this set can be given by reference to the properties or behaviours of past members. The functions of such entities are the properties or behaviours that would be cited in these explanations. The paradigm example of an entity with a direct proper function is an adaptive biological trait or component, like a heart; hearts are reproduced in similar forms, and this can be explained by the contribution they make to animals' survival and reproduction. The function of the heart, at least pending discussion of a very important objection below (section 7.2), is to pump the blood, because this is what past hearts have done which explains the existence of present hearts, by explaining how they have contributed to the survival and reproduction of our ancestors. The heart does not have the function either of doing things which hearts have done occasionally in the past, which contributed to survival and reproduction but only by chance, or of doing things which hearts have consistently done, but which have not made significant contributions to survival and reproduction (like making rhythmic thumping sounds); in both cases this is because describing these behaviours would not make for a good explanation of hearts' present existence. So direct proper functions are very tightly related to *teleological explanations*, which are explanations of entities' existence and properties in terms of what things of their type are capable of doing, but these functions are still determined by actual causal histories.

However, many entities that are not components of biological systems which have been shaped by Darwinian natural selection also have functions. For one thing, behaviours (as opposed to components), both of whole organisms and of parts of organisms, can be adaptations and can therefore have direct proper functions. In addition to this, processes other than natural selection can arguably create the right conditions for teleological explanations, and hence for direct proper functions. An example of particular interest is reinforcement learning: if a response to a stimulus will tend to be repeated when it leads to reward, and not otherwise, then present states and processes that contribute to this response will have direct proper functions, because we will be able to explain their existence by explaining how they contributed in the past to getting reward, and hence to their own persistence. Also, as well as direct proper functions, entities may also have *derived functions*. Derived functions occur when systems with direct functions produce further items, that are themselves supposed – according to the functions of their producers – to do certain things. An example of an item with a derived function is an antibody for a novel pathogen; this antibody will have the function of binding to and contributing to the destruction of that specific pathogen, because the function of the system that produced it was to produce antibodies that behave in this way. This type of antibody thus has a function even though there are no ancestral antibodies of the same type. Functions from learning and derived functions are potentially important to teleosemantics, because representations are frequently novel. This account of functions rightly leaves open the possibility that *devices* (by which I mean, entities with functions) can have more than one function. For example, the human tongue has functions in both speech and eating; it makes more than one kind of contribution to survival and reproduction.

Millikan's next step is to claim that representations must have *producers* and *consumers*. This move is attractive on the grounds that representation seems to be fundamentally communicative, and it also helps in developing a clearly-specified criterion for indicative content that is not based primarily on the notion of information (the advantage this brings will become clearer shortly). Producers are devices which have the function of causing representations to come into existence; but the more important notion is that of a consumer. Consumers are devices that are not just causally affected by representations, but which have the function of behaving in different ways, depending on the occurrence and properties of those representations. Consumers are said to have *relational functions* with respect to the representations that they consume, and to be *adapted*[17] by these representations, when they occur. To determine whether or not a given device or behaviour is a representation, we need to say whether it has a consumer, which means being able to distinguish devices with relational functions from those with constant functions. To illustrate this distinction, the skin, bones, ligaments and tendons all seem to have constant functions – they work just by holding things in place in various ways. However, the muscles have relational functions – what they are supposed to do depends on activity in the neurons that innervate them, or possibly on more distal

[17] This is a technical term, introduced by Millikan, that I will use repeatedly. Representations adapt their consumers when they occur, and the functions of consumers are described by saying what they are supposed to do when adapted by representations. It should not be confused with the more common biological use of 'adapted', as in 'adaptation', meaning a device or behaviour that contributes to survival and reproduction.

states, like the location of a target object. This means that firings of the neurons that innervate the muscles are at least candidates for representational status. For example, firing of the musculocutaneous nerve adapts the biceps, which has the function of contracting when adapted by events of this type. The biceps muscle has a relational proper function, and is a consumer of activity in this nerve, because it has the function of behaving in a different way when the musculocutaneous nerve is inactive.

A consequence of the claim that representations have consumers is that representations must vary, or to be more precise, must have varied in ancestral cases. That is, any given representation must sometimes occur and sometimes not, or must have different properties from one occasion to another. Otherwise there could be no device with the function of varying its behaviour, depending on the state of the representation. The different behaviours that the consumer performs will contribute to survival and reproduction under different conditions, so for the systems concerned to work, representations must co-occur sufficiently reliably with these conditions, and this co-occurrence is achieved by the producer. So at this point we already have an interesting account of the nature of representation. According to this version of teleosemantics, representations are things which fulfill their functions by influencing the behaviours of consumers, and specifically by helping to calibrate the behaviour of the consumers to further conditions. In many cases, they contribute to the survival and reproduction of organisms in this way, and the 'further conditions' are things going on 'out in the world'.

This is the framework within which Millikan gives her criteria for the directions of fit, and her theory of content. If one of the functions of a representation is to cause its consumer to behave in a specific way or to bring about a specific outcome, then they have imperative content; they tell their consumers to behave in that way, or to bring about that outcome. So on the present view, honeybee waggle dances seem to be representations with imperative content. Their consumers are other bees which watch the dance, or perhaps systems within these bees; either the bees have the function of behaving differently depending on whether they see a dance, and on its properties, or systems within them have the function of causing them to behave differently depending on the dance. And these dances contribute to the survival and reproduction of the bees because they cause the watching bees to fly in search of nectar at specific directions and distances from the hive, which correspond to features of the dance (the number of waggles and the angle at which it is performed). So the dances have the function of causing flights that correspond to them according to this pattern, and consequently have imperative content, according to the present criterion. They say things like: *Fly to the place at angle A and distance x from the hive!*

The criterion for indicative content involves a further technical notion, which is that of a *normal condition* for a type of behaviour. Normal conditions are those under which behaviours have taken place in the past, which would be mentioned in the best explanations of the persistence of those behaviours. So they are the conditions under which the behaviours have typically succeeded – *not* those under which the behaviours are most often performed. A representation has indicative content if one of its functions is to occur at the same time as some specific normal condition for the behaviour it prompts in its consumer, and its indicative content is that normal condition. That is, indicative representations contribute to the success of the systems in which they are embedded by occurring at the same time as states of affairs that are relevant to what their consumers should do. They say that these states of affairs obtain, and are true when they do, and false otherwise. A key consequence of this account of indicative content is that it means that consumer behaviours in response to true representations will typically succeed, while those in response to false representations will typically fail. This criterion apparently implies that waggle dances also have indicative content, because there being nectar available at the places the dances direct co-operating bees to fly to is a normal condition for the success of their flights, and occurring when this condition holds is among their functions.

One reason why normal conditions are interesting is that they figure in explanations of how producer-consumer systems come to be reproduced repeatedly (Godfrey-Smith 2013). Consumers evolve to respond to the behaviour of producers precisely when those producers behave in ways that sufficiently reliably correspond to conditions that make the difference to the success of the consumers' behaviour. This can also influence the evolution of producers, since producers will be selected for if they behave in ways that facilitate successful behaviour by consumers, provided that the evolutionary interests of producers and consumers are sufficiently closely connected. Another reason is that on the present view representations will almost always carry information about the states of affairs they represent, in the sense of making those states of affairs more likely (see Shea 2007), but will not in general represent what they carry most information about – the states of affairs with which they are most closely correlated. This is in part because false positives may in some cases be less costly than false negatives (Godfrey-Smith 1992). For example, the theory implies that beaver tail-slaps have the indicative content that *danger is present*, because the escape behaviour that these signals cause has the success condition that the beavers in the community are threatened; escape behaviours where no danger is present have small but real costs. However, beavers may have evolved to produce tail-slaps even when the evidence of danger is weak, if the costs of escaping when danger is not present are sufficiently small, relative to the costs of not escaping when danger is present. So tail-slaps may be better correlated with rustling in the trees than with the presence of danger, but still represent the latter on the teleosemantic view.

A useful way to see the value of these ideas is to consider what they add to the thought that representation has something to do with information. For the purposes of this discussion, I will adopt the standard definition of information in this context, which is that one event or state of affairs carries information about another if it makes the other more likely (i.e. if the probability of the latter given the former is greater that the unconditional probability of the latter; Shannon 1948). I will also focus on indicative representations, which are widely taken to be paradigmatic. If we assume that this is the case – that indicatives are paradigmatic – then representation certainly does have something to do with information, because indicatives are only useful if they make what they represent more likely (leaving aside complex cases involving deception). But the naïve theory that whatever carries information about some state of affairs represents it is hopeless, because information is ubiquitous. The great strength of teleosemantics is that it helps with the problems this causes in an elegant, principled way.

First, carrying information about some state of affairs is obviously not sufficient for representing it, because everything carries information about something, and not everything is a representation. Even carrying information that is used and useful is not sufficient; the presence of clouds carries information about the chance of rain, but does not represent it. Teleosemantics takes a major step towards solving this problem by appealing to functions. It is far from obvious that there are any entities that have the function of carrying information about some state of affairs, but are not representations. And this move is highly intuitive, because it seems that what distinguishes representations from other phenomena which we (and other animals, and our sub-systems, and other things) use to draw inferences about further states of affairs is that representations are created for this purpose. Also, the idea that consumers are necessary for representations helps to clarify this thought. Nothing can contribute to the success of a wider system just by carrying information, unless the information-carrying event also causes further events that benefit the wider system, and these effects are mediated by some further system. It is the presence of consumers that makes it possible to identify information-carrying, rather than causing some beneficial effect at an appropriate time, as the distinctive function of representations.

Second, thinking about what information a representation carries is not particularly helpful in identifying its content, but the notion of normal conditions helps to deal with this problem. Although the information carried by specific representations is important for understanding how they work (Skyrms 2010, Sutton 2013), the consumer-focused approach advocated by teleosemantics allows us to both explain how producer-consumer pairs using representations to co-operate emerge, and state truth-conditions for indicatives. Any representation will carry information about many different states of affairs, for several reasons. For example, consider the beaver’s tail-slap. We have already seen that this carries information about the presence of danger, and about rustling in the trees. But it also carries information about more or less specific states of affairs, such as *wolves being present* and *animals being present*; about more proximal causes of the signal such as specific patterns of stimulation to the beavers’ sense organs; about effects of the signal, such as the beavers’ performing escape behaviours; about background conditions for the signal, such as the presence of oxygen in the environment; and about disjunctions of possible causes of the signal, such as *danger being present or the wind rustling the trees*. So if we are to identify the truth-conditions of representations from all of these, we need a principled reason to select one state of affairs over the others. Millikan’s idea offers such a solution. The normal conditions for consumer behaviours are different from the other states of affairs that representations carry information about because they explain successful behaviour; they make it the case that true representations explain success and false ones explain failure. They are also good for explaining why producers and consumers are configured the way they are, because it is the fact that representations carry information about normal conditions that explains why it is useful for those representations to be produced, and for consumers to respond to them in the ways they do. So the three main innovations of Millikan-style teleosemantics – functions, consumers and normal conditions – all seem to be highly promising insights for understanding the nature of representation.

7.2 Indeterminacy

In this section, I will discuss the first of the three objections to teleosemantics to be addressed in this chapter. Among the most challenging and best-known objections to teleosemantics is that it fails to identify determinate contents for many, if not all representations (Fodor 1990), and a similar problem also confronts teleosemantic accounts of direction of fit. So here I will explain why indeterminacy problems in general are not fatal for teleosemantics, showing how they arise both for accounts of content and accounts of direction of fit, and in the next chapter I will return to the issue as it applies to my own theory. Although there are various ways to develop this objection, they are all derived from the same problem, which is that for any device there is more than one way to explain how it contributes to the survival and reproduction of a wider system. Since functions are the properties or behaviours that are cited in such explanations, this variety of explanations seems to imply indeterminacy of function, which would in turn imply indeterminacy of content and direction of fit. For the sake of easier expression, I will write as though teleosemantics is only concerned with biological functions, but similar points are likely also to apply to non-biological functions.

This objection is often discussed using the example, introduced by Fodor (1990), of a frog that is disposed to try to catch and eat any small dark object that flies near it. We assume that the frog produces internal states of a certain type when it sees things like this, that in some way encode the position, direction and speed of the objects, and that these internal states cause the frog to flick out its tongue in the appropriate direction. We can also assume for the sake of argument that this is a reflex, so no other internal representational states are involved in controlling the action, and we can ignore the question of how the internal states represent position, speed and direction, in favour of thinking about what they represent as moving in that way – so we will also ignore any potential imperative content. The question is whether a given internal state of this type – call it R – represents the presence of *food*, *flies* or *small dark moving things*, or perhaps something else again.

At least looked at pessimistically, the problem is particularly acute because there are principled arguments concerning R's content that seem to point in different directions. Recall that according to the theory, R's indicative content is some specific normal condition for the successful performance of the behavior that R causes; that is, a condition that would be cited in a good explanation of how the R-caused behaviour typically contributed to the frog's survival and reproduction. So the content of R is something like *there is a _____ at location x travelling with speed y and angle z*, and we are interested in what fills in the blank. Among the possibilities are:

i) *fly*: For the sake of argument, let's assume that this mechanism was typically used to catch flies in the frog's ancestral environment, and that flies are a natural kind. Then it seems to be a good explanation of how the tongue-flicking behaviour, as prompted by R, contributed to the frog's survival and reproduction to say that Rs occurred when flies were present at locations systematically correlated with the different forms that R can take. Still, this explanation works because we have a certain amount of background knowledge: we know that flies are usually good food for frogs, and that flies are, in virtue of their size, solidity, colour, and typical rate of movement, fairly easy things to see – so we can understand the ability of the frog's visual system to generate Rs.

ii) *small, dark moving thing*: If we replace talk about flies with talk of small, dark moving things in our explanation of how tongue-flicking works, then we no longer need the background knowledge that flies are easy to see, so our explanation has the advantage that it makes it more transparent how the task to be performed is possible. This is, in general, an important theoretical virtue. However, this replacement does have the cost that the explanation now relies on the background knowledge that in the frog's ancestral environment, small, dark moving things were often flies.

In order to see the full case for this line, we need to step back from the frog example, and notice that typically internal devices in biological systems contribute by working together with many others. They are parts of hierarchically-organised systems, with corresponding hierarchies of functions. For example, the molars ultimately contribute to our survival and reproduction by maintaining our health; they are part of a system that does this by ingesting nutrients; and also of a sub-system that takes food into the body and prepares it for the absorption of individual nutrients; and of a sub-sub-system that breaks the food into smaller and softer pieces; and of a sub-sub-sub-system that grinds the food; and within this system they act as the grinding surfaces, rather than producing the necessary movements. So among the things the molars do which contribute to survival and reproduction, and are hence among their putative functions, are *maintaining health*, *ingesting nutrients*, *grinding food*, etc.. Karen Neander (1995) points out that from this hierarchy, we can identify the function that is specific to a device, which is the one that it most immediately performs – perhaps *acting as a grinding surface* in the case of the molars. A useful way to think about this, she suggests, is to think about what kind of failure would imply malfunction on the part of the specific device we are interested in. Now, if this is a good way to overcome functional indeterminacy, it pushes us towards using the most immediate function of R to fix its content, and Neander argues that the most immediate function of R is detecting small, dark moving things. This is something R does that requires minimal co-operation either from other parts of the system, or from a benign environment.

iii) *healthy frog food*: An alternative way to adjust the explanation given in i) is to try to avoid using the background knowledge that flies are good food for frogs, by replacing talk of flies with talk of healthy frog food (Goode & Griffiths 1995). This makes it less transparent how the system concerned is able to detect the presence of what it represents, but more transparent how catching this thing will contribute to survival and reproduction. In addition to increasing the transparency of that aspect of the explanation, this also increases its robustness, because there are possible cases in which frogs catch flies but this makes them less healthy, or in which they catch things other than flies that make them more healthy. For example, Papineau (1998) mentions flies which are poisonous or cause allergic reactions. Moving in this direction (which might take us further, to thinking of R as representing *reproduction-enhancer* – Papineau 1998) tends to make it the case that the truth of indicatives ensures the success of the behaviours they cause, and is therefore in keeping with the spirit of teleosemantics, which in its canonical form takes indicative content to be determined by success conditions rather than by the information that representations carry.

In addition to these, a further possibility is that R represents the presence of a characteristic pattern of firing on the frog's retina. A potential advantage of this view is that we might say that the function of the frog's eye is to produce the characteristic firing pattern when flies are present, and that the function of the system that produces R is to produce it when this pattern occurs – this would be to think of the eye and the producer of R as co-operating sub-systems of a wider system, whereas the other proposals seem to treat the wider system as R's producer. This is a slightly different version of the indeterminacy problem, because it takes R to represent a different event rather than the same event under a different description. Overall, in the face of this sort of example, it is hard for teleosemantic theorists to deny that there is a serious threat to their theory of representation from functional indeterminacy. The relevant phenomena admit of many good explanations backed up by principled explanatory strategies.

Before giving a response to this objection, I will outline one further example, in less detail, in order to show that the same problem is also relevant to direction of fit. Imagine now that the frog also uses an internal state S, which causes it to croak, and that there are a range of different situations in which it does this: it does it whenever it sees a rival that might invade its territory, as a warning, and also when it gets dark, to advertise its presence to potential mates, and also when predators approach its spawn, as a way of distracting them. Now S clearly has imperative content on the teleosemantic account, in that it tells the frog's motor systems to produce a croak. But whether or not S also has indicative content seems to be indeterminate, because some explanations of how it works would describe a normal condition for croaking with which it has the function of co-occurring, but others would not. One way of describing the normal condition for croaking would be disjunctive, giving S the indicative content: *a rival is near or a potential mate is near or something is threatening the spawn*; but a disadvantage of this is that there will be no system in the frog for identifying when this disjunctive condition holds, other than the systems for detecting each disjunct. Another way of describing the normal condition would just be to say that the situation is such that croaking is appropriate, but this does not explain in any substantive way how croaking succeeds. We should be wary of indicative content like *the situation is such that croaking is appropriate* because if we allow it in all cases we risk having no account of how pure imperatives are possible. I discuss this kind of case and the issue of indeterminacy specifically in direction of fit further in the next chapter; the key point for now is that I cannot dismiss the present objection just by saying that my interest is in direction of fit and not content.

A second point worth noting before turning to my main response to the objection is that it is not at all clear what level of determinacy of content a good theory should provide in any given case. We should not assume unquestioningly that representational content in simple biological cases will always be easy to capture in human languages, or that contents which can be described easily in such languages are therefore perfectly determinate, or that perfect determinacy is a virtue. In particular, the existence of vague predicates in human languages suggests that this issue is not straightforward.

I turn now to my main response to this objection, which is that the objection only shows that biological functions have not yet been adequately defined by teleosemantic theorists, not that they are themselves indeterminate. So while the objection may show that teleosemantics needs to say more to specify exactly what determines the functions that ground content, it does not show that a radical change of direction is needed. In order to bring out this response, I will first describe how the teleosemantic account of functions could be strengthened to avoid indeterminacy, then argue that we do not currently have sufficient grounds to abandon realism about determinate biological functions.

On the first of these points, attempts have already been made by Neander (1995, described above) and Carolyn Price (1998) to give more precise accounts of functions. I will describe Price's work shortly, but first we should note that the present account of functions is weak – it says only that a function is any property or behaviour of a device that contributes to the success of a wider system, and hence explains its existence – and that resources are certainly available to strengthen it. In particular, I described the case for indeterminacy made above as a ‘pessimistic’ perspective on the point that principled approaches to explanation pull us in different directions, and there does seem to be an optimistic interpretation of this point available as well. Given that there are a range of explanatory virtues that we can use to evaluate proposed functions, the materials are there for an account of function that maintains the connection with explanations, but gives more functional determinacy. We simply need to explain how to balance considerations like making it easy to see how systems contribute, making it easy to see how it is possible for them to perform their functions, and avoiding duplication of functions (for instance, not attributing the same functions both to systems and their own sub-systems).

Price’s account follows broadly this approach, and while it may not entirely solve the indeterminacy problem, it at least shows us how to go about solving it. She claims that in addition to being cited in an explanation of how a device contributes to the survival and reproduction of a wider system, functions also satisfy the following four conditions:

- *Immediacy*: if an activity *a’* of a given device contributes to the success of a broader system only by facilitating a further activity *a* of the same device, then doing *a’* is not part of the function of that device. For instance, growing is not part of the function of a flower, even though the flower’s growing is a condition of its contributing to the success of a plant.
- *Independence*: the function of a device is something that device is capable of doing on its own. The function of the heart is not to distribute oxygenated blood to the muscles, because it can only do this with the co-operation of the lungs and blood vessels.
- *Abstractness 1*: the function of a device is to produce an effect,[18] rather than to produce that effect via a specific mechanism or process. The function of an axon is not to quickly propagate an action potential to a set of synapses using a myelin sheath, because axons could in principle make the same contribution to neural systems in a different way.

[18] Many teleosemantic theorists take the view that functions must involve producing effects, and therefore that representations cannot have functions like carrying information about, or occurring at the same time as, particular states of affairs (see Millikan 1990). They argue that functions must be constituted by causal contributions to successful outcomes. But in my view co-occurrence with normal conditions is the way that indicatives causally contribute to success, even though it cannot readily be characterised in terms of producing an effect.

- *Abstractness 2*: the function of a device is correctly specified in a way that assumes that the device's fellow components are also performing their functions, whatever those functions are. The function of the heart is not to pump clot-free blood, because other mechanisms have the function of ensuring that the blood is clot-free.

A notable feature of these four conditions is that with the exception of *Immediacy*, they are instances of a more general principle, which is that different devices that are parts of the same system do not have the same function.[19] This principle has the obvious explanatory virtues that it avoids leaving any device apparently superfluous, or making any uninformative functional claims. Respecting this principle requires careful attention to the system/sub-system and fellow-component relations in which devices stand, as Price also emphasises. An example will illustrate the links between Price's conditions and the general principle.

The arteries, veins, and the heart all have distinct functions, which contribute to the functioning of the circulatory system, of which they are all sub-systems; and the circulatory system works together with the respiratory system. *Independence* tells us that the arteries, veins and heart have functions that they perform on their own, without relying on each other: so the heart pumps the blood, the arteries channel the movement of the blood to the muscles, and the veins channel the movement of the blood back to the lungs. *Abstractness 1* tells us that the circulatory system as a whole does not just do the conjunction of these things, but instead does something more abstract, that could in principle be achieved in more than one way: it circulates the blood between the muscles and the lungs. And *Abstractness 2* tells us that the circulatory system does not have the function of carrying *oxygenated* blood to the muscles, because if the circulatory system was already doing this, then the role of the respiratory system, which is not a sub-system of the circulatory system, would not be clear. The function of the respiratory system is to oxygenate the blood and remove carbon dioxide from deoxygenated blood; that of the circulatory system is to move the blood around; and the function of the wider system that involves both is to supply oxygen to the muscles. So together, *Independence*, *Abstractness 1*, and

[19] Apart from in systems that actually do have multiple distinct devices for doing the same thing; it's important not to slip into Panglossian assumptions when thinking about biological functions.

Abstractness 2 allow us to identify with a high degree of precision just what roles each device plays in a complex system.

Price goes on to argue that her account allows us to settle the dispute I described about the contents of R, the frog's internal state. On her view, R represents the presence of flies. It does not represent the presence of small, dark moving things, because the function of the consumer of R is to catch flies, not small, dark moving things. The connection between being small, dark and moving and being nutritious for frogs is too contingent for the proposal that the consumer has the latter function to explain how it contributes to the survival of the organism. Also, the first abstractness condition rules out the possibility that the function of this consumer is to catch flies *by* catching small, dark moving things. R does not represent the presence of healthy frog food, because the second abstractness condition tells us that we should assume that the fellow components of a device are performing their functions successfully, and it is the performance of other systems within frogs that explains why flies are healthy food for them. However, it would be wrong to put too much stress on whether these arguments succeed, partly because we do not know what level of determinacy we should be aiming for in this particular case, and partly because it is only one of a huge number of possible cases.

Instead, Price's account shows that it is possible to be much more precise about which properties and behaviours of devices are their functions. As I have suggested, it focuses mainly on the relationships between the functions of different components of the same system, and it may be the case that further refinements could be achieved by focusing on other issues. For instance, it may be that there are important relationships between functions and natural kinds which are yet to be illuminated, and that this could tell us more about why R represents the presence of flies – if it does. But what is crucial for my purposes is not showing exactly how the teleosemantic account of functions should be modified, but showing that there is reason for optimism about its prospects, and Price's account gives us good cause for optimism.

The second part of my response to the indeterminacy objection is to argue that the objection does not succeed in undermining a robust realism about functions. One might think that the reason the indeterminacy objection arises is that functions are defined in terms of explanation, and that good explanations are a fundamentally pragmatic, discourse-relative matter. The thought would be something like this: that biological devices have the functions they do, and therefore representations in biological systems have the contents they do, in virtue both of the ways that organisms are and their ancestors were, and in virtue of the way that we are, as observers and theorists. This would be because what makes a good explanation depends on who is interested and why. So functions and contents are indeterminate from our point of view, because what makes a good explanation for us is indeterminate, partly because we have various or indeterminate interests when thinking about biological systems.

There is a good reason to be suspicious of this line of thought, because it threatens to show that evolutionary biology can only reveal facts about, at best, how *we can understand* how organisms came to have the traits they do, rather than straightforwardly about how they *did* come to have those traits. Scientists regularly study both what particular traits have been selected for, and what generalisations it is possible to make about patterns of selection. This means that they study functions, because the functions of devices are those properties and behaviours for which they have been selected. For example, a current debate in evolutionary biology concerns the early evolution of feathers, which was a crucial stage in the development of modern birds; it is debated whether the first function of feathers was aerodynamic or to do with display, insulation, or some combination of factors (Zhou 2014). To take another example, it has recently been argued that the phenomenon of 'insular dwarfism' – that is, the tendency of large mammals living on islands to develop smaller forms – is due to a fitness advantage caused by faster reproduction in smaller forms (Raia & Meiri 2006). So these authors suggest that small body size has the function, in this group, of allowing faster growth to reproductive maturity. A current debate in biology of particular philosophical interest concerns the proportion of the human genome which is functional, and again this illustrates that respectable scientists are concerned with functions in Millikan's sense (Graur et al. 2013, Doolittle et al. 2014). So we should certainly not rush to conclude that there are no interest-independent facts about biological functions.

However, it remains plausible that explanations are interest-relative, so it is important that we are able to give a historical definition of functions which does not mention explanation. Given the close relationship between functions and selection-for, we can borrow the following definition of selection-for, from Sober (1984, p. 100):

> *Selection-for*: There is selection for trait *T* in a population if and only if having trait *T* causes organisms to have increased reproductive success in that population.

Following Sober, we can define direct proper functions in this way:

> *Direct Proper Functions*: A direct proper function of a device is a property or behaviour of that device which causes the wider systems in which it is embedded to have increased reproductive success.

This definition replaces the claim that functions explain success with the claim that they cause success, but apart from removing the appearance of interest-dependence, it does not obviously change the account very much. In particular, something like the indeterminacy problems can still be raised, although now they appear in a somewhat different light. Just as both co-occurring with the presence of flies and co-occurring with the presence of small, dark moving things are candidate explanations of the contribution of R, so they are also both candidate means by which R may have caused increased reproductive success. Instead of a challenge about the plurality of explanation, the indeterminacy problems now appear as a challenge about distinguishing causation from correlation, and about correctly attributing causal contributions to the components of complex systems. These are certainly hard philosophical problems, but they are not distinctive to the present case, and it would be a radical response to them to argue that there are no objective facts about causation in cases of the present kind.

Despite these points, it remains striking that thinking about explanation can be so productive in seeking to get a grip on functions. But this need not lead us to the conclusion that claims about determinate functions are significantly more dubious than other causal claims, because it is true in general that thinking about explanation is (at least) a useful heuristic for identifying causes. For instance, consider again the outcome devaluation experiments described in chapter 2; when we aim to identify the cognitive system that causes this pattern of behaviour, it makes sense for us to look for the features of the world that best explain it. Explanatory virtues are widely, and it seems rightly, taken to be a good guide to causation. So facts about biological functions do not seem to be contingent on our interests, and we can therefore conclude that, as it stands, the indeterminacy objection fails to defeat the teleosemantic project.

7.3 Liberality and Explanation

After the indeterminacy objection, the next-most famous objection to teleosemantics is probably the Swampman objection (Davidson 1987). Before discussing Swampman, however, it will be convenient for us to consider a more recent objection, put forward by Tyler Burge (2010), Michael Rescorla (2013) and Peter Schulte (2015). Burge, Rescorla and Schulte are primarily concerned with teleosemantic claims about representational status; they argue that the version of teleosemantics described in section 7.1 is too liberal, alleging that many of the simpler representations that teleosemantics identifies are not representations at all. Although my commitment is to a theory of direction of fit, not of representational status, this line of objection does pose a threat to my view. The problem is that my theory of direction of fit is designed to apply to a particular class of representations which are (in some respects) among the simplest to satisfy the teleosemantic account of representational status. So if Burge, Rescorla and Schulte are right, it may be that none of the entities to which my theory of direction of fit applies are representations at all. To avoid the consequence that desires are not representations, they would have to also reject my account of desire. Alternatively, if the objection succeeds, it may be that many or most of the entities to which my theory applies are not representations, but a few are. That would also cast doubt on my theory, since it would imply that it is only narrowly applicable.

The argument that these critics of standard teleosemantics present is simple: they claim that there is no explanatory advantage to using representational terms to describe the class of entities which teleosemantics implies are representations, but which they claim are not. Schulte (2015) gives the example of the hormone vasopressin. Vasopressin is produced by the hypothalamus in response to excessive blood osmolarity and transported to the kidneys, where it causes an increase in the amount of water reabsorbed into the blood. Since vasopressin apparently has a function, a producer and a consumer, and it causes its consumer to do something that is beneficial only if a further condition holds, vasopressin counts as a representation on the version of teleosemantics described in section 7.1. But, Schulte claims, describing vasopressin as a representation, or talking about its content, truth or falsity, or direction of fit, can add no explanatory value to the kind of explanation just given; he writes that thinking of vasopressin as a representation does not help us to understand either its effects, or why it is present. Since vasopressin has no obvious claim to be a representation other than that it satisfies the teleosemantic criteria, it makes a good test case. If it is possible to identify ways in which using representation-talk to describe vasopressin is valuable, then the objection fails, but if not, then it will apparently succeed.

In my view, the objection fails. There are two ways in which describing vasopressin as a representation can contribute to good explanations. First, Papineau (1993) and Shea (2007) argue that talk of representations is explanatorily valuable because we can explain how systems succeed by saying that they acted on true representations. Papineau focuses on cases involving belief and desire, so he has explanations of this kind in mind: *Eric succeeded in getting a beer by going to the fridge, because he truly believed that there was beer in the fridge*. But we can construct a similar explanation involving vasopressin: *On this occasion, the behaviour of the kidneys succeeded in maintaining a healthy concentration of solute particles in the blood plasma, because the vasopressin signal accurately represented the osmolarity of the blood*. Both authors emphasise the distinction between explanations of this kind, which explain outcomes that happen as a result of behaviours, and explanations of behaviours themselves. They suggest that explanations of the latter sort do not rely on talk of representations, since the internal states concerned could be functionally characterised, as tending to cause particular behaviours when tokened in particular combinations, and equally adequate explanations of the behaviours would be available. In contrast to this, explanations of outcomes in terms of internal states require attention to the co-ordination of those internal states with external conditions. So for example, to explain why the kidneys increased water reabsorption, we only need to mention that vasopressin was released, and to know that vasopressin release tends to have this effect. But to explain why the action of the kidneys led to a good outcome, we also need to know that the vasopressin was released under the right circumstances, and one way to capture this is to describe vasopressin as truly representing that those were the circumstances.

One might think that this use of representation-talk is still unnecessary; we can explain the success of the kidneys' action just by saying that on this occasion, the blood osmolarity was too high. However, while it is correct that in order to explain the fact that a given process produced a particular outcome on a particular occasion, it is sometimes crucial that an element of that process co-occurred with some external condition, that is only one aspect of the form of explanation that I am describing here. Explanations of success that make use of representation-talk also have other significant implications. In general the co-occurrence of some event with an external condition, which leads to a beneficial outcome, might take place by accident. But in contrast, an explanation that describes an entity as a true representation also implies that the entity was *supposed* to co-occur with the external condition that it is said to represent, and therefore says something about how the process in question works. For example, saying that the release of vasopressin truly represented that the osmolarity of the blood was too high goes beyond just saying that the vasopressin release occurred at a time when this was the case. So these explanations both identify the outcomes they explain as having been produced by mechanisms of a certain kind, and show that they are successes, rather than just beneficial outcomes.

We explain successful outcomes, then, by indicating the functions of crucial parts of the processes that lead to these outcomes, and saying that they have been successfully performed. One variety of this form of explanation cites the employment of true representations; this specific variety is valuable because it identifies the successful outcome as having been produced by a process involving elements with the function of co-occurrence with external conditions, which influence the behaviour of consumers.[20]

Second, describing vasopressin as a representation can help explain what it is for, and how it contributes to the operation of a wider system – although this does

[20] One potential problem with this response to Burge, Rescorla and Schulte is that on the standard teleosemantic account of representation, explanations of this kind are arguably rather shallow (Godfrey-Smith 1996, Shea 2007). The problem arises because according to that account, what it is for a representation R to have content C is for the behaviours that past instances of R caused in their consumers to have typically succeeded when state of affairs C obtained. This means that when we say that a behaviour succeeded as a result of a true representation that C, what we are saying amounts to just that it succeeded as a result of things being as they were on past occasions when behaviours of the same type were successful. As Shea puts it, the explanation is of the 'dormitive virtue' type – it's like explaining why a sleeping tablet works by citing its dormitive virtue. However, as Shea explains, this problem can be ameliorated by an adjustment to the teleosemantic account that makes little difference to which entities are counted as representations.

depend on how much the person being given the explanation knows in advance. Still, to someone who had heard that there is a hormone called ‘vasopressin’ but knew nothing else about it, it would be very informative to be told that it is a signal which represents high blood osmolarity, and causes an appropriate response by the kidneys. To someone who had no idea at all what vasopressin was, it would be very informative to be told just that vasopressin is a certain sort of chemical signal found in the human body – much more so than being told just that it is a chemical in the body.

These explanations work primarily by identifying vasopressin as a member of a very general functional category, which is the category of things that work by co-occurring with conditions that are relevant to the behaviours of their consumers. So it is worth emphasising that this is a significant category, by showing that there are some things with functions, that have the function of causing co-operating devices to undergo certain changes, but which are not members of this category – that is, which are not representations according to standard teleosemantics. For example, contractions of the biceps cause the angle between the forearm bones and the humerus to become more acute, and these contractions arguably only perform their function when they do this at the right times – when they are caused to do so by firing in the musculocutaneous nerve. But contractions of the biceps are not intuitively representations, and teleosemantics gets this result, because the bones of the forearm do not count as consumers. This is because the forearm bones do not change what they do when they are acted on by biceps contractions; the definition of a consumer is of something that has the function of changing its behaviour under a certain condition, and the forearm bones do not satisfy this definition. They are merely acted on by biceps contractions. In contrast, the kidneys do have the function of changing their behaviour when vasopressin is released. So the second way in which representation-talk is useful in the case of vasopressin is that it identifies vasopressin as member of a significant functional category, and thereby helps to explain what vasopressin is, and the nature of the contribution that it makes to the system of which it is a part.

My response to the liberality objection offered by Burge, Rescorla and Schulte is therefore that even in the simplest cases where teleosemantics identifies entities as representations, there is something to be gained from describing them as such.

Representation-talk can help explain why successful behaviours succeed, and what representations are for.

7.4 Swampman

We can now turn to the third objection to teleosemantics, which was proposed by Donald Davidson (1987), although anticipated by both Millikan (1984) and Papineau (1984). The objection asks us to consider Swampman, a perfect physical replica of Davidson, formed by chance in a swamp, who emerges and takes Davidson's place in society. Since neither Swampman nor any of his parts have any history of selection, teleosemanticists are apparently committed to claiming that nothing going on in his 'brain', nor any of the sounds he produces or marks he makes, have any representational content. Swampman presents a challenge to teleosemantics not just because it is intuitive that he has beliefs, desires and the rest, but because any account of Davidson's behaviour that used representational notions would *seem* to work just as well for Swampman. If such explanations really would work just as well in Swampman's case, then there may be something that Swampman has in common with the rest of us that accounts for this, and we might think that *this* feature – whatever it is – is more fundamental to representation than teleology. Since my account of direction of fit relies on the teleosemantic notion of function this objection presents a real challenge to my view; if swampman has internal states with directions of fit, then direction of fit may reduce to something other than historically-derived functional properties.

Among the premises that the Swampman objection relies on are the claims that Swampman's internal states lack functions, and the claim that explanations employing representational notions work just as well for Swampman as for his physical duplicate Davidson. On closer observation, however, both of these premises are questionable, so two possible responses to the objection are available; I will discuss them in turn.

It is clear that when he first emerges, Swampman's internal states lack the historical properties that are necessary for functions on the present account. However, Paul Griffiths (2009) has argued that the very possibility of biology shows that there must be biological functions that supervene only on organisms' present, non-historical properties. If Griffiths' argument succeeds, it may show that

Swampman’s internal states do have functions, and hence that a version of teleosemantics could be formulated that would avoid the objection.

Griffiths’ argument has two parts. First, he accepts Millikan’s (2002) view that in order to investigate how biological systems work, why they are the way they are, and how they interact with one another, a notion of function is necessary that goes beyond the brute causal dispositions of systems and their parts (sometimes called ‘Cummins functions’; Cummins 1975). This is because without such a rich notion of function, it does not seem to be possible to distinguish parts of organisms from parts of the environment, parts of organisms from one another, or normal from pathological processes. To take one of Griffiths’ examples, there would be no reason to privilege how and why kangaroos eat grass as a subject of biological study over how and why they are consumed by bushfires. In order to draw these distinctions, he claims, we need to take an evolutionary perspective – to consider the contributions that different processes make to organisms’ survival and reproduction. Second, he claims that among the subjects of study for which such a perspective is necessary is the study of selective histories; that is, the very subject that standard teleosemantics takes to be necessary to reveal biological functions. Specifically, he suggests that if facts about functions could only be known by studying history, then we could not study any given period in the history of selection, because to do so we would first have to know about prior history. So Griffiths claims that there is an irresolvable tension between two of Millikan’s ideas, because if she is right that functions are necessary to identify the objects and processes which are biology’s subject, then she cannot also be right that it is necessary to do historical biology in order to learn about functions.

In Griffiths’ view, then, the success of biology as a discipline shows that it must be possible to identify biological functions just by studying how things are the present. He therefore proposes a *forward-looking* account of biological functions, according to which the functions of components and behaviours are the things they do that currently promote future survival and reproduction. Given that Swampman is a physical duplicate of Davidson, then, Griffiths’ view seems to imply that Swampman’s internal states do have biological functions (although the Swampman objection is not the target of Griffiths’ work).

Unfortunately, this is not a wholly satisfactory response to the objection. I suspect that many readers, like me, will feel uneasy at the suggestion that having come into existence entirely by chance, Swampman's organs and behaviours and internal states could have functions. This unease is vindicated by another aspect of Griffiths' view, which means we have no need to assess his argument for the importance of forward-looking functions in biology. Griffiths also claims that our ability to make sense of biological phenomena is contingent on our understanding that organisms are the products of, and continue to be subject to, evolution by natural selection. It is because kangaroos have been shaped by these forces that we are right to pay greater attention to their dispositions to eat grass than to their dispositions to be burnt in fires, even though we can recognise the difference between these dispositions without knowing anything specific about kangaroo history. So Griffiths' argument may show only that the components and behaviours of systems with the right kinds of histories have functions, even though what those functions are is determined in a forward-looking way. Swampman is an unusual case, because he and his possible descendents will behave in the future as though they were ordinary humans, but it would still seem to be a mistake to treat him as a normal object of biological study – in particular, we cannot explain why he has a 'heart' by saying how this object will contribute to his survival and reproduction. At least pending further argument, we should conclude that Swampman's organs and internal states do not have functions, even though they are very much like Davidson's, which do.

The second premise needed to sustain the Swampman objection is the claim that explanations that employ representation-talk work just as well for Swampman as they do for Davidson. The thought behind this premise is something like the following: assuming that Swampman has beliefs and desires, and that the sounds he produces are words, would help us to predict and understand his behaviour just as effectively as the same assumptions would for Davidson. But this thought is questionable, because while we can be confident that Swampman's behaviour could be *predicted* just as effectively as Davidson's, it seems to beg the question to suggest that we could *understand* it equally effectively. This is because false claims with true consequences do not constitute good explanations of those consequences. If I explain why a saucepan is hot by saying that it is burning, I have not given a good explanation, not because burning things are not generally hot, but because the saucepan is not burning; its heat has a different cause. Similarly, if we explain why Swampman goes to the fridge by saying that he wants a beer, this will be a good explanation only if Swampman actually does want a beer. This line of thought is relevant even if the only way of determining which things are representations is to appeal to explanatory considerations, because predictive power may not be the only factor that makes representational explanations valuable. So if there are explanatory advantages to refraining from attributing content to Swampman's internal states, then contrary to appearances, it may be that explanations that employ representation-talk do not work as well for Swampman as for Davidson. This suggests a bullet-biting strategy for responding to the objection – accepting that Swampman lacks beliefs and desires – which has been adopted by teleosemantic theorists including Millikan (1996) and Neander (1996).

Millikan argues that, roughly, there is no possible theory of representation that does not refer to history, and that there is no 'real kind' that unites human psychological states with Swampman's psychological states. If this is correct, then using representational terms to describe Swampman would be a poor form of explanation, because it would categorise together entities that are in fact very different. Neander argues along somewhat similar lines, that since teleosemantics identifies an interesting, real kind its advantages outweigh the apparent disadvantage of excluding Swampman. However, it is also possible to go beyond these arguments, and detail some further ways in which attributing beliefs and desires to Swampman fails to offer the same explanatory value that it does in Davidson's case. These points rely on the assumption that Swampman's internal states lack functions, but this is an assumption we can legitimately make, because if it is false then the objection fails anyway.

In the previous section, I argued that among the distinctive advantages of representation-talk are explaining success and failure in action, and explaining what representations are and what they are for. Focusing first on the latter point, it is clear that if Swampman's internal states lack functions then we cannot explain what they are for, never mind why he possesses them, by describing them as representations of any sort. There is no reason except chance why Swampman is the way he is, and no part of him is for anything. But what's more, this is also relevant to explanations of Swampman's behaviour. When we say that Davidson went to the fridge because he wanted a beer, we do not imply just that he had some set of internal states which were disposed to cause this kind of action; instead, we put the action in a much broader context. We imply that some mechanism in Davidson working roughly as it is supposed to caused him to go to the fridge. We imply that his action will succeed if and only if he gets a beer, not just because he has an internal representational state which makes getting beer his goal, but because what states like this are for is setting goals for action. We allude to the fact that Davidson is the kind of thing whose movements are rightly seen as purposive, because of the way he has come to exist. And none of this is true of Swampman. So ultimately the Swampman objection and the liberality objection fail for the same reason: that representation-talk has rich implications, which go beyond causal dispositions and co-occurrence relations. For the liberality objection this matters because it applies even to simple signals, and for the present objection it matters because despite Swampman's apparent sophistication, the objection can only succeed if he lacks functions, so the rich implications are not available in his case.

Two further objections to teleosemantics are similarly well-known, but not relevant here because they focus on content, rather than direction of fit or representational status. Paul Pietroski (1992) describes a group of creatures, the kimu, who develop a mutation that allows them to see red via some internal state K. For some reason, they become attracted to the colour red, and those with the mutation tend to gather at the tops of hills at dawn to see the sunrise. This behaviour happens to protect them from predatory snorfs, so the mutation spreads through the population. According to Pietroski, Millikan's theory implies that K represents the absence of snorfs, but this cannot be right, he claims, since if snorfs and red things occurred together, kimus that encountered this scene would token K and would be drawn to it. Even if it succeeds, this objection suggests at most a shift in emphasis from the consumer's needs to the producer's abilities in the theory of indicative content. Christopher Peacocke (1992) argues that teleosemantics implies that we cannot think about matters that have not, and could not have, had any impact on our evolutionary history. But in fact we do have thoughts about such matters, such as my belief that the pole star is 434 light years from Earth. The issues this claim raises are quite distant from those I am concerned with, since our ability to think about such remote matters seems to be connected to our capacity for conceptual thought, and I have not touched on the role of concepts in representation. If this objection succeeded, it would not imply that teleosemantics is ill-suited to providing a theory of direction of fit, but only that it is not up to the task of analysing the contents of representations produced by sophisticated thinkers.

In this chaper, I have described the three key components of the teleosemantic framework: the theory of functions, the claim that representations mediate between producers and consumers, and the proposal that normal conditions for consumer behaviour can be used to pick out contents. I have argued that these ideas all provide major steps forward in thinking about representation, and I have defended teleosemantics against three objections, each of which has been thought to show that these steps take us in the wrong direction. Next, I will argue for a novel theory of direction of fit within this framework.

Chapter 8: The Discretion View

8.1 The Discretion View and The Canonical View

The new theory of direction of fit which I advocate is called the *Discretion View*, and premise IV of my overall argument is an immediate consequence of this theory. In this chapter I describe the Discretion View, and argue that it is superior to what I will call the *Canonical View*, which is the standard teleosemantic theory of direction of fit. The Discretion View accurately captures a distinction between two ways in which representations can work, which correspond to the directions of fit, while the Canonical View does not. More specifically, there are some representations which the Canonical View identifies as having both directions of fit, but which in fact work in the same way as pure indicatives; so the Canonical View ascribes imperative content too readily. In this section I introduce the Discretion View, and in section 8.2 I specify its scope, by defining the class of *biological representations*. Then in section 8.3 I give my argument for the Discretion View, and in section 8.4 I address two ways in which issues relating to indeterminacy affect the theory.

Before we start, it may be useful to recall the standard teleosemantic theory of direction of fit: representations have imperative content if and only if they have the function of causing their consumers to perform specific behaviours (or bring about specific states of affairs), and have indicative content if and only if they have the function of co-occurring with specific normal conditions for the behaviours they prompt in their consumers. These are the claims that I am now calling the 'Canonical View'.[21]

The Discretion View is inspired by a proposal by Lewis (1969). Lewis's subject is systems of signals that have been set up by explicitly-agreed conventions. His proposal can be illustrated by a simple example:

[21] See footnote 25 for discussion of the place of this theory of direction of fit in the teleosemantic canon.

> Hero and Leander are lovers who live on opposite banks of the Hellespont. They arrange that on nights when Hero is alone, she will light a lamp, and Leander will swim across.[22]

The lamp's being alight is a representation, and we can ask about its direction of fit. Does it say something about Hero's condition – *I'm alone*, say – or about what Leander should do – *come and see me* – or both? Lewis answers that it depends on whether either party was supposed to use their discretion in using the signal, according to the convention by which it was established. There are three possibilities:

> If Hero is to use her discretion about when to light the lamp, but Leander is always to come when he sees it, then it has only imperative content – it says only something like *come and see me*. The thought here is that since Leander knows that Hero is using her discretion, the lamp doesn't tell him anything specific about her circumstances; but since he is not supposed to use his discretion, it does tell him what to do.
>
> If Leander is to use his discretion about what to do when he sees the lamp, but Hero is always to light it when she is alone, then it has only indicative content – it says only *I'm alone*. The lamp's being alight tells Leander something specific about Hero's circumstances, but does not tell him what to do.
>
> If neither party has discretion, then the signal has both directions of fit – it says both *I'm alone* and *come and see me*.

The criterion for indicative content that Lewis is proposing is therefore that the producer lacks discretion about when to produce the signal, and the criterion for imperative content is that the consumer lacks discretion about what to do when they receive the signal. According to Lewis, when both criteria are satisfied, the signal is 'neutral' – it may equally well be described as either imperative or indicative. But I will take it that such signals have both directions of fit.

[22] In the myth of Hero and Leander as it's now told, the lamp was not used for this purpose, but to guide Leander across the channel. I'm changing the story.

To make Lewis's proposal applicable to the case of desire, we need to reconfigure it in teleosemantic terms. The criterion for indicative content is that the producer lacks discretion; we might put this in more teleosemantic terms by saying that there is some specific state of affairs under which the producer is supposed to produce the signal. This is equivalent to the canonical criterion for indicative content, which is that the representation has the function of co-occurring with a specific normal condition for the behaviour it prompts in its consumer. There are two apparent differences; one is that the discretion criterion concerns the function of the producer, while the canonical criterion (as I have presented it) concerns the function of the representation itself, and the other is that the canonical criterion cites normal conditions. But neither of these is more than apparent. First, a representation will have the function of occurring under some condition if and only if its producer has the function of producing it under that condition. And second, the point about normal conditions does not substantially change the criterion, because it can only be the function of the producer to produce the signal under a certain condition if its doing so would be mentioned in the best explanation of how it works, and that will only happen if the signal co-occurs with the condition on typical occasions on which it contributes to the success of the system. So these conditions have to be relevant to explaining successful behaviour in any case, which is to say that they have to be normal conditions.[23] It follows that the criterion for indicative content suggested by Lewis's proposal is the same as the canonical one.

Because the Discretion View adopts this criterion, the dispute between the two views only concerns the nature of imperative content.[24] Lewis's criterion for imperatives is that their consumers must lack discretion. In teleosemantic terms, that comes to the claim that the consumers of imperatives must have the function of behaving in specific ways, whenever they are adapted by those signals. (Recall that in Millikan's terminology, a consumer is said to be adapted by a representation when the representation occurs, and this occurrence cuts down the range of things

[23] Both Millikan and Papineau insist on describing indicative content as fixed by normal conditions, and this might seem to be a unnecessary complication in light of the argument of this paragraph. Their approach is explained by their view that the function of a device must be one of its *effects*, and that co-occurrence cannot be an effect (e.g. Millikan 1990, Papineau 1998). I don't share this view about functions, but the normal conditions approach in any case has the advantage of making it more explicit that indicatives represent success conditions.

[24] Price (2001) also criticises the canonical criterion for imperative content, but retains the indicative one.

that the consumer might do in accordance with its function.) In contrast, the canonical criterion is that a representation has imperative content if it has the function of causing some specific behaviour in its consumer. This is weaker than the Discretion View's criterion, as we will shortly see. First, however, I will give statements of the various criteria for the directions of fit, for ease of reference:

> *Indicative Criterion (both views)*: A representation has indicative content if and only if it has the function of co-ocurring with some specific normal condition for the behaviours that it causes its consumer to perform, in accordance with the consumer's function.

> *Imperative Criterion (Canonical View)*: A representation has imperative content if and only if it has the function of causing some specific behaviour on the part of its consumer.

> *Imperative Criterion (Discretion View)*: A representation has imperative content if and only if its consumer has the function of behaving in some specific way whenever it is adapted by the representation.

One point of clarification in order is that any given behaviour can be accurately described in many ways – just as an intentional action can be described as moving one's arms, working a pump, or drawing water – but the behaviours caused by an imperative only need to be of one type under some description, as long as that type would be mentioned in the best explanation of how the representation contributes to the wider system. For example, suppose that a simple organism has a number of different feeding strategies, but pursues feeding in general when and only when a certain internal state occurs. This internal state might well have imperative content on both views, even though the organism would perform different behaviours when it was tokened, because they would all be feeding behaviours. In stating the Discretion View's criterion for imperative content, I have used the expression 'behaving in some specific way' (rather than 'performing some specific behaviour'), because my intention is that a representation can have imperative content even if its consumer only has the function of entering into some *mode* of behaviour when adapted by that representation (see section 8.4).

A typical case in which the Discretion View and the Canonical View have different consequences is the following. Consider a neural system that causes a simple organism either to feed, to drink, or to seek warmth, according to a fixed-priority rule. Imagine that this system is the consumer of three signals – a food signal, a water signal, and a warmth signal – produced by other systems, and that when it receives more than one of these signals at the same time, it always gives feeding priority over both drinking and seeking warmth, and drinking priority over warmth-seeking. We can further assume that all three signals have indicative content; they say *food is needed now*, *water is needed now*, and *warmth is needed now* respectively. This example distinguishes the two views, because all three signals have imperative content on the Canonical View, but only the food signal does on the Discretion View. On the Canonical View, what it takes for a signal to have imperative content is that it has the function of causing the consumer to bring about a specific outcome. This is true of all three signals, because each of the three typically contributes to the organism's survival and reproduction when and only when it causes the consumer system to do something specific: to get food, water and warmth respectively. On the Discretion View, though, the water and warmth signals lack imperative content, because when they occur the consumer system does not always have the function of behaving in one specific way. Instead, how it should behave according to its function depends on other factors, such as whether the food signal is also occurrent. So in the sense I have defined, the consumer has discretion about what to do when adapted by the water and warmth signals. The Discretion View thus implies that these signals do not tell the consumer what to do – they merely inform it about the organism's needs.

In general, then, the kinds of representations that distinguish the Canonical and Discretion Views are ones that sometimes fail to perform their functions, even though their consumers function perfectly, and the functions in question are to cause the consumers to behave in particular ways. To get representations of this sort, a fairly complex recipe of features is required: the consumers must be capable of being adapted by more than one input at a time (not necessarily all representations); the representations must have the function of causing the consumer to behave in specific ways; yet the consumer's function must require it to behave in those ways, when adapted by the relevant representations, only under a proper subset of the possible ways in which they may be adapted by other events at the same time. Clearly, though, the example shows that such cases are possible.

A noteworthy feature of the example is that the Canonical View arguably handles it better than the Discretion View. On the Canonical View, all three signals are bi-directional, whereas on the Discretion View, the feeding signal is bi-directional and the other two are pure indicatives. This is a mildly awkward result, but my argument is not based on the ability of the Discretion View to get neat or intuitive results in specific toy cases. Instead, I will argue that the Discretion View accurately picks out a deep distinction between two ways in which representations can work.

Finally, to get a more complete picture of the relationship between the Discretion View and the Canonical View, we need to see whether there are any representations that have imperative content on the Discretion View, but not on the Canonical View. Combined with the example just given, the answer to this question will tell us whether the discretion criterion (as I shall call it) is strictly more demanding than the canonical criterion, or just different. The answer is that there are no such representations. For consider a representation with a consumer that has the function of behaving in a specific way, whenever the representation adapts it. Then there is some specific form of behaviour by its consumer which is caused by the representation, and which would be described as such in the best explanation of how the system works; so it is a function of the representation to cause this behaviour. So the discretion criterion is strictly more demanding, in the following sense: every representation that has imperative content on the Discretion View also has imperative content on the Canonical View, but not every representation that has imperative content on the Canonical View also does so on the Discretion View. Given that every representation of the kind we are concerned with has either imperative content, indicative content, or both, what this means is that there are some representations that are pure indicatives on the Discretion View, but bi-directional on the Canonical View. The two views agree about which representations are pure imperatives, because this is determined by their shared criterion for indicative content.[25]

[25] On first glance, Millikan's (1984) canonical statement of teleosemantics may appear to endorse the Discretion View, not what I am calling the Canonical View. Millikan writes that, 'In the case of imperative intentional icons, it is a proper function of the interpreter device, as adapted by the icon, to produce something onto which the icon will map in accordance with a specific mapping function…' (p. 99). However, in a later account of her view, Millikan (1995, p. 189) writes that, 'A representation

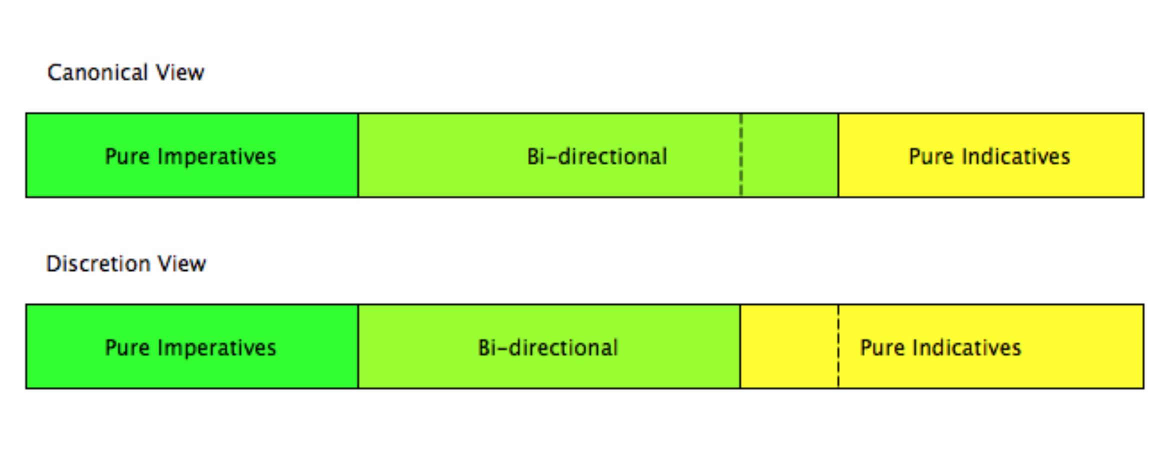

Fig. 3. The Canonical View and the Discretion View. Because the Discretion View has a stricter criterion for imperative content, more representations are pure indicatives, rather than having both directions of fit.

8.2 Biological Representations

The Discretion View is a *partial* theory of direction of fit, in the sense that it does not aim to say what characterises the directions of fit in all cases; instead, it is concerned only with a specific class of representations. In my view, the most promising way to develop a good overall theory of direction of fit (and of representation more generally) is to start by developing such partial theories, and then try to put them together. This is appropriate because the range of properties we can appeal to when theorising about representations in human languages, for instance, is very different from the range we can appeal to when theorising about representations in many other biological systems, and the challenges that theories will face in the two areas may also be very different. For example, in the case of language we might – possibly – appeal to grammatical properties or speakers' intentions, and face challenges arising from the variety of human languages and

is directive [i.e. imperative] when it has a proper function to guide the mechanisms that consume it to produce its satisfaction condition,' and immediately goes on to describe desires as directive (/imperative), a result which follows only on the Canonical View (see also 1984, p. 140). This apparent discrepancy can be resolved by noting an ambiguity in the technical term 'adapted'. Millikan's original definition of this term does not clearly distinguish between two possible meanings, respectively entailing that: (i) a representation adapts its consumer whenever it occurs; and (ii) a representation adapts its consumer whenever it will be causally responsible for the consumer's behaviour, provided the consumer acts according to its function. I am using 'adapts' with meaning (i), and Millikan's statement of her criterion for imperative content is equivalent to the canonical criterion, providing that she is using 'adapts' with meaning (ii). It is also a foundational tenet of Papineau's teleosemantic theory that desires have imperative content (1993, 1998), and he writes that teleosemantics *in general* identifies imperative contents with 'the conditions that [representations] are biologically supposed to produce' (1998, p. 1).

cultures. In other biological cases we might appeal to functions derived from natural selection, and face challenges from indeterminacy claims and swampman-like thought experiments.

This approach might raise concerns either that there is no one phenomenon of direction of fit – since different theories of direction of fit can be correct in different areas – or, conversely, that partial theories of direction of fit will not generally be correct, but merely useful stepping stones towards the correct universal theories. But neither of these results need follow. It is quite possible that there is a single phenomenon of direction of fit, and that partial theories could correctly characterise parts of it. These theories would be co-extensive with parts of the true universal theory of direction of fit, but likely expressed in different terms. For the sake of comparison, consider what would be involved in giving a theory of what it is to be responsible for some outcome, in the sense of being an apt object of the reactive attitudes as a result of that outcome (Strawson 1962). Since both individuals and institutions can be responsible for outcomes, one would wish to give a theory that covered both. But it would be reasonable to give distinct theories of individual and corporate responsibility first, then try to work out what they have in common, even on the assumption that responsibility in this sense is a single phenomenon.

The cases to which the Discretion View is intended to apply are *biological representations*. Biological representations are those to which teleosemantic theories in general are most naturally suited. They have clearly-identifiable producers and consumers, with specific functions, which co-operate with one another to promote the success of some wider system. Their functions and those of their producers and consumers come from biological selective processes, rather than from the explicit intentions of designers or from analogous processes within human cultures; however, these functions may be derived rather than direct, in the senses described in section 7.1. Because they have consumers – which by definition are systems with functions that involve behaving differently, depending on whether they occur – biological representations are all at least candidates for having directions of fit, rather than merely standing for objects, properties or states of affairs. Biological representations include both many representations that occur within organisms, and many that are used for communication between organisms.

One might think that there would be many borderline cases of biological representations, since many representations are the products of interactions between biological and cultural factors. For instance, one might ask whether the desire for one's country to become a democracy is a biological representation or not, since some aspects of its function – those it shares with desires in general – seem to come from biological processes, whereas others – those that give it its distinctive content – seem to be as culturally-determined as those of any representation. But here I am defining a technical notion of biological representation for a specific purpose, and I stipulate that it is the aspects of the functions of representations, producers and consumers that are relevant to determining direction of fit that must be derived from biological processes. Typically, these will be more general aspects of the functions of producers and consumers, rather than those which are concerned with specific representations.

A relatively clear example of a class of representations which are not biological in the present sense is sentences in human natural languages. This is not because natural language is not a biological phenomenon, but because the functions of producers and consumers with respect to sentences of natural languages are not fixed by biology to a sufficient extent to ground theorising about types of natural-language representation. The theory developed here is not suitable for understanding direction of fit in natural language sentences.

It would count in favour of the Discretion View if similar theories of direction of fit for other kinds of representation were also successful, but I will not attempt to develop such theories here. In particular, there are significant obstacles facing the discretion approach even in the context of conventional signals like those discussed by Lewis; I describe some of these obstacles at the end of section 8.4. The subject of the next section is biological representations, but for convenience's sake I will often drop the 'biological'.

8.3 An Argument for the Discretion View

My argument for the discretion view consists of three steps. First I will argue that all biological representations (and many others) work by co-occurring with states of affairs that are relevant to the behaviour of their consumers. Then I will argue that there are two kinds of such states of affairs, which correspond to genuinely different ways in which representations can work, which in turn correspond to the directions of fit. Finally, I will show that the Discretion View accurately captures this taxonomy of ways in which representations can work, but the Canonical View does not. So as I have said, my argument will work by showing that the Discretion View accurately captures a deep distinction between two ways that representations can work.

First Step: All representations work by co-occurrence with relevant states of affairs

It is uncontroversial that indicative representations work by co-occurring with relevant states of affairs. What it is for a representation to have indicative content, on the account agreed by all parties, is for it to have the function of co-occurring with normal conditions for the behaviours it causes its consumer to perform, and normal conditions are by definition relevant states of affairs. This idea is also attractive from outside the perspective of teleosemantic theory: indicatives are true when they co-occur with the states of affairs which are their contents, and false otherwise, and they are useful insofar as they convey information about these states of affairs to other systems that may profitably use it to modify their behaviour.

At least some imperatives also seem to have this feature. Consider a simple, conventional signal which is intuitively a pure imperative: a bugle call which tells the soldiers in a camp to muster on the parade ground. This bugle call might be blown whenever the relevant officer thinks the soldiers need to be brought together, either in specific situations in which it is agreed convention that they should muster (like at dawn every morning to begin their day), or in quite unexpected ones, that nonetheless justify this action (like alien invasions). But it remains the case that the signal will only be useful if it sufficiently reliably co-occurs with states of affairs of a specific kind: those which have the property of making it appropriate for the soldiers to muster in the parade ground, given the unit's projects and aims. It will also only be useful if the ability of the soldiers to independently recognise such circumstances does not match the ability of the responsible officer to do so; if the soldiers could all tell when they should muster by themselves, then the signal would be otiose. So some representations that appear to be pure imperatives work by co-occurring with states of affairs that have the following two properties: they are relevant to determining what the consumer should do, and in some sense the producer has better access to them than the consumer does.

The kind of division of labour that makes the imperative useful in this case can also occur in biological cases. For instance, an organism might have two complex sub-systems in its cognitive architecture; one for determining what sort of action to perform, and another for executing actions. The first system would need to be good at identifying situations in which specific action-types are appropriate, while the second system would need to be good at generating actions of these types. It could be that the only signals sent between these systems were imperatives; the first system telling the second what to do. But these signals would be useful because of their correlation with the different sorts of situations identified by the first system. They would work by making the behaviour of the second system sensitive to states of affairs that were relevant to determining what it should do, which only the first system was suited to detecting.

More generally, imperatives are typically thought of as representations with satisfaction conditions that concern the behaviour of their consumers, subsequent to receiving the signal. They call for specific responses from their consumers, and thus contribute to the success of wider systems by causing their consumers to behave in specific, valuable ways. But the point I wish to call attention to is that this cannot be done at random, unless the behaviours would be equally valuable whenever they were produced; and in this unusual situation, the contribution of the randomly-produced representations to the successful functioning of the system would be minimal. So in general, an important part of the way that even pure imperatives contribute to the success of wider systems is by co-occurring with states of affairs that are relevant to the behaviour of their consumers: specifically, by occurring when matters are such that specific behaviours by the consumers are appropriate.

An argument on similar lines follows from the definition of a consumer. What it is for a device to be a consumer of a representation is for the device to have a relational function with respect to that representation, and what this means is that what the consumer has the function of doing depends on whether or not the representation is occurrent at the time. From this, it follows that the behaviours that representations cause in their consumers must be appropriate under some circumstances, and not under others – because otherwise there would be no reason for the behaviour of the consumer to be conditional on anything else; the representations would be of no value. Given that this is the case, and that the consumers have the function of behaving differently in response to the representations, it must be that case that the representations co-occur sufficiently reliably with the states of affairs that make the consumers' behaviours appropriate that the consumers' functions are adaptive. And furthermore, this co-occurrence must be a large part of what the representations contribute to the system, because the consumers rely on it when they adapt their behaviour to the representations, and the producers' role is to generate the representations at the right time.

The value of both indicatives and imperatives therefore depends crucially on their co-occurring with states of affairs that are relevant to the behaviour of their consumers. Co-occurring sufficiently reliably with these states of affairs is a necessary condition for its being adaptive for representations to be produced, and for consumers to adapt their behaviour to them. Even if there are some representations that make other substantive contributions to the successful operation of the systems of which they form parts, it is in virtue of this co-occurrence property that they are useful *qua* representations.[26]

Second Step: Ways in which representations can work

So far I have argued that in some sense, the representations I am concerned with all work in the same way: by co-occurring with relevant states of affairs. There is a *prima facie* tension between this idea and the idea that representations can have either of two directions of fit, because it is natural to think that the two directions of fit are two different ways in which representations can work. This tension is resolved, however, when we notice that there are two different kinds of relevant states of affairs, and hence two different ways in which representations can work. The two kinds of relevant states of affairs are: (i) normal conditions for consumer

[26] A possible objection to this claim is that some representations are useful primarily because they are isomorphic to what is represented. As it stands this suggestion is ambiguous: it could mean that there is an isomorphism between meaningful parts of the representation and parts of the thing represented (as on a map), or it could mean that there is an isomorphism between the members of a set of possible repreesentations and a corresponding set of possible contents (as in Millikan's example of the honeybee's waggle dance). Either way, I suspect that such isomorphisms are useful because they make it possible for representations to be interpreted by their consumers – that is, for consumers and representations to be mutually configured so that representations will affect consumers in adaptive ways. But this is necessary in every case, and even the simplest signals (like beaver tail-splashes) are trivially isomorphic to what they represent. On the other hand, it is also no use for a representation to be isomorphic with something if it occurs at the wrong time. So while it is correct that in some cases signals are more remarkable for occurring at the right time (like beaver tail-splashes), and in others they are more remarkable for being interpretable (like maps), it is far from clear that isomorphism is an *alternative* function to co-occurrence. For more on this see Shea (2013).

behaviours; and (ii) states of affairs of the form *X's being appropriate*, where X is some behaviour the consumer can perform. I will first explain how these two kinds of states of affairs relate to each other, then how they ground genuinely different ways in which representations can work, and finally how these correspond to the directions of fit.

As I will use the term, the state of affairs of a particular behaviour's being *appropriate* is the state of affairs of that behaviour's being the correct one for the consumer to perform, *given how it and co-operating systems work*. This means that appropriateness is a property that only one behaviour can have on any given occasion. It does not mean that the appropriate behaviour is always the most biologically beneficial one that the consumer could perform, because there could be a more beneficial behaviour available in some circumstances which it would not be adaptive for the organism to attempt to discriminate – because either the processing costs or the risks of failure would be too great. It also does not mean that it is appropriate for a system to behave in a given way when and only when it is the function of that system to behave in that way, because in general functions depend on what co-operating systems are doing, rather than how things are in the environment, and co-operating systems may fail to perform their functions successfully. Most pertinently, if the producer of a representation that works by co-occurring with the state of affairs of some behaviour's being appropriate produces this representation at the wrong time, then it will be the function of the consumer to perform the behaviour concerned, but it will not be appropriate for this behaviour to be performed.

Given this definition, the state of affairs of a behaviour's being appropriate cannot be a normal condition for that behaviour, unless this type of state of affairs can also be characterised in another way. Normal conditions are, by definition, states of affairs that have obtained on typical past occasions on which behaviours succeeded, which explain that success. But it can never explain the success of a behaviour *B* to say that the state of affairs of *things being such that B is appropriate* obtained at the time, because it is close to trivial that if the behaviour succeeded then it was appropriate in the circumstances. Normal conditions are *substantive* success conditions for consumer behaviours.

However, many representations do work by co-occurring with states of affairs which are both normal conditions, and make particular behaviours appropriate, because they form parts of systems that use relatively simple techniques for working out what to do. For example, the beaver's tail-slap works by co-occurring (not on every occasion, but sufficiently reliably) with the presence of danger, which is a substantive success condition for the behaviour of making for the lodge. But as we imagine the case, the presence of danger is also sufficient to make heading for the lodge appropriate; beavers don't, we imagine, take anything else into account in working out whether to do this. As I will explain shortly, it is in virtue of this point that there are biological representations with both directions of fit.

The alternatives to this kind of case are for representations to work by co-occurring with (substantive) normal conditions, which are not sufficient to make specific behaviours appropriate; or for them to work by occurring when specific behaviours are appropriate, but for the systems that produce them to determine this in sufficiently complex ways that they should not be described as co-occurring with normal conditions.[27,28] In the former case, what it is the function of the consumer to do must depend on other inputs, so the consumer must be relatively (but only relatively) sophisticated. In this kind of case we can think of producer systems as *serving* the consumers of representations, because they provide information that the consumers need, but do not thereby control what the consumer does. This way of thinking becomes more attractive as consumer systems become increasingly sophisticated relative to producers. However, representations that work in this way are common. Even the fixed-priority system that I described in section 8.1 uses representations that work in this way – for instance, the 'water' representation which could be trumped by the 'food' signal.

In the latter case, what the consumer must do does not depend on other inputs, so at least to this extent it should not be sophisticated (although it may use sophisticated methods to execute the appropriate behaviour). On the other hand, the producer of the signal must have a fair degree of sophistication, since it must be

[27] Is it also possible for representations to work by co-occurring with states of affairs which are neither normal conditions, nor make specific behaviours appropriate? If there could be such representations, this would imply that my account of direction of fit is too restrictive, since it wrongly leaves some representations with neither direction of fit. I argue that this is not possible, clarifying my view in the process, in section 8.4.

[28] Any biological system that determines whether a specific behaviour is appropriate will do so by determining whether some substantive condition holds; but it is not always the case that this condition should be given in stating the function of the representation or its producer, because it may be that to do so would involve explaining how the producer works, and the function of a system must be distinct from means by which the system performs that function. This is roughly the condition on functions that Price (1998) calls *Abstractness 1*. For more see sections 7.2 and 8.4.

capable of working out what behaviour on the part of the consumer would be appropriate in a way that goes beyond just establishing whether some normal condition holds. In this kind of case we can think of producer systems as *controlling* the systems that consume their representations, although again the attractiveness of this way of thinking depends on the degree of difference in sophistication between the systems.

We can see, then, that representations can work either by occurring when specific behaviours are appropriate, or when specific conditions relevant to behaviour obtain, or (usually in simpler systems) by doing both of these things simultaneously. We only get this clean taxonomy of functions if we insist on a notion of appropriateness which means that only one behaviour can be appropriate at a time. Otherwise there would be some representations which we would describe as working in the first way, but which actually determined behaviour only in combination with other inputs, like those that work by co-occurring with normal conditions; while others would themselves be sufficient to determine consumer behaviour.

It is hard to resist using the language of direction of fit to describe representations of these three kinds: those that work by occurring when specific behaviours are appropriate tell their consumers what to do; those that work by occurring with substantive normal conditions tell them how things are; and those the work in both ways, do both. Since co-occurrence with normal conditions is the commonly-agreed criterion for indicative content, there is no reason not to think of representations that work in that way as indicatives. These representations have clear accuracy conditions. Also, it makes sense to think of representations that work by co-occurring with states of affairs of the form *X's being appropriate* as imperatives. They certainly tell their consumers what to do, and we can also specify how these representations work by stating that they are imperatives, and giving their imperative content – that is, specifying behaviours that their consumers can perform, or states of affairs that their consumers can bring about, that have a special connection to the representations. Saying that they are imperatives implies that they work by co-occurring with some behaviour's being appropriate, and giving their content says which behaviour this is. Representations that work this way have satisfaction conditions, because they call for specific behaviours to be performed. It is also an advantage of associating this distinction with direction of fit that it leaves room for bi-directional representations.

Third Step: The two views and the ways that representations can work

As I have noted, this taxonomy of ways in which representations can work fits neatly with the shared criterion for indicative content. So all that remains to be shown is that it also fits neatly with the Discretion View's criterion for imperative content. After that, I will conclude my argument by briefly explaining what is wrong with the Canonical View.

First, suppose that a representation satisfies the discretion criterion, so its consumer has the function of behaving in a specific way whenever it is adapted by this representation. This means that when the representation occurs, the consumer will do its part in promoting the success of the wider system by performing this specific behaviour. So the role of the representation must be to occur at times at which the consumer's behaviour is supposed to occur, given the way that the co-operating producer and consumer systems work. In other words, the representation itself must work by co-occurring with the state of affairs of the behaviour's being appropriate. Now suppose that some other representation works in this way; if this really is the nature of its contribution to the system, it must be that this representation alone is sufficient to prompt the consumer to perform a specific behaviour. So it follows that the consumer lacks discretion with respect to this representation, and hence that the representation satisfies the discretion criterion. Representations work in the way constitutive of the imperative direction of fit if and only if they satisfy the Discretion View's criterion.

To see the contrast with the Canonical View, we can consider its criterion for imperative content. On that criterion, it is sufficient for a representation to have imperative content that it has the function of causing its consumer to behave in a specific way. Crucially, this criterion is weaker than the discretion criterion, because it only requires that the representation causes the consumer to behave in a specific way *on those occasions on which the representation itself causes the successful behaviour*, whereas the discretion view's criterion requires that the representation causes the consumer to behave in a specific way *on every occasion on which the representation occurs, and the consumer performs its function*. Consequently, on the Canonical View there can be representations which have imperative content even though on most occasions on which they occur it would be extremely costly for their consumers to behave in the way they say, because other much better alternatives are available. This is possible because on those occasions, provided things go well, other representations will be causally responsible for the consumer's behaviour, overriding the influence of the so-called imperatives.

The more general objection that the present argument generates to the Canonical View, however, is that those representations that satisfy the canonical criterion but not the discretion criterion work in the same basic way as pure indicatives. They contribute by co-occurring with states of affairs that are relevant to how their consumers should behave, but do not determine this – do not suffice to make specific behaviours appropriate. So the Canonical View attributes different directions of fit to some representations that should be classified together. As well as this, for the same reason, the Canonical View attributes both directions of fit to some representations that do suffice to make specific behaviours appropriate, and also to others that do not – it classifies as belonging to the same group two sets of representations that in fact work differently. The Discretion View offers a superior account of direction of fit, because unlike the Canonical View it accurately distinguishes fundamentally different ways in which representations can work.

This completes my argument for the Discretion View; in the next section, I elaborate further on the theory, in the course of defending it against two related objections.

8.4 Challenges Relating to Indeterminacy

The indeterminacy objection to teleosemantics is the claim that biological functions are insufficiently determinate to ground representational properties, because many equally good explanations are possible of how biological systems work. It constitutes a challenge to teleosemantic theories of direction of fit, just as much as to theories of content. Although I have already given my main response to this objection in section 7.2, and that response was sufficiently general to apply to issues concerning direction of fit, part of how teleosemantic theorists should respond to this objection is by clarifying and adding detail to their theories. So I will discuss the indeterminacy objection in this section, partly for the sake of further developing the Discretion View. I will also discuss a further, closely related challenge, which suggests that because the Discretion View's criterion for imperative content is relatively demanding, it wrongly implies that some representations lack either direction of fit. Finally, I will outline how similar issues arise for Lewis' theory of direction of fit in conventional signals, and explain the significance of this point for my own theory.

To see how the indeterminacy objection applies to the Discretion View, we can consider the simple system for selecting actions that I described in section 8.1. That toy system consists of a producer which is capable of detecting whether the organism is in need of food, water or warmth; a consumer that is capable of generating actions suitable for satisfying these needs; and three possible representations that mediate between them. The consumer works according to a fixed-priority rule: its function is always to seek food if it receives the feeding signal, and to seek water and warmth only when the signals for these needs occur, but the feeding signal does not. As I described things then, the Discretion View implies that the water and warmth signals only have indicative content, because what the consumer should do when it is adapted by these representations depends on other inputs. In contrast, the Canonical View implies that these representations have both directions of fit, because they contribute to the organism's success only when they cause specific behaviours.

Incidentally, we can now see why the Discretion View does better than its rival in analysing this case: it is because there is an important difference between the feeding signal, which determines the behaviour of its consumer, and the water and warmth signals, which do not. When the producer generates the water and warmth signals there is only an extremely limited sense, if any, in which it is working out what the consumer should do, as opposed to detecting substantive states of affairs which are relevant to the consumer's behaviour. But when it generates the feeding signal, it is working out whether the organism should seek food. That said, the contrast is less stark here than it would be in cases involving more sophisticated systems.

The indeterminacy objection arises here because there are other possible ways to describe this system. Most pertinently, it might be objected that there *is* a specific behaviour that the consumer has the function of performing whenever it is adapted by the water signal: the behaviour of *seeking water, unless the feeding signal is also received*. Unless there is some principled way of deciding whether the consumer has this conditional function, it will be indeterminate whether the representation concerned satisfies the discretion criterion, and hence indeterminate whether it has the imperative direction of fit. Although the present example does not allow it to be demonstrated so easily, the very same issue also affects the Canonical View. Also, the problem generalises. For any representation that appears to lack imperative content, it will be possible in principle to give a conditional or disjunctive description of a specific way that its consumer should behave, whenever it is adapted by this representation. And equally, for any representation that appears to lack indicative content, it will be possible in principle to give a conditional or disjunctive description of a specific, substantive condition under which its producer has the function of producing it. Why doesn't this mean either that direction of fit is indeterminate on the present account, or that every representation has both directions of fit?

In the case of the fixed-priority action-selection system, I suspect that it is in fact indeterminate whether the water and warmth signals have imperative content. It may be that the distinctions between the different ways in which representations work begin to break down in the simplest cases. In this case, the Discretion View implies that whether the water signal has imperative content is a matter of whether the consumer's function is best described by a look-up table, one line of which reads *when adapted by water signal, seek water unless food signal is also occurrent*, or by the combination of a general principle: *always perform the behaviour with the highest priority, of those associated with occurrent signals*, and a list of behaviours and signals in order of priority. If the former is correct, the consumer will lack discretion, but if the latter is correct, it will have discretion, in the sense that the Discretion View defines. Which of these is the consumer's function depends on which more accurately describes the property of the consumer that has caused it to be reproduced, by contributing to the success of the organism of which it is a part. But in this particular case we seem to have two descriptions of the very same property, so there may be no fact of the matter whether the discretion criterion is satisfied. This signal may be 'neutral', in Lewis's sense. Still, because this is a very simple case, this indeterminacy does not pose a threat to the Discretion View. To test (and perhaps refine) the Discretion View, we need to see whether the indeterminacy persists more generally.

To this end, we can consider a new example. Suppose that the entire human action-selection system produces representations as outputs, which are consumed by a further system which is responsible for directly controlling the muscles to produce the required movements. The representations would, we can imagine, stipulate that motor programmes should be performed – for instance, that the consumer should cause the body to pick up a pen, throw a ball, or walk downstairs. At first pass, both the Discretion View and the Canonical View imply that most of these representations will be pure imperatives. They lack indicative content, because there is no specific normal condition under which they have the function of occurring. The challenge facing the teleosemantic approach to direction of fit is that this may be indeterminate; that there may be a good sense in which these representations do work by co-occurring with specific normal conditions. So we can consider whether the *walk-downstairs representation* (as we might call it) has the function of occurring together with any specific normal condition.

To try to construct such a condition, let us suppose further that the human action-selection system consists solely of sub-systems for habitual and goal-directed control, and a further sub-system that converts their outputs into a single instruction for the motor-control system. In this case, the walk-downstairs representation should only be produced when the current situation is either (i) of a type that has been associated in the agent's experience with walking downstairs leading to reward, or (ii) of a type that the agent believes implies that walking downstairs will contribute to the satisfaction of his or her desires more than other actions currently available, or both. So it might conceivably be suggested that it is the function of the action-selection system to produce the walk-downstairs representation when this complex condition holds. This would mean that the walk-downstairs representation would have indicative content, because this appears to be a substantive normal condition for walking downstairs. However, there are at least two problems with this proposal.

First, in this particular case the proposal cannot be correct, because it involves conflating the function of the action-selection system with the functions of its sub-systems. If the action-selection system tests whether the agent has a habit conditioned on current stimuli, and examines what actions would currently best promote the agent's desires, what is there left for the habitual and goal-directed systems to do? This is part of the point of Price's *Abstractness 1* condition: that the functions of systems are to do things, rather than to do them by specific means, because this allows us to identify accurately the contributions of different parts of complex systems (see section 7.2). In this case, the function of the system is to produce the walk-downstairs representation when walking downstairs is appropriate, rather than to test for its appropriateness of this action in a specific way. The problem with the proposal is particularly clear in this case because of the way that the present system combines the results of two sub-systems, but the point also holds more generally.

Second, although the proposed indicative content for the representation is both distinct from the state of affairs of the behaviour's being appropriate, and capable of explaining the success of this behaviour, closer scrutiny shows that it is not a normal condition for walking downstairs. To see this, it is helpful to compare the present example to the case of the beaver's tail-slap, which is genuinely bi-directional. Let's say that the proposed normal condition in the present case is *that the environment is such that the agent's desires and habits imply that she should walk downstairs*; this is supposed to be the normal condition for walking downstairs. In the tail-slap case, *danger's being present* is the normal condition for the beavers' action of making for the lodge. These two conditions do not stand in the same relationship to the two behaviours. The normal condition in the tail-slap case is the success condition for the subsequent behaviour that the producer aims to detect, whereas the putative normal condition in the action-selection case is a more proximal condition, that is relevant to the means by which the producer detects the state of affairs that really matters. It is analogous to *there being rustling in the trees*, if we imagine that this is how beavers detect predators. This means that it could much more easily happen that the 'normal condition' is satisfied in the action-selection case, but the agent does not benefit from walking downstairs, than that there could actually be danger about in the forest, and the beavers not benefit from going to the lodge.

For these two reasons, the overall function of the human action-selection system is simply to work out which action would currently be most beneficial for the agent. Its function with respect to the walk-downstairs representation is to produce it when walking downstairs would be most beneficial. This pattern is typical of general-purpose mental systems that produce and consume a range of representations: they have general functions, of which their functions with respect to specific representations are instances. These considerations also show, crucially, that the teleosemantic approach to direction of fit does imply that there are representations with determinate directions of fit, including pure imperatives. The existence of pure indicatives follows from similar considerations about how to understand consumer functions.

We can now turn to the second challenge to the Discretion View. As I have often noted, the Discretion View's criterion for imperative content is significantly more demanding than the Canonical View's criterion. A natural way to argue against the Discretion View is therefore to argue that this criterion is too demanding. Also, it is not obvious that there couldn't be biological representations with both producers and consumers that had discretion – that is, that were both produced and consumed in flexible ways, sensitive to a variety of factors. If there were such representations, the Discretion View would imply that they lacked direction of fit, perhaps showing that the Discretion View's criteria were too stringent. So I will now consider a case that appears to be of this form.

Imagine that blue tits have a dedicated internal system for determining when and whether they should engage in nest-building, which produces a signal that influences the behaviour of a general-purpose action-selection system. When adapted by this signal, the action-selection system has the function of causing the bird to gather and arrange nesting material at all times during daylight hours, except when a particularly urgent need or fine opportunity arises to do something else – for instance, to feed or to escape from a predator. And imagine further that the producer of the system does not follow a simple rule to determine whether the time is right to build a nest – instead, it takes into account the weather, day-length, presence of potential mates, and availability of food and nesting resources in the area in a complex way. Given the considerations I just presented regarding the human action-selection system, it looks as though the producer in this case will have the function only of producing the signal when conditions are suitable for nest-building, rather than when some substantive normal condition for nest-building holds, so the signal will lack indicative content. But at the same time, what the consumer should do when adapted by the signal depends on other inputs – it should cause nest-building behaviour only at times when feeding, for instance, is not a better option.

How things work out in this kind of case depends primarily on the function of the consumer; specifically, on how the consumer uses the signal. First, it could be that the consumer performs more than one kind of behaviour that is typically adaptive only under the conditions with which the signal has the function of co-occurring, in which case these would amount to a normal condition. In our example, it could be that the tit's action-selection system both causes nest-building, and a particular style of faster, riskier feeding behaviour when it receives the signal. In this case, the signal would have indicative content, even though produced by a relatively sophisticated process. Second, it could be that the consumer enters a particular mode of behaviour when it is adapted by the signal, in which case it would have imperative content. In our example, it may well be appropriate to describe the tit's action-selection system in this way; and also it's possible that the alternative feeding style would be part of this mode of behaviour, in which case the signal might have both directions of fit. In that case, its content would be something like *conditions are good for nest-building; go into nest-building mode*.

The key question now is whether these possibilities are exhaustive, and not just in the blue tit case but more generally. It seems that there is one further possibility: that the consumer would have only one behaviour available to it to which the signal is relevant, but that it should have discretion about whether to perform this behaviour. For instance, it might be that the only difference the nest-building signal makes to the blue tit's behaviour is that it causes it to engage in nest-building when no other action promises much reward, whereas without the signal it would never do this. In this case, however, the representation will have the function of co-occurring with a normal condition for the behaviour it causes in the consumer, because it can't be the case that the condition with which it co-occurs is sufficient for that behaviour to be appropriate – if it were, the consumer would not have discretion. In the case of the blue tit, the signal would have indicative content like *conditions are generally suitable for nest-building*, which *is* a normal condition for nest-building at specific moments, if nest-building at specific moments is also adaptive only when further conditions are also satisfied. In general, where the consumer has discretion, the representation will always have the function of co-occurring with a genuine normal condition.

This completes my defense of the Discretion View. One further obvious objection to the Discretion View remains, which is that it implies that desires are pure indicatives, and the remaining two chapters will each be concerned, in part, with versions of this objection. I complete this chapter by briefly returning to Lewis' theory of direction of fit in conventional signals.

If Lewis' theory was correct, that would lend support to the teleosemantic Discretion View, since we should expect the correct theories of direction of fit in different domains to be broadly similar. However, I am agnostic about Lewis' theory, for reasons analogous to some of the points already discussed in this section. We can return to the example of Hero and Leander.

To make things concrete, suppose that the explicit convention the lovers agreed gave neither of them discretion: they agreed that Hero would light the lamp whenever she was alone for the night, and Leander would swim across whenever he saw it. Now suppose that they both followed this convention, and as a result Leander was killed attempting to swim the Hellespont in a storm. Even if this was they had said they would do, Hero should not have wanted or expected Leander to set out in a storm, and they might both have been glad if, on the night of the storm, either Hero had chosen not to light the lamp, or Leander had chosen to ignore it. So a question arises about how the theory should deal with the possibility that the convention should conflict, to the parties' knowledge, with their shared interests. One option for answering this is of course to say that the signalling system could be thought of as inadequate or poorly-chosen. Another is to say that in this case one or other of the lovers really did have discretion, according to the convention, even though this went unsaid when it was agreed, so the signal only had one or other direction of fit; more generally, the approach would be to allow tacit clauses in conventions. A third is to say that in this case either the imperative content, the indicative content, or both, were conditional or conjunctive (*come and see me if you can* or *I'm alone and the strait is safe*, say), but that again this depended on tacit clauses. And yet a fourth would be to argue that all human conventions include tacit clauses to the effect that they should only be followed if no major unexpected obstacles arise, and that these should be distinguished from other forms of discretion and are not relevant to determining content and direction of fit. These points certainly show that the theory needs to be developed in more detail, especially because it is plausible that human conventions often involve tacit clauses. But it is particularly noteworthy that it will be possible to describe these tacit clauses either as attributing discretion to one or other party, or as implying that the parties have fixed tasks, which are more complex than is explicitly stated; so the theory risks indeterminacy between these possibilities, and hence indeterminacy about direction of fit.

In addition to this, Lewis' theory also needs to be further developed to deal with cases in which both parties seem to have discretion. Lewis (1969, p. 146) chooses not to engage in analysis of cases of this kind. So I am unable to claim support for my Discretion View from this source.

In this chapter, I have presented and argued for the Discretion View, a theory of direction of fit for biological representations. According to this view, representations have imperative content when their consumers lack discretion, and indicative content when their producers lack discretion. So the Discretion View entails my premise IV, that biological representations with consumers that have discretion have only the mind-to-world direction of fit. My argument in favour of the Discretion View is that it accurately distinguishes two ways in which representations can work – either by occurring when some specific behaviour on the part of the consumer is appropriate, or by occurring when some substantive condition, relevant to the behaviour of the consumer, holds. The Discretion View makes the difference between saying how things are and saying what to do a real difference in kind.

Chapter 9: The Direction of Fit of Desire

9.1 Desires are Pure Indicatives

In this chapter, I combine the results of part I, concerning the nature of desire, with those of the last three chapters, on direction of fit, to bring my argument to a conclusion. In this section, I argue that desires are pure indicatives. Then in the following section, I discuss the content of desires; that is, what they say about how things are. The principal problem here is to get a clearer grip on the notion of reward. In the third section, I recap the argument of parts I and II.

In part I, I defended an account of desire which yielded the following three premises:

> I. Desires are outcome values.
> II. The goal-directed control system works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.
> III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.

These three premises outline the key features of desires which are relevant to their direction of fit. In addition to this, in the chapter just gone I argued for the Discretion View, which is the conjunction of the following two criteria for direction of fit in biological representations:

> *Indicative Criterion*: A representation has indicative content if and only if it has the function of co-ocurring with some specific normal condition for the behaviours that it causes its consumer to perform, in accordance with the consumer's function.

> *Imperative Criterion*: A representation has imperative content if and only if its consumer has the function of behaving in some specific way whenever it is adapted by the representation.

The imperative criterion entails my fourth premise:

> IV. Biological representations with consumers that have discretion have only the mind-to-world direction of fit.

So to complete my argument it remains only to show that desires are biological representations with consumers that have discretion – that is, that they fail to satisfy the imperative criterion. For the sake of completeness, though, it will also be worthwhile to show explicitly that desires do satisfy the indicative criterion. In this section I will first explain why desires count as biological representations, then show that they satisfy the indicative criterion, then finally show that they fail to satisfy the imperative criterion.

However, things are a little more complicated than that might make them sound, because desires come in two forms: standing and occurrent. So before we can address the issue of their directions of fit, we need to establish what the relationship is between these two kinds of desire, considered as representations. For example, consider my desire to visit Yellowstone National Park in the United States. Is my standing desire to do this the same representation as the occurrent desire(s) to do so that I sometimes experience?

The answer to this question is 'no', because standing and occurrent desires are produced and consumed by different systems, exist for different periods of time, and contribute in different ways to adaptive action-selection. The role of standing desires is to store information about the levels of reward provided by outcome-types over relatively long periods, and they are produced and modified by reward signals, according to the process discussed in sections 3.4 and 3.5. It is interesting to note that the producer of standing desires may not be localised within the brain, or even an integrated, dedicated system. It may be that standing desires are formed and updated just in virtue of the way that the OFC responds to receiving perceptual signals and reward signals. If the same parts of the OFC are involved in maintaining standing desires and generating occurrent ones, and if the perceptual signals and reward signals are also used for other purposes (such as updating habits), then all of the elements of the producer of standing desires also have other functions. However, this is not a problem for teleosemantic analysis, since this system, such as it is, does have a distinct function with respect to the production of standing desires.

Standing desires are consumed in the production of occurrent desires. Occurrent desires are produced by a mechanism which causes activity in the OFC and ventral striatum, which represents the reward value of currently salient outcomes. The level of this activity is influenced by the strength of the relevant standing desires; by the level of salience of the outcomes; and also by factors such as the organism's occurrent basic drives and the degree to which the outcome concerned is currently represented as associated with other desired or aversive outcomes. It appears, then, that occurrent desires do not represent the average reward values of types of outcomes, but instead the levels of reward that outcomes are expected to provide on specific occasions. This is another reason to see them as distinct representations. The primary consumer of occurrent desires, meanwhile, is the goal-directed action-selection system. This system uses occurrent desires and instrumental beliefs to calculate predicted reward values for salient possible actions. It does not control action directly, but instead produces many outputs representing the values of actions, which are used together with the outputs of the habit system and perhaps other signals in the ultimate determination of behaviour (see ch. 4, especially section 4.1). Desires are also consumed in the process of generating reward signals, which in turn are used to update desires and habits, but it is not clear whether standing or occurrent desires are used for this purpose.

There is also no particular reason to assume that standing and occurrent desires will have the same direction of fit, so I will consider them separately. However, the discussion of the last two paragraphs does show that desires are biological representations. Desires have producers and consumers, which have functions determined by biological selective processes; and in order to establish their directions of fit we need only look to biologically-determined aspects of their functions. For instance, although understanding the object of my desire to visit Yellowstone may require us to consider the influence of language and culture on my mind, understanding the direction of fit of this mental state is a purely biological matter. So we can now turn to the main task of this section, which is to use the Discretion View to establish the direction of fit of desire.

According to the Discretion View, representations have indicative content if and only if they have to function of co-occurring with normal conditions for the behaviours they cause in their consumers. Occurrent desires satisfy this criterion because the goal-directed system normally performs its function, which is to accurately predict the level of reward that salient actions will provide, when the strengths of the occurrent desires that it consumes correspond accurately to the levels of reward available from the outcomes which are their objects. For example, suppose I was trying to decide whether to take a trip to Montana. In that case, booking a flight to Montana would be a possible action for me, that was salient at the time, and the function of my goal-directed system with respect to this action would be to calculate how much reward I would get by taking it. This value would be calculated by multiplying the strengths of my occurrent desires by the credences that I held that these desires would be satisfied if I took the flight, and adding up the results. If my desire to visit Yellowstone was occurrent at the time (which it should be, since it would be rational for me to associate going to Montana with visiting Yellowstone), then this desire would affect the output of the goal-directed system, and normally (in the teleosemantic sense) this output would only be correct if the amount of reward I would actually get from going to Yellowstone corresponded to the strength of this desire. Things could work out well otherwise, but only if the error in my desire was, by good fortune, compensated for elsewhere – for instance, if my desire to go to Yellowstone was a little too strong, this could be compensated for if I underestimated the chances that I would get to Yellowstone by exactly the same degree.

There are two crucial points to notice here, both of which relate to my talk of ‘accuracy’. First, the outputs of the goal-directed system are capable of being accurate or inaccurate only because they are themselves representations with indicative content. These representations are consumed by what we might call the *ultimate action-selection system*, which directly controls human behaviour. The ultimate action-selection system normally causes adaptive behaviour only when the signals it receives from the goal-directed system and others co-occur with the right states of affairs of the form *things being such that action A will lead to x units of* *reward*. If the outputs of the goal-directed system do not correspond with how things are in the world, then it can only be by good fortune if the ultimate action-selection system, and therefore the organism, behaves in the most adaptive way available at the time. This is where the accuracy conditions for the outputs of the goal-directed system come from, which in turn determine the accuracy conditions for occurrent desires.

Second, these accuracy conditions also depend on the existence of a systematic *code*, which relates physical properties of the various signals involved to states of affairs 'in the world'. Consider again my occurrent desire to visit Yellowstone, and suppose that it involves a specific level of neural activity y (y here is taken to be the value of some physiological parameter). If this occurrent desire is to represent some outcome's having a particular reward value z for me, then there must be something about me that makes it the case that y stands for z, rather than some other level of reward, or something else entirely. What makes this the case is that through my evolution and development, my brain has come to employ a particular code when using levels of neural activity to represent reward values; it is the fact the my occurrent desire is a signal in this code that makes it the case that it represents my visiting Yellowstone as having reward value z. The outputs of the goal-directed system must also use a similar, but distinct code. The case is comparable to that of the honeybee's waggle dance, which is widely discussed in the literature on teleosemantics; in that case, even novel dances have determinate content, because they represent the location of nectar via a code that has been established by the operation of selective processes on dances of the same kind (Shea 2013).

It is worth noting too that while occurrent desires have the function of co-occurring with normal conditions, these are not normal conditions of the most typical kind. Returning again to our example, consider the output that my goal-directed system generates when it is adapted by my occurrent desire to go to Yellowstone; this output is a signal representing the predicted reward value of booking a flight to Montana. This behaviour on the part of the goal-directed system could normally contribute to adaptive behaviour even if I entirely lacked the desire to go to Yellowstone, and even if going to Yellowstone would be in no way rewarding for me, because booking a flight to Montana could be rewarding for me to the same degree for quite different reasons. So this case is somewhat more complicated than the case of the beaver's tail-splash, for instance, in which there is one straightforward normal condition for the behaviour the representation causes. However, we can resolve this issue by re-describing the behaviour of the goal-directed system. Although there may be no fixed normal condition for representing the action of booking a flight as having a particular reward value, there is a fixed normal condition for *producing this representation in response to an occurrent desire to go to Yellowstone, of a given strength*. The normal condition is that going to Yellowstone will be rewarding to the corresponding degree. This normal condition arises from a general description of how the goal-directed system works; because the way it works in general relies on the accuracy of occurrent desires, a normal condition for its operation on particular occasions is always that outcomes actually will be as rewarding as the occurrent desires say they will. This issue cannot be taken to be an objection to the claim that occurrent desires have indicative content, because the same points are also true of beliefs.

Occurrent desires, then, have indicative content because part of how they contribute to the overall action-selection process is by co-occurring with states of affairs of roughly the form *outcome O has reward value x*.

Standing desires also have indicative content, because they too work by co-occurring with states of affairs of roughly this form, although they make their contribution at a different stage of the process. For example, consider my standing desire to visit Yellowstone. The consumer of this standing desire is the system that generates occurrent desires, and its behaviour will be affected by the standing desire only when visiting Yellowstone is a salient outcome. When that outcome is salient, this system will produce an occurrent desire to visit Yellowstone, with a strength determined by the strength of the standing desire, together with a range of other factors. Normally, this occurrent desire will only be of the correct strength when the standing desire too is of the correct strength; it would take good fortune for this to happen otherwise. So in my example, visiting Yellowstone's being of a certain reward value for me is a normal condition for the production of an occurrent desire, of a certain strength, for that outcome. The existence of this convoluted system, in which one representation is used in the production of another with almost the same content, is explained by the value of having both standing and occurrent desires, detailed in section 3.2.

Like occurrent desires, standing desires have indicative content as elements in a sequence of representations that eventually leads to behaviour, and it is important that token standing desires belong to a type that represents using a fixed code. In addition to these points, it is also important that both standing and occurrent desires are produced by systems which are sensitive (to some extent) to the states of affairs that seem to be normal conditions for the behaviours they prompt in their consumers. This claim is closely related to one part of premise III, stated above and defended in part I, and I will discuss it in more detail in the next section. The reason for putting it off until then is that it requires a more precise account than any I have given so far of the nature of reward.

We can now turn to the question of whether desires have imperative content on the Discretion View. This will be the case if and only if their consumers have the function of behaving in specific ways, whenever they are adapted by specific desires. As I have argued in chapter 8, this would also mean that any given desire would have the function of co-occurring with a state of affairs that made a specific behaviour on the part of its consumer appropriate.

Occurrent desires lack imperative content, because how the goal-directed system should behave when adapted by a given occurrent desire also depends on the other inputs that it receives at the same time. Whether or not a particular representation of the value of an action should be produced, when a given occurrent desire adapts the system, depends on whether that action is salient at the time; what instrumental beliefs are present concerning that action; and on the strengths and objects of other occurrent desires, since the action may affect the chances of many outcomes. This is true even in cases in which the actions and outcomes concerned are extremely tightly linked, or even identical – which is possible since there is no reason why the performance of an action should not itself be a desired outcome. Even in this case, the agent will typically have other occurrent desires, and beliefs about how the action concerned will affect the chances that these desires will be satisfied, either of which could vary and affect the output.

Standing desires also lack imperative content, because the system that consumes them is the producer of occurrent desires, and we know that at any one time the goal-directed system should only be adapted by a small number of occurrent desires, relative to the number of standing desires that are stored. This is a significant part of the point of occurrent desires, as I argue in section 3.2. So whether a given standing desire should cause the production of a corresponding occurrent desire depends on whether its object is salient at the time – the standing desire exists, and adapts the consumer, whether or not this is the case, and continually over long periods of time. Also, the strength of the occurrent desire does not depend solely on the strength of the standing desire. So it is not the case that the consumer of standing desires has the function of behaving in a specific way, whenever it is adapted by a given standing desire.

These conclusions are striking, and it is clear that they would not have been reached without the Discretion View's strict criterion for imperative content. However, I have argued for that criterion in detail in chapter 8. So instead we should turn to a different issue, which is the extent to which the claims that standing and occurrent desires lack imperative content are contingent on the details of the account of desire developed in part I. Are my conclusions robust in this sense? I will discuss one change to my account which would not affect these conclusions, and another more radical one which would.

First, in section 4.1 I consider two different ways in which the goal-directed and habitual systems might interact. On one model, which I tentatively defend in that section, the two systems each calculate reward values for a range of possible actions, and these values are amalgamated in determining how we act. I have so far assumed this model in the present chapter. On the alternative model, though, there is a further system which is responsible for delegating control of behaviour at any given time to one or other of the two systems. On this model, the goal-directed system would produce only a single output, which would be a signal capable of driving action. So this model is closer to how desires are usually thought of; as inputs to a system the outputs of which are actions themselves. However, even if the alternative model is correct, occurrent desires still lack imperative content (and the situation concerning standing desires is unaffected). This is because when a given occurrent desire adapts the goal-directed system, on the alternative model, which action the goal-directed system has the function of causing depends on the other occurrent desires and instrumental beliefs present at the time. In personal-level terms, what we as agents should do when we experience desires always depends on what other desires we have at the time, and on our beliefs about the causal structure of the situation. So if the Discretion View is correct, the conclusion that desires lack imperative content is robust across a range of possible accounts of desire.

However, one limit of this robustness is that my conclusion does rely on the truth of the claim in premise III that more than one desire can be an input to the goal-directed system at any one time. If it were the case that only one occurrent desire could adapt the goal-directed system at a time, then depending on some other details, it may be the case that this representation would have imperative content – would tell the goal-directed system to bring about the desired outcome in the best available way. My argument for this claim – which I take to be highly plausible in any case – is given in section 4.3.

This completes my argument for premise V, which was the claim that if premises I-III are correct, then desires are biological representations with consumers that have discretion. So I have now presented arguments for all of the following five premises:

I. Desires are outcome values.

II. The goal-directed control system works by promoting the performance of the action that has the greatest expected reward value, based on outcome values and representations of action-outcome relationships.

III. Outcome values are inputs to the goal-directed control system, which are produced and modified by a system which is to some extent responsive to evidence for the reward values of outcomes, and it is normal for more than one outcome value to act as an input to the goal-directed control system at any one time.

IV. Biological representations with consumers that have discretion have only the mind-to-world direction of fit.

V. It follows from I-III that desires are biological representations with consumers that have discretion.

From these premises, it follows straightforwardly that desires have only the mind-to-world direction of fit.

An apparent implication of this result is that desires are capable of being true or false. I accept this implication, although I admit that it sounds strange to describe desires in this way. Several factors may contribute to this strangeness, but one which is worth emphasising is an important difference between beliefs and desires. This is that the content of beliefs as representations is the same as what is believed, but the content of desires as representations is not the same as what is desired. We refer to desires by describing the outcomes that we desire, in sentences such as ‘I want to visit Yellowstone’. I do not deny that what is desired here is the outcome that the agent visits Yellowstone; I simply claim that this is distinct from the content of the desire, which is something like *My visiting Yellowstone has reward value x*, for some value *x*. This means that truth and falsity for desires, unlike for beliefs, is not a matter of the truth or falsity of what is desired. For this reason, I distinguish between the *content* of desires, which is the way that they represent things as being, and the *objects* of desire, which are the outcomes that desires concern - the things which we desire. Desires have both truth-conditions – given by their content – and satisfaction-conditions – given by their objects – so we have no need to radically change the way in which we talk about desires.[29]

9.2 The Nature of Reward and the Content of Desire

In this section, I discuss desires’ representational content. According to the orthodox view, desires are pure imperatives with propositions describing outcomes as their contents. For example, on the orthodox view the desire to eat ice-cream is a pure imperative with the content: *I am eating ice-cream*. On my view, in contrast, this desire is a pure indicative with the content: *My eating ice-cream has reward value x*, for some value *x* (abstracting away here from differences between occurrent and standing desires). My focus in this section will be on the nature of reward. Specifically, I will try to show that there is a property of rewardingness which outcomes possess, and which desires might track. For my purposes it is particularly important to show that there is a way of understanding the rewardingness of outcomes for agents other than as the degree to which those outcomes are desired; if this was the only plausible way to understand reward, my view about the direction of fit of desire would be very seriously undermined.

[29] The fact that when we talk about desires we usually only mention their objects might be thought to be evidence against my view, and in favour of the view that desires have only imperative content. In particular, we explain behaviour by describing the objects of agents’ beliefs and desires. But this line of thought is not compelling, because to give a full explanation of someone’s behaviour in belief-desire terms it is necessary to describe the strengths of their desires, as well as the objects of those desires. So if we can infer what desires represent from what we need to say about them to give belief-desire explanations, then the strengths of desires must also contribute to their contents – and it is hard to see how this could be possible, if desires were pure imperatives.

My aim in this section is only to show that it is possible to give a plausible and cogent account of reward that satisfies some basic criteria. I will not attempt to give a full defence of the claim that desires represent levels of the property I will describe, because I do not have a theory of content to hand, and because there is presently too much uncertainty about how standing desires are formed and modified for this task to be completed (see section 3.5). I will also leave aside a further very important question about the content of desires, which is what determines their objects – for instance, what it is about certain instances of activity in my OFC that makes them desires *to eat ice-cream*, rather than, say, *to go to the moon*.

Being rewarding is a property that outcomes have relative to individuals at times; and rewardingness is a matter of degree. Actions can also be described as rewarding, depending on their tendency to cause rewarding outcomes, and both types and tokens of both outcomes and actions can be rewarding to determinate degrees. It is normal in psychology and neuroscience to describe outcomes and even objects as *rewards* when they are, or their consumption is, positively rewarding, but this talk of rewards tends to collapse the distinction between outcomes which the individual *takes to be* rewarding, and those which *actually are* rewarding. My approach will maintain this distinction, while also entailing that the rewardingness of an outcome for an agent may be significantly (although not directly) affected by the strength of the agent's desire for that outcome.

We can begin by considering the hypothesis that what it is for an outcome to be rewarding is for it to be biologically beneficial; that is, conducive to the individual's survival and reproduction. This hypothesis is attractive in part because reward is a 'common currency' for comparing actions and outcomes that have advantages and disadvantages of different kinds. Reward signals update desires and habits, and the goal-directed and habitual systems are general-purpose systems for behavioural control. These systems are responsible for much of our behaviour, and this behaviour will be maximally biologically beneficial if our desires and habits accurately represent the level of biological benefit that outcomes and actions will bring us, provided that the systems also work correctly in other ways. So it is tempting to conclude that the normal condition for successful action on a desire of a given strength is that the biological benefit available from its object corresponds to that strength. Assuming that orthodox teleosemantics gives the correct theory of content for representations such as desires, this would mean that reward is biological benefit.

However, there are powerful arguments against this view. Crucially, it looks as though the function of the entire brain is to control the organism's behaviour so as to generate the maximum possible biological benefit. So to attribute the same task to the desire and habit systems seems to confuse the relationship between systems and sub-systems (Price 1998; see section 7.2). For instance, consider the role of standing basic drives: these seem to be features of the brain and body that cause us to direct our actions towards outcomes which are of particular importance for survival and reproduction. But standing basic drives co-operate with the desire and habit systems in the overall process of behavioural control; so it would more accurately describe their relationship to think of drives as dispositions to find beneficial things rewarding, and reward itself as something like drive-satisfaction.

Thinking of reward as biological benefit also makes some features of the way reward signals are generated, and hence of how desires and habits are formed, rather puzzling. Because reward signals are generated when desires are satisfied, as well as drives, it is possible for us to acquire strong desires which have only tenuous connections with either drive-satisfaction, or biological benefit, without any malfunction occurring. Humans in particular commonly have strong desires for outcomes which seem to have very little impact on our ability to survive and reproduce – desires for certain kinds of aesthetic experience, for instance. The point here is not that the contents of a representation must be such that misrepresentation would involve malfunction (Neander 1995) – that would rule out the hypothesis that beaver tail-slaps represent the presence of danger, since the mechanism used to produce tail-slaps presumably does not malfunction when it responds to mere *signs* of danger, and have similar effects in many other similar cases. Instead, the point is that there is room for an account of reward that better explains what the system that produces reward signals is doing.

For these reasons, a more promising approach is to think of reward in the following way: what it is for an outcome to be rewarding for an agent is for it to promote the aggregate satisfaction of that agent's desires and basic drives. This approach has the significant virtue of making sense of how the goal-directed and habitual systems work. In particular, reward signals used in updating desires and habits measure current levels of reward by measuring aggregate levels of desire- and drive-satisfaction. I will first give a more precise, formal presentation of this view, then explore some of its properties.

First, let us define some expressions. Noting that both types of outcomes and token outcomes have reward values, let $R^{type}_{A,t}(o)$ be the function that takes outcome types to the reward values of those outcome types for agent A at time t (using the letter o for outcome types), and let $R^{token}_{A,t}(u)$ be the function that takes outcome tokens to the reward values of those outcome tokens, again for agent A at time t (using u for outcome tokens). Let $d_{A,t}(o)$ be the strength of A's standing desire or standing basic drive for outcome type o at time t, and let α_i be a parameter modifying the function d according to the strength of the agent's occurrent desire or drive for outcome type o_i at the time. Here negative desires and negative drives – that is, learnt and innate aversions – should be included. Then assuming A has standing desires or drives for outcomes $o_1, o_2, \ldots\ o_n$, we can define the reward value of an outcome type for A at time t as follows:

$$R^{type}_{A,t}(o_k) = \sum_{i=1}^{n} (\Pr(o_i|o_k) - \Pr\,(o_i|\neg o_k))d_{A,t}(o_i) - \ d_{A,t}(o_k)$$

That is, the reward value of an outcome type is the degree to which it promotes the satisfaction of all of the agent's other desires and drives, taking into account the conditional probabilities relating the outcome type to other desired outcomes, along with the strengths of the agent's desires. Taking u_k to be a token of outcome type o_k, we can further define the reward value of a token outcome for A at time t in the following way:

$$R^{token}_{A,t}(u_k) = (\sum_{j=1}^{m} \alpha_j d_{A,t}(o_j) - \ \alpha_k d_{A,t}(o_k)) - \sum_{i=1}^{n} \Pr\,(o_i|\neg o_k)\alpha_i d_{A,t}(o_i)$$

In this formula, the outcome types indexed by 'j' are those of which tokens occur simultaneously with or subsequent to u_k, and $\alpha_i d_{A,t}(o_i)$ designates the strength of the agent's occurrent desire or drive for outcome o_i at time t, if they have one, or the strength of their standing desire or drive for that outcome if not. So according to this formula, the reward value of a token outcome is the difference between the degree to which the outcome satisfies the agent's occurrent desires and drives, except for their desire for that very outcome, and the likely degree to which these desires and drives would have been satisfied, if the outcome concerned had not occurred.

To gain an initial grasp of these two formulae, it is useful to think of each as composed of three elements. First, there is a 'positive' element: in the case of the formula for outcome types, this is the sum of the strengths of the agent's standing desires, weighted by the conditional probabilities of their being satisfied, given the outcome concerned; in the case of the formula for outcome tokens, it is the sum of the strengths of the agent's occurrent desires for the outcomes that actually occurred together with or subsequent to the token outcome concerned. These elements give initial expression to the idea that the reward value of an outcome is its tendency to promote the aggregate satisfaction of the agent's desires and drives. This is then refined by two further 'negative' elements. The sum of the agent's desire strengths (standing or occurrent, depending on the formula), weighted by the probabilities of their being satisfied if the outcome concerned does not occur, is subtracted. Intuitively, the reason why this is necessary is that if some outcome makes the satisfaction of an agent's desires and drives less likely, then that outcome has negative reward value. So reward value is a measure of how some outcome *affects* the probabilities of others. The other 'negative' element is the strength of the agent's desire for the outcome itself, which is subtracted because merely desiring an outcome is not in itself sufficient to make that outcome rewarding. This term occurs with weight 1 in the 'positive' element, so the effect of this 'negative' element is just to cancel it out – to ensure that the strength of the agent's desire for the outcome concerned has no direct effect. Intuitively, this goes back to one of the most basic ideas underlying my claims in this thesis, which is that it would be highly maladaptive for us to desire at random.

There are several important points to note about the account of reward set out in these two formulae.

First, an apparent constraint on an adequate theory of reward is that the reward value of an outcome type should be approximately equal to the average reward value of tokens of that type. To see that my account satisfies this constraint, let $u_k^1, u_k^2, \dots u_k^l$ be a large number of tokens of outcome type o_k, and let $a_i^h d_{A,t}(o_i)$ be the strength of A’s occurrent desire for o_i on the occasion of u_k^h. Then the constraint will be satisfied if:

$$lR_{A,t}^{type}(o_k) - \sum_{h=1}^{l} R_{A,t}^{token}\left(u_k^h\right) \approx 0$$

That is:

$$l\left[\sum_{i=1}^{n} (\Pr(o_i|o_k) - \Pr\,(o_i|\neg o_k))d_{A,t}(o_i) - d_{A,t}(o_k)\right] \\ - \left[\sum_{h=1}^{l} \left\{ (\sum_{j=1}^{m} \alpha_j^h d_{A,t}\left(o_j\right) - \alpha_k^h d_{A,t}(o_k)) \\ - \sum_{i=1}^{n} \Pr(o_i|\neg o_k)\, a_i^h d_{A,t}(o_i) \right\}\right] \approx 0$$

Rearranging gives us:

$$l\sum_{i=1}^{n} \Pr(o_i|o_k)\, d_{A,t}(o_i) + \left[l\sum_{i=1}^{n} \Pr(o_i|\neg o_k)\, a_i^h d_{A,t}(o_i) - l\sum_{i=1}^{n} \Pr\,(o_i|\neg o_k) d_{A,t}(o_i)\right] \\ + \left[\sum_{h=1}^{l} \alpha_k^h d_{A,t}(o_k) - l d_{A,t}(o_k)\right] - \sum_{h=1}^{l}\sum_{j=1}^{m} \alpha_j^h d_{A,t}\left(o_j\right) \approx 0$$

Of these, the two terms in square brackets are both approximately zero, assuming that the variations in occurrent desire strength between the various occurrences of the outcome type cancel each other out. This entails that:

$$l\sum_{i=1}^{n} \Pr(o_i|o_k)\, d_{A,t}(o_i) - \sum_{h=1}^{l}\sum_{j=1}^{m} \alpha_j^h d_{A,t}\left(o_j\right) \approx 0$$

which is true, again relying on the assumption about desire strength, because over a large number of occurrences, the outcome types designated by o_j will occur together with o_k with a frequency proportional to the conditional probability of their occurrence, given o_k. So my account satisfies this apparent constraint.

A second apparent constraint on a theory of reward is that for an agent that lacks desires, the theory should entail that reward just is the tendency to promote the aggregate satisfaction of basic drives. In purely habitual creatures, this is what we should expect the habit system to aim to maximise. And again, the present theory satisfies this constraint.

Third, this account coheres tolerably well with the account I have defended of how standing and occurrent desires are formed and modified. On this account of reward, in order to become more accurate as representations of reward, standing desires should be strengthened when new evidence is received, suggesting that they are more strongly probabilistically associated with the objects of other desires and drives than the previous evidence had suggested; and this is an empirically plausible account of the function of the desire-updating system (bearing in mind the uncertainty about that system detailed in section 3.5). Occurrent desires for outcome types should be strong when occurrent desires and drives for positively probabilistically-associated outcomes are strong; and again this seems to be how occurrent desires work – they form mutually-reinforcing networks, and are boosted by occurrent drives when incentive learning has taken place.

Fourth, on this account it is both possible for desires to be inaccurate *qua* representations, and for them to be very accurate (*perfect* accuracy is unlikely since we are dealing with continuously variable properties – it is unlikely for anyone to have a perfectly accurate desire, just as it is unlikely for one to have a perfectly accurate belief about e.g. one's own height). For example, consider an addict's desire for heroin, after they have ceased to find the drug pleasurable. This desire will be very strong, but the reward value of heroin for the addict will be determined by the extent to which taking heroin tends to promote or frustrate the addict's other desires and basic drives. Given that taking heroin will make the satisfaction of many of their other desires and drives much less likely than refraining from taking it would, this reward value may be strongly negative (although refraining would make highly aversive withdrawal symptoms very likely). So, as we would hope, the account entails that the desires for drugs of stereotypical addicts massively overestimate the reward values of those drugs. Assessing whether desires are likely to be accurate on this account is more difficult, because we know so little about how they are formed and updated. But assuming that this process relies on a reward prediction error signal and that desires and drives are used to measure reward levels for generating this signal, then it is plausible that our desires could sometimes be accurate. Depending on exactly how the system works, the signal may well be positive when outcomes are associated with objects of desires and drives to a greater degree than the agent expected; so the strengths of desires will tend to reflect the degree to which their objects are associated with the objects of other desires and drives.

Fifth, this account entails that outcomes will be rewarding to different degrees for different people, in a plausible way. One way in which differences between individuals of this kind might be thought to emerge is in virtue of its being the case that desiring an outcome in itself is sufficient to make that outcome rewarding. However, one of the strengths of my account is that it does not support this suggestion. It cannot be part of the function of the reward system to generate desires for outcomes that are probabilistically independent of the objects of our basic drives (although it is possible that such desires might occasionally be formed), but such desires could be accurate if desiring an outcome was sufficient to make it rewarding. We can see this point more clearly by thinking about how strong a *new* desire should be: in this case it is clear that only the extent to which the outcome concerned promotes the satisfaction of one's existing desires and drives should matter.

This approach still leaves room for variation between individuals from several sources. Although desiring some outcome does not *in itself* make that outcome more rewarding, humans have the capacity to experience and to desire pleasure, and what causes us pleasure is determined at least in part by what we desire. So most outcomes we desire will be rewarding for us in so far as their occurrence will cause us pleasure, which we desire. In addition to this, we vary in the strengths of our basic drives, and in the probabilistic relationships which hold between outcome types, where the outcomes concerned involve us. On the former point, it is highly likely that there is variation between individuals in the strengths of our innate drives for, for example, sugary foods or social success. These differences will have substantial knock-on effects in the rewardingness, for us, of a wide range of outcomes. On the latter point, our different skills, abilities, personalities and circumstances make a great deal of difference to whether certain outcomes will or will not promote the satisfaction of our basic drives and desires. To take just one example, for a talented athlete taking part in competitive sports will be strongly associated with receiving acclaim and enhanced social status, but for a less talented athlete this probabilistic (in this case causal) link will be much weaker.

Sixth, and finally, a possible objection to this account is that it gets the reward value of the objects of basic drives wrong. To take a particularly clear example, it surely cannot be the case that the reward value of having sex for an organism is determined by the extent to which this promotes the satisfaction of their other desires and basic drives. It is obviously appropriate for most organisms to have strong drives for sex regardless of the effect this has on their ability to get food, for instance. The reason why this objection does not succeed is that the role of basic drives is not to track the reward values of outcomes; instead, basic drive strengths should be expected to approximate the degrees to which their objects promote the organism's chances of survival and reproduction. So my view is that desires are pure indicatives representing the reward values of outcomes, which is to say that they represent the degrees to which these outcomes tend to promote the aggregate satisfaction of our other desires and basic drives.

9.3 Summary of the Argument So Far

In this final section of part II, I recap the argument so far. Over the last nine chapters, I have steadily built up a positive argument for my claim that desires have only the mind-to-world direction of fit, which I will survey here. The remaining chapter – chapter 10 – addresses one implication of my view, and responds to a possible objection. Rather than summarising each of the chapters in order, I will start from my conclusion and work backwards, explaining how the arguments and evidence I have presented support that conclusion and answer some of the questions that my view raises.

My conclusion is that desires have only the mind-to-world direction of fit. That is, they have indicative content but not imperative content. This claim may prompt several questions. One question which I have just discussed is what desires say; my provisional answer to this question is that desires represent the levels of reward that outcomes provide, where reward is the tendency to satisfy the agent's desires and basic drives. This may look worryingly self-referential, but the truth-conditions of individual desires are not fixed by those very desires. Instead, for desires to succeed as indicative representations their strengths must accurately reflect the extent to which their objects are positively probabilistically associated with the objects of the agent's other desires and basic drives. Reward is the 'common currency' used by the goal-directed and habit systems in assessing actions and outcomes, and the evidence from neuroscience and psychology suggests that these systems aim for the satisfaction of agents' desires, as well as their basic drives. These points raise very interesting questions about the relationships between reward and individual well-being, and between desires that are accurate as representations and those that it is rational or fitting to have, but I have not addressed these questions.

A further issue is that my conclusion may strike some readers as likely to involve a misappropriation of the notion of direction of fit. Isn't the whole point of this notion to illuminate a deep difference between desire and belief? This is an issue which I discuss in more detail in chapters 6 and 10. The key points of my view are: that direction of fit is a much more widespread phenomenon than the focus on desire and belief suggests; and that the primary notion of direction of fit is of a distinction between two kinds of representation, one of which aims to say how things are, while the other says what to do or what should be brought about. The Discretion View, the theory of direction of fit which I developed and defended in chapter 8, is unequivocally an account of the phenomenon understood in exactly this way.

However, the most important question my conclusion raises is why we should believe it. There is a powerful intuitive case for the idea that desires have indicative content, representing their objects as good in some way, because it would be so strange if we desired at random. But I have not developed this case in detail, and it does not support the more controversial part of my conclusion, which is that desires lack imperative content. Instead, my argument is based on an investigation of the nature of direction of fit in biological representations, conducted with no particular emphasis on desire. My conclusion may be seen as a corollary of the Discretion View, which is the thesis that what it takes for a representation to have indicative content is for it to have the function of co-occurring with specific normal conditions for the behaviour it causes in its consumer, and what it takes for a representation to have imperative content is for its consumer to have the function of performing a specific behaviour when adapted by this representation. This is a teleosemantic theory of direction of fit, so we can ask both why the teleosemantic approach is justified, and why this particular theory should be accepted.

I addressed the first of these two topics in chapters 6 and 7. In sections 6.4 and 7.1 I gave arguments in favour of a teleosemantic approach. Thinking about functions is a good way to think about many aspects of representation, particularly in the case of biological representations, because their status and properties as representations are determined by the roles they play in biological systems. To analyse representations we need to understand these roles, which means understanding how the systems in which they are embedded work, along with the nature of the contribution that representations make. The idea that representations are used to co-ordinate the behaviour of producers and consumers is a fundamental breakthrough in this project. Furthermore, the two directions of fit seem to be distinguished by what different kinds of representations are for. Those with the mind-to-world direction of fit are for keeping track of, and conveying information about, how things are, while those with the world-to-mind direction of fit are for specifying tasks and goals. So the directions of fit seem to be kinds of functions for representations. A further advantage of the teleosemantic approach for my purpose is that the kinds of properties that it appeals to in the analysis of representations are ones that desires have, which can be discovered by empirical methods.

My argument for the Discretion View was given in chapter 8. If we look at the issues with an open mind, I argued, we can see that for representations to have imperative content, their consumers must lack discretion. This is because (biological) representations can work in either or both of two ways: by co-occurring with some state of affairs which is relevant to how their consumers should behave, but which does not trivially determine appropriate behaviour; or by occurring when a specific behaviour on the part of the consumer is appropriate. In the first case, representations tell their consumers how things are, and in the second they tell them what to do. Many representations make a contribution which combines these two ways of working, and thus have both directions of fit.

What the Discretion View means for the direction of fit of desire is that *if* desires are biological representations with consumers that have discretion, then they are pure indicatives. So my argument also relies crucially on the account of the nature of desire that I gave in part I; this is a sense in which my conclusion is much more than a corollary of the Discretion View. Two aspects of the account of desire that I developed in part I make particularly important contributions to my argument. First, it is of course important that on my account, the systems that consume desires have discretion. What the goal-directed system should do when it is adapted by a particular occurrent desire depends not only on the object and strength of that desire, but also on the other desires that adapt the system at that time, and on the agent's instrumental beliefs. I argued that the goal-directed system is typically adapted by several occurrent desires at any given time in section 4.3. Standing desires are consumed by the system that generates occurrent desires, and this system has discretion because it is continually adapted by all of the agent's standing desires, but has the function of producing occurrent desires only for those outcomes that are salient at the time. The strengths of occurrent desires are also only partly determined by the strengths of standing desires.

Second, one of my major goals in part I was to show that desires are the members of psychological natural kind. In order to establish this conclusion, I argued that outcome values form a psychological natural kind, and have many of the most important properties that desires are normally thought to have. Outcome values combine instrumental beliefs to motivate a high proportion of human actions; they come in occurrent and standing forms; they have a wide range of objects, at least in humans, and are formed and modified by a process which is responsive to evidence to about the right extent; and they interact in roughly the ways that desires are thought to with habits, emotions and intentions. One reason why this is important is that outcome values have characteristic functions, established by biological processes, and their producers and consumers also have such functions with respect to outcome values. This means that outcome values, and hence desires, are biological representations, so the Discretion View applies.

Finally, I have also responded to two lines of objection that threatened to undermine my argument at relatively early stages. One was the objection that an empirically-informed philosophical theory of desire has already been developed, by Tim Schroeder (2004), and that he both settled on a theory that was extensionally distinct from mine, and argued (although not in detail) that desires have only the world-to-mind direction of fit. In chapter 5, I argued for the superiority of my account of desire over Schroeder's; the key difference was that on his view, even standing basic drives count as desires. This is most fundamentally a problem because standing basic drives are, in the terminology I introduced then, mental rules rather than mental states. The second line of objection was that teleosemantics has been discredited by the indeterminacy, swampman and liberality objections, and consequently cannot provide the framework for a correct theory of direction of fit. In chapter 7 I offered detailed responses to all three of these objections, and I also argued for the determinacy of direction of fit under the Discretion View in section 8.4.

In this chapter, I have shown that the combined results of parts I and II imply that both standing and occurrent desires are pure indicatives, or have only the mind-to-world direction of fit. I have also presented an account of the nature of reward which satisfies certain basic constraints, and serves to illuminate the content of desires. In the last section, I recapped some of the key claims and arguments of chapters 1-9.

Part III: Motivation

Chapter 10: The Humean Theory of Motivation

10.1 Introduction to Part III

One of the most influential discussions of direction of fit in the philosophical literature is by Michael Smith (1987, 1994), who appeals to what he takes to be the directions of fit of desire and belief in arguing for a claim known as the *Humean Theory of Motivation (HTM)*. HTM is in essence the claim that motivation by reasons requires the presence of desires, and cannot be generated by beliefs alone. Here in part III, I discuss the relationships between my view, Smith's argument, and the Humean Theory.

In this part of the thesis, my topic is the implications of the work of parts I and II for theories of reasons and rational motivation. The possible implications of empirically-informed investigations of processes leading to action for such theories are very extensive, because it is widely thought that reasons must in some sense be capable of motivating us; or to put it another way, because whatever acting for reasons involves, it must be something that humans actually do. This means that empirical questions such as those addressed in part I have fairly close connections to central issues in ethics and meta-ethics. For example, *internalism about reasons* is the thesis that all reasons for action are derived from those of the agent's mental states that are capable of motivating him or her. This claim is now most strongly associated with Bernard Williams (1979), but also plays a very important role in the moral theories of Hume and Kant, among many others. If internalism about reasons is true, then both empirical and philosophical research will be necessary to learn how reasons for action arise. Another important example of the potential importance of empirical research concerns perhaps the most famous argument for *noncognitivism about moral judgments*, which is the thesis that moral judgments are not beliefs. This argument, which is also prominent in Hume's moral philosophy, relies on the premises that moral judgments are intrinsically motivating, and that beliefs alone cannot motivate us (a version of HTM). On the face of it, both premises are empirical claims about the kinds of mental processes which lead to motivation and action.

However, in this chapter I will discuss only issues relating to the form of Humeanism advocated by Smith and James Lenman (1996). In the next section I describe Smith's version of the Humean Theory of Motivation, and outline his argument. In sections 10.3 and 10.4 I give a careful analysis of this argument, showing that there are some parts of Smith's Humeanism that are contradicted by my view, but other parts that are consistent with it. The crucial problem with Smith's argument is that he falsely assumes that mental states that set goals for action must have the world-to-mind direction of fit. In section 10.5 I discuss, and reject, a possible response that a Humean might make to my arguments, and in section 10.6 I consider the possibility that the dispute between the Humean and I is merely verbal.

One central purpose of this chapter is to respond to a likely objection to my view. This objection is that the Humean Theory of Motivation requires that desires and beliefs have opposite directions of fit, so my conclusion can be correct only if HTM is false. My response to this objection is that HTM itself is compatible with my view, although certain elements of Smith's argument for HTM are not. Largely as a result of Smith's influence, philosophical thought has tended to connect the representational property of direction of fit too closely with properties of mental states relating to reasons, motivation and explanation. So a further aim in this chapter (which I pursue simultaneously) is to disentangle these ideas.

10.2 The Humean Theory of Motivation and Smith's Argument

For the purposes of this chapter, I will take the Humean Theory of Motivation to be the following claim:

> *Humean Theory of Motivation (HTM)*: For an agent to be motivated by a reason, they must be motivated by a desire that they believe the action under consideration would help to satisfy.

There are two important points about this version of HTM which stand in need of further explanation.

First, it is important that this is a claim about *motivation by reasons*. To get clearer about this idea, it is helpful to contrast *motivating reasons* with *normative reasons*. Normative reasons are often described as considerations that count in favour of some course of action or other. They are the reasons that we ought to be sensitive to, and which contribute to making it the case that particular actions are right or justified, morally or otherwise. In contrast, motivating reasons are the considerations that can correctly be cited when giving rationalising explanations of motivation and action. Rationalising explanations explain actions by showing how they were rational or intelligible responses to the agent's situation. However, considerations which show how actions were rational or intelligible are only motivating reasons when they in fact helped to produce the agent's motivation – that is, when the agent was motivated by these considerations. So an agent has a motivating reason R if and only if: (i) R makes some action A a rational or intelligible thing for the agent to do; and (ii) the agent is motivated to perform A in part because they believe or notice R.

When things go well, our motivating reasons are normative reasons. Normative reasons are certainly considerations that are capable of making actions appear rational or intelligible, and it is common for us to be motivated by these reasons. For example, suppose I wanted to travel to London, and knew that the only train left at 10 o'clock. The fact that the only train left at this time would be a normative reason for me, given my desires, to go to the station just before 10. But it could also easily be a motivating reason for me to perform this action, if I had the relevant belief. However, things do not always go well, and sometimes we have motivating reasons which are not normative reasons. Suppose now that my belief that the London train left at 10 o'clock was false. Then, at least according to some views of reasons, I would have a motivating reason to go to the station just before 10, because I would be motivated to do this by a consideration that made this action rational or intelligible. But this consideration – that the London train left at 10 – would not be a normative reason for me to go to the station then, because it would not be true. It is also possible for us to have normative reasons which do not motivate us, for instance because we are unaware of them.

These points show that the present version of HTM is a theory of the kind of motivation of which rationalising explanation is possible. It claims that all motivation of this kind works in the same way as the travelling-to-London example just given: the agent has a desire for some outcome; they believe that some consideration implies that a certain action would help to satisfy this desire; and they are consequently motivated by this consideration to perform the action. However, it is important that this leaves open the possibility that we can also be motivated in other ways, as long as this motivation cannot be the subject of rationalising explanations. If it weren't for this point, the fact that we can be motivated by habits, in the absence of any relevant desire, would be a very strong objection to HTM. But given that HTM concerns motivation *by reasons*, the theory can be plausibly defended on the grounds that when we act out of habit we are not motivated by reasons. The fact that HTM concerns motivation by reasons is also important for Smith's argument, as we will shortly see.

The second point to note about this version of HTM is that it is weaker than some alternatives, in two noteworthy respects. A difference from Smith's own formulation of HTM is that this version does not claim that belief-desire pairs which generate motivation *are* motivating reasons, but only that they are necessary for motivation by reasons. I am using the slightly weaker formulation in order to avoid a controversy about whether mental states can be reasons (see Alvarez 2010).

In addition to this, in meta-ethics the Humean Theory of Motivation is often taken to be the claim that beliefs alone cannot motivate us. The crucial difference between this formulation (call it *HTM**) and the version of HTM I am considering is that HTM* rules out two ways in which beliefs could be solely responsible for motivation, which are left open by my HTM. These are that there could be some beliefs that are *at the same time desires* (sometimes called 'besires'), and that there are some beliefs that are capable on their own of causing new desires to be formed. Variants of these two strategies have been adopted as ways of explaining how moral judgments can be beliefs which are capable of motivating us independently of our antecedent desires, by philosophers including Nagel (1970), McDowell (1979), Darwall (1983) and Altham (1986). They are rejected by Humeans, who argue that beliefs and desires are 'distinct existences', so there can be no 'besires' (Smith 1987, 1994), and that desires cannot be rationally derived from beliefs alone (Lenman 1996). But this aspect of the Humean view – the insistence on the ontological and rational separateness of beliefs and desires – is not the aspect that conflicts with my view of the direction of fit of desire. Instead, the conflict comes from the Humean idea that motivation requires states with the world-to-mind direction of fit, and the strategies that HTM* opposes are precisely ways to make this Humean idea compatible with cognitivism about moral judgments (that is, the view that moral judgments are beliefs). So my concern is with the idea that desires are necessary for motivation, not with the idea that beliefs are not sufficient.

We can now turn to Smith's argument for HTM. He employs the following three premises (1994, p. 116):

A. Having a motivating reason is, *inter alia*, having a goal.
B. Having a goal is being in a state with which the world must fit.
C. Being in a state with which the world must fit is desiring.

From this Smith concludes that having a motivating reason involves having a desire, and from here it is only a short step to HTM, since HTM is the claim that being motivated by a such a reason involves being motivated by a desire. By 'being in a state with which the world must fit', Smith explains, he means being in a state with the world-to-mind direction of fit. In his (2011), he writes that this means having within oneself a representation of how the world is to be (p. 154). The reason why beliefs cannot provide motivating reasons is that they only purport to represent how things are, and thus cannot provide us with goals, a thought expressed vividly by James Lenman (1996, p. 300):

> A belief is a representation of the way we think the world is, a desire a representation of how we would have it be. Given this understanding of beliefs and desires respectively, it is only when we have some desire that we have anything that is even a candidate for a reason to interfere with the world… You can have representations of the way things are until you are blue in the face… but you don't begin to have a *reason* to interfere so long as you couldn't care less about the way things are. This possibility of caring less only arises when you start to prefer some possible ways for the world to be over others. And the direction of fit of any such preferences is world-word and not vice-versa.

The conflict between my view and the Humeanism of Smith and Lenman is therefore clear. I have claimed that desires have only the mind-to-world direction of fit; but Smith and Lenman claim that states with the world-to-mind direction of fit are necessary for motivation, and that these states are desires. The view which I argued for in detail in parts I and II contradicts premise C of Smith's argument, so if I am right, the argument fails. However, the implications of my view go further than this, as I explain in the next two sections.

10.3 Two Components of Humeanism

As I have mentioned, in this chapter one of my main aims is to respond to a possible objection to my view. The objection is that my conclusion is in some way incompatible with HTM, and now that we have seen Smith's argument, we can see better how such an objection may go. For suppose we accept HTM, along with the first two premises of Smith's argument:

> HTM: For an agent to be motivated by a reason, they must be motivated by a desire that they believe the action under consideration would help to satisfy.
> A: Having a motivating reason is, *inter alia*, having a goal.
> B: Having a goal is being in a state with which the world must fit.

Given premise A, the best explanation of why being motivated by a reason requires being motivated by a desire seems to be that desires set goals for action. In fact, it is hard to resist this conclusion, since it seems that what it is for some outcome to be a goal for action for an agent is – at least roughly – for them to be motivated by their belief that the action will help to bring about that outcome. So having a goal amounts to having a desire. This means that, given premise B, desires must have the world-to-mind direction of fit.

Although not deductive, this argument seems to show that my position can only be sustained if we reject either the Humean Theory of Motivation, premise A, or premise B. In this section I will argue that we should reject premise B. We can usefully think of Smith and Lenman's Humeanism as made up of two components. The first component includes HTM itself, premise A, and a further claim which I will mention shortly, all of which are compatible with my view, and all of which are concerned solely with the nature of motivation by reasons. The second component includes premises B and C, links ideas concerning motivation by reasons with direction of fit, and is in my view mistaken.

To reject the Humean Theory of Motivation would be the wrong strategy for dealing with the present objection. This is because an equally powerful argument against my conclusion can be generated by combining premises A and B with the much weaker claim that being motivated by beliefs and desires is *one way* of being motivated by reasons. If this weaker claim is true, then again it seems that desires must play the role of setting goals for action, given that having a motivating reason requires having a goal (premise A). So again it follows from premise B, the claim that having a goal consists in having an internal representation with the world-to-mind direction of fit, that desires must have the world-to-mind direction of fit. All that we lose by replacing HTM with the weaker claim is the result that desires are the *only* goal-setting states. We should accept that being motivated by beliefs and desires is one way of being motivated by reasons, because describing agents' beliefs and desires is the paradigmatic form of rationalising explanation. So we must reject either premise A, or premise B, or both.

Smith (1987, 1994) describes both A and B as 'unassailable', although he does not argue explicitly for B. The reason why he takes A to be unassailable is that it is an immediate consequence of the *teleological account* of rationalising explanation, to which he subscribes:

> *Teleological Account of Rationalising Explanation*: All rationalising explanations work by showing how the action concerned could be seen as appropriate, given some goal the agent had.

Those who accept the teleological account, like Smith, think that what it is to make an action appear rational or intelligible is to show how it could be seen as a rational means to a goal, aim or end that the agent had. Given that motivating reasons are considerations that can be cited in giving rationalising explanations, it follows that one can only have a motivating reason if one has a goal.

The teleological account is controversial, so there may be some reason to doubt premise A. In particular, consider the position of someone who believes that there are normative reasons for action that apply to all of us, regardless of our goals, desires or other attitudes. A philosopher with this view might well choose to argue that when someone is motivated by such a reason, it is possible to explain why they are so motivated just by giving this reason, without appealing (explicitly or implicitly) to any goal the agent had. However, just as was the case with HTM, we should not respond to the present objection by rejecting premise A, because the argument will still go through if it is replaced by a much weaker premise, which is much harder to deny. Consider the *Teleological Account of Belief-Desire Explanation*:

> *Teleological Account of Belief-Desire Explanation*: Rationalising explanations which cite the agent's beliefs and desires work by showing how the action concerned could be seen as appropriate, given some goal the agent had.

This claim is very highly plausible, and it follows that when an agent is motivated by a reason in virtue of having some belief and desire, that agent must have a goal. This is the weaker alternative to premise A that we need to make the objection go through.

My conclusion is therefore faced with the following objection:

> [Alternative to HTM]: Being motivated by beliefs and desires is one way of being motivated by reasons.
> [Alternative to A]: When an agent is motivated by a reason in virtue of having some belief and desire, that agent must have a goal.
> B: Having a goal is being in a state with which the world must fit.
> Conclusion: Desires are states with which the world must fit.

I find the first two premises of this argument genuinely unassailable, and again the inference from the premises to the conclusion is compelling, although not deductive. So my claim that desires have only the mind-to-world direction of fit is, in my view, incompatible with Smith's premise B. I therefore take my arguments of parts I and II, together with the present discussion, to amount to a powerful case against B. To put the point in Lenman's terms, this argument shows that we do care about things in virtue of certain representations of the way things are.

We can therefore distinguish two components of Smith and Lenman's Humean Theory. The first is a set of claims about motivating reasons and rationalising explanation that includes HTM itself, the teleological account of rationalising explanation, and Smith's premise A. These are united by the idea that having goals –

and therefore having desires – is necessary for motivation by reasons. My claim that desires have only the mind-to-world direction of fit is compatible with this component of Humeanism, because I accept that desires generate motivating reasons, in virtue of the fact that desiring an outcome is a way of having a goal. If this is right, the truth of the claims that make up this component turns on: whether there are other ways of having goals, besides having desires; and more fundamentally, whether there are other ways of having motivating reasons, besides having goals. Answering these questions is beyond the scope of this thesis, and I am therefore agnostic about HTM, the teleological account of rationalising explanation, and premise A.

The second component of Smith and Lenman's Humeanism includes premises B and C. That is, it includes the claims that goal-setting states must be representations with the world-to-mind direction of fit, and that all and only desires are such mental representations. So the key idea here is that, given the first component, representations with the world-to-mind direction of fit have a particular and crucial role to play in rational motivation. This component concerns how representational properties of mental states link up with their roles in rational motivation. If desires have only the mind-to-world direction of fit, then – given the plausible further claim that desires set goals for action – Smith and Lenman's views on this topic cannot be correct.

Although premises B and C are necessary for Smith's argument for HTM, it seems entirely coherent to accept HTM while rejecting these premises. For instance, Sinhababu (2009) defends the Humean Theory of Motivation on the grounds that there are no compelling cases of human motivation or action (for reasons) that cannot be elegantly explained by citing beliefs and desires. He lists five properties of desire which enter into his explanations of difficult cases, all of which fit readily into the account I developed in part I. In the next section I will further defend the claim that HTM is compatible with my view, by giving an account of how desires set goals for action which does not rely on the claim that they have the world-to-mind direction of fit.

10.4 How Desires Set Goals

As I have explained, although I am agnostic about HTM itself, I do accept that desires generate motivating reasons, and that they do so by setting goals for action. In the first half of this section I will outline a conception of goals, and of what it is for mental states to set them, which supports this claim and is also compatible with my view that desires have only the mind-to-world direction of fit. In the second half I will consider Smith's reasons for accepting his premise B.

The notion of a goal which is relevant here is constitutively tied to the teleological form of rational explanation. Teleological explanations explain actions or motivation by showing how they are rational or intelligible responses to the agent's situation, given the agent's goals. So what it is for an agent to have a goal is for there to be some end such that we can rationalise possible actions on the part of the agent, by showing how they might be thought to further this end. However, there are important ways in which this characterisation of what it is to have a goal may be interpreted too broadly.

First, crucially, it is not enough that we can make sense of an agent's action by attributing some end to them, for this to be among their goals; rather, it must be the case that their pursuit of the end concerned contributed to the rational process that led to their action. For example, imagine a man travelling by train from Glasgow to Fort William. It would certainly be possible in principle to rationalise this action by attributing to the man the goal of taking one of Britain's most scenic train journeys, but this is not sufficient to show that it was *his* goal. It may be that he had no interest in scenic train journeys, and simply wanted to visit someone in Fort William by the cheapest means available to him. So it seems that agents' goals are determined by their mental states, since they are determined by the factors that can contribute to the various processes of rational deliberation that they might go through.

Second, agents' goals are inputs to, not outputs from, processes of rational deliberation. To continue with the example, it is possible to imagine that someone might have the goal of travelling by train from Glasgow to Fort William, and this would allow us to make at least some sense of their doing so. But in a more typical case, doing this would be chosen as the best means to achieving some other goals, such as enjoying the scenery or spending time with a friend. Teleological explanations that cite less specific ends such as these will reflect more accurately what the agents concerned really care about. However, like all explanations, teleological explanations work best when they focus on features of the situation which are neither too specific, nor too general. In keeping with this, it will also usually be wrong to think of agents as having extremely general goals such as doing what is best for themselves, since these will typically fail to reflect what is distinctive about the agent concerned, and will not be explicitly considered in rational deliberation.

Our goals, then, are roughly the ends that we use our rational abilities, and in particular our capacity for means-ends reasoning, to pursue. So the mental states that set our goals are those that determine these ends.

With this in mind, desires set human agents' goals because humans tend to be motivated to do things that they believe will increase the likelihood of the outcomes they desire. The goal-directed system implements a rational process in which goal-setting states – desires – combine with instrumental beliefs in the selection of actions. Desires are inputs to the goal-directed system, rather than outputs of it, but they also vary between individuals and over time. Teleological explanations of human action and motivation are apt partly because many human actions are caused by the action of this system, and because it almost always functions at least well enough that the action concerned will be intelligible as the pursuit of one of the outcomes the agent desires. However, this is not to say that desires are the only mental states that are capable of setting goals. For instance, it may be that evaluative judgments can also do so, if they are capable of motivating us through processes of practical reasoning that are suitably independent of the goal-directed system.

These points may not be surprising, but they are worth making for the sake of emphasising that they are consistent with my claim that desires have only the mind-to-world direction of fit. They show that if the Discretion View is correct, then there is a set of possible biological functions which is sufficient to entail that the mental states that have them set goals for action, but which is not sufficient to entail that these mental states have imperative content.

It may strike readers familiar with Smith's work that he does have an argument for his premise B (i.e. the claim that having a goal is a matter of being in a state with the world-to-mind direction of fit), and that I have not explicitly responded to it so far in this chapter. If Smith has an argument for this claim, it centres on his account of direction of fit, which includes the proposal that states with the world-to-mind direction of fit towards a proposition *p* are roughly (Smith's qualification) those that tend to endure in the presence of a perception that *not p*, and dispose the subject to bring it about that *p* (1987, p. 54). Smith is explicitly proposing a functional-dispositional characterisation of direction of fit in mental states, and consequently also of desire, since he takes desires to be defined by their direction of fit. But it would also be reasonable to think of him as proposing a functional-dispositional characterisation of *having a goal*, since this is on my account a matter of being disposed to undergo processes of practical reasoning concerning a particular end.[30] And presumably this account of having a goal would be the same as his theory of the world-to-mind direction of fit. So Smith's argument for premise B might be that he has given a theory which unifies the goal-setting property of mental states with the representational property of having the world-to-mind direction of fit.

Smith and I may well agree, then, about what it is for mental states to set goals for action, and hence for agents to have goals. But there is no need to go any further than we just have in trying to understand Smith's reasons for accepting premise B, because they will certainly rely on his view of direction of fit, and this is contradicted by the Discretion View. According to Smith's theory of direction of fit, desires have the world-to-mind direction of fit, even if we adopt my account of desire and drop Smith's idea that desires are defined by their direction of fit. Desires in my sense are such that they tend to endure in the presence of perceptions that represent that they are unsatisfied, and they dispose us to act so as to bring about their objects. So if it succeeds, my argument for the Discretion View shows that Smith's theory of direction of fit is false. This point is particularly clear since my argument focuses on the extensions of the directions of fit. So whatever Smith might say in defence of premise B, my response would be the same: that desires set goals, and the correct theory of direction of fit implies that desires have only the mind-to-world direction of fit, so premise B cannot be right.

[30] Note that my although my account of desire is a natural-kind view, my account of having a goal is functional. This leaves open the possibility that states other than desires could set goals for action. I am not sure whether having a goal is a matter of teleological function, as representational properties are, or a matter of 'function' in the sense employed in traditional functionalism in the metaphysics of mind (which is also the sense in which Smith uses this term). But I do not think this issue is important for present purposes.

10.5 Going Up a Level

In this section, I will consider and reject a possible line of response that Humeans might attempt against my arguments. One way in which the disagreement between Humeans and anti-Humeans is often brought out is by considering cases in which we might ordinarily describe people as motivated not by their desires but by their beliefs. For example, we hear of philosophers who 'have no desire' to attend committee meetings, but believe that it is their duty to do so. According to the anti-Humeans, the motivation these philosophers experience does not depend on their having any particular desires, but only on their beliefs about their duty. One response Humeans can make might be called 'going up a level'; they respond that the philosophers will be motivated by beliefs about their duty only if they desire to do their duty. Going up a level in this way is also a possible response to my claim that desires lack the world-to-mind direction of fit. Very roughly, the idea would be that even though we seem to be motivated by states that don't fit the Humeans' theory (in the meta-ethical case, because they are beliefs; in the present case, because they have the wrong direction of fit), there are in fact states that do fit the theory at 'the level above'.

There are two possible forms that such a response might take. The more ambitious form would argue that while the states that I have called desires do have only the mind-to-world direction of fit, they are not in fact desires. Instead, on this proposal, they would be beliefs about reward. We would each desire reward, and would represent it as to be brought about, and reward – rather than the outcomes that we represent as rewarding – would be the goal that we pursue when our behaviour is controlled by the goal-directed system. These claims would each be grounded on the point that we are in some sense disposed to pursue reward, in virtue of the way that the goal-directed and habitual systems work; these systems cause us to perform the actions that we represent as most rewarding, when they work according to their functions. If this proposal were correct, then all of the Humeans' claims could be true, consistently with my analysis of the representational properties of the states I have called desires. A more conservative form would drop the idea that what we *desire* is reward, but would continue to insist that we represent reward as to be brought about. If the more conservative form was preferred, the Humean could continue to maintain both that representations with the world-to-mind direction of fit are necessary for rational motivation (presumably in order to determine goals), and that desires were also necessary, a point which I have not disputed. If the bolder form was preferred, they could further claim that the imperative representations concerned *were* the necessary desires. In either case, they could maintain valuable parts of their doctrine which are threatened by my arguments.

However, neither form of this response is plausible, as I will now argue. We do not in general desire reward; nor do we represent it as to be brought about; and nor is it generally among our goals.

The reason why it is wrong to suggest that we desire reward goes back to a distinction drawn in chapter 5. There I argued that standing basic drives are not desires, because desires are mental *states*, while standing basic drives are mental *rules*. The same point applies here, because the fact that we are disposed to perform the actions that we represent as most rewarding is grounded on the existence of a mental rule (or perhaps of more than one mental rule), not a mental state. Mental rules are, by definition, innate features each of which causes exactly one kind of transition between mental states. The goal-directed system is innate, and disposes us to pursue reward by causing transitions from desires and instrumental beliefs, to further representations which tend to control the behaviour of downstream systems for selecting and generating actions. This system is centred on a mental rule that governs transitions from desires and instrumental beliefs to these further representations. We do not learn to pursue reward via the goal-directed system, and our disposition to do so is best explained in biological, rather than psychological, terms. To the extent that our disposition to pursue reward is also mediated by the habit system and downstream action-selection systems (the ‘Action Selector’ in the diagram in section 3.6), the mental rules that govern these systems will also be partially responsible for the disposition in question. But again, there is little doubt that these systems are innate, and consequently that the disposition to pursue reward is grounded on mental rules, not mental states.

The second suggestion by the Humean is that we possess mental representations with the world-to-mind direction of fit, that represent reward as something to be brought about. The fact that our disposition to pursue reward is grounded on mental rules is one reason to doubt this; it is appealing to think of rules and representations as distinct and complementary features of computational systems. In addition to this, it is doubtful whether the psychological feature that underlies our disposition to perform rewarding actions could be a representation, because typically representations either vary over time or sometimes occur, and sometimes fail to occur. This allows them to influence the behaviour of co-operating systems, by carrying information about some state of affairs which obtains at some times and not at others. Representations which do not do this rarely, if ever, occur, because systems do not need signals to help them to adapt to unchanging states of affairs. There may be exceptions to this rule, in which constant representations are used in the implementation of certain computational processes, but we have no reason to think that this is such a case. Instead, the feature we are concerned with seems to be a paradigmatic example of a rule for how representations are to be handled, rather than a representation itself.

Finally, there is the thought that our only goal is reward. The Humean has little reason to defend this idea without either of the previous two claims, because I have agreed that the presence of goals is a part of what makes it the case that we are motivated by reasons when we act on our desires. Without the previous two claims, this point would not allow the Humean to defend either the idea that goals are set by states with the world-to-mind direction of fit, or that these states are desires. But in any case, this thought has little to recommend it. One problem is that we are less rational in our pursuit of reward than in our pursuit of the outcomes we desire, because of the ways in which our desires can and cannot be updated. Because our desires can only be changed by reward signals, we fairly frequently desire outcomes despite believing that those outcomes are unhealthy, or incompatible with other outcomes that we desire. So we are motivated to pursue these outcomes despite having good evidence that they are not, on balance, rewarding. Because we are irrational in pursuing reward, to this extent, rationalising explanations that take reward to be our goal will be less successful than explanations that take us to be aiming at satisfying our desires. Of course, we are sometimes motivated to perform actions that we have good reasons to believe will not promote the satisfaction of our desires, because we are not perfectly rational in updating our instrumental beliefs; and this does affect the quality of rationalising explanations. A rationalising explanation is less powerful if it involves attributing to an agent a belief which it is hard to understand their having. But this problem is significantly more severe if our goal is reward. A related issue is that people's desires vary considerably, and this is apparently crucial to explaining differences in their behaviour, but it is hard to understand this variation in desires as a product of rational processes.

Another, simpler problem is that we care about the things we desire, in a way that we do not care about reward. Given that reward is aggregate drive- and desire-satisfaction, it would not normally make sense for someone to prefer that their desires were satisfied, over getting reward. But it certainly seems that the reason we would care about aggregate drive- and desire-satisfaction, if we thought about it, is that we care about the objects of our desires and drives. Consider a cyclist with a strong desire to race in the Tour de France. If our goal is reward, then the cyclist has a reason to want this, which is that racing the Tour would be rewarding for him, because he desires it and things associated with it. He should feel no special connection with this objective, because any other strong desire he might have would also offer the prospect of reward. But this is not an attractive picture of the cyclist's concerns; it is more plausible to suggest that if he has a reason to want to ride the Tour it is that it is the world's most iconic and challenging race. To a significant extent, though, the cyclist would deny that he wants to ride the Tour for any reason other than that it's the Tour. That is what both desires and goals are like – they are not wholly rational, because they do not serve other desires and goals, although they are often connected to them. Thinking of reward as our goal draws the line between the rational and the non-rational factors in motivation in the wrong place. So 'going up a level' would not be a good way for the Humean to respond to my arguments.

10.6 Dissolving the Disagreement?

In this chapter, I have argued that the Humean view put forward by Smith and Lenman goes wrong in accepting what I have called the second component of their view, which links desire and goal-setting to the world-to-mind direction of fit. Smith and Lenman claim, in my view falsely, that having a goal is a matter of representing some outcome as to be brought about, and that desiring also involves a representation of this kind. As I will explain in this final section, though, it is possible to interpret Smith and Lenman as having a different phenomenon in mind when they talk about direction of fit from the one I have investigated. That is, it is possible that the Humeans and I talk past each other when we talk about direction of fit. And on this alternative interpretation of Smith and Lenman's views, I do not disagree with them about the relationship between having a goal, desiring, and (what they call) the world-to-mind direction of fit.

One aspect of the way that Smith and many other philosophers think of direction of fit is that it is a way to distinguish between beliefs and desires. If this idea is really at the heart of Smith's thinking about direction of fit, then it may be that he would be prepared to give up the idea that the directions of fit are two kinds of representation, rather than the claim that desires, which set goals, have one direction of fit, while beliefs, which keep track of how things are, have the other. In this case, direction of fit would be fundamentally a property of mental states, rather than of representations. Exactly how Smith's views would relate to mine in this case depends on the details of the alternative conception of direction of fit. Specifically, it goes back to an issue that I discussed in section 6.3.

As I described in that section, there are two different ways in which philosophers elaborate on the idea that mental representations are essential components of states such as beliefs and desires. The first way, which I have adopted, takes states such as beliefs and desires to *be* mental representations. This view implies that unless the terms 'content' and 'direction of fit' are ambiguous, the content of a belief or desire is the content of the mental representation with which it is identical, and the direction of fit of a belief or desire is the direction of fit of that same mental representation. The second takes beliefs and desires to be *attitudes to* mental representations, which themselves lack directions of fit. These mental representations merely stand for states of affairs, somewhat like the italicised expression in the sentence, 'If *Argentina has the strongest squad*, then they deserve to win the world cup.' On this view, the content of a mental state is the representation it involves, and is hence distinct from the content of the representation itself; and the direction of fit of the mental state is a property of the attitude, and is not the same as the direction of fit of the representation, since the representation lacks direction of fit. It is clear that which of these two views one accepts will make a difference to how one thinks of direction of fit, especially if one takes it to be primarily a property of mental states rather than representations. So I will first discuss the position of a Humean who accepts the first view, then that of one who accepts the second.

On the first view, beliefs and desires are representations, just as sentences, maps, traffic signals, many paintings, and the communicative signals of non-human animals are too. I take it to be accepted on all sides that there is a property of these non-mental representations that can reasonably be called direction of fit, that divides them into various kinds, depending on whether they have imperative content, indicative content, both or neither. If this is right, then it is very likely that beliefs and desires have directions of fit in this sense. The claim that I have defended in part II is that desires have only the mind-to-world direction of fit, *in this sense of direction of fit*. So one way in which the disagreement between the Humean and I could be dissolved is if they have a different sense of direction of fit in mind. The only thing that the Humean's sense of direction of fit could amount to, on these assumptions, is just what they take to distinguish beliefs from desires. The idea would be roughly that what it is for a mental state to have the world-to-mind direction of fit is for it to set a goal for action – that is, for it to be of a kind that interacts in the right way with instrumental beliefs in processes of rational deliberation. What it is for a mental state to have the mind-to-world direction of fit might be, say, that it is of a kind that is correctly formed and updated by theoretical reasoning. How this related to the classification of mental representations in respect of the properties they share with non-mental representations would be a further issue.

If this is the sense of direction of fit that Humeans have in mind, then I have no disagreement with them. I share their view that desires set goals for action, and hence that in this sense they have the world-to-mind direction of fit. I am agnostic about whether desires are the only mental states like this, and about whether HTM and the teleological account of rationalising explanation are correct. However, we should note that this interpretation comes with a significant cost. It means that parts of the way that Smith and Lenman express their Humeanism, which emphasise that beliefs and desires are representations and that direction of fit is a matter of representing how things are or what to do, are seriously misleading. This is because on this view direction of fit is not about what beliefs and desires are like *qua* representations.

On the second view, direction of fit is a property of attitudes to representations. On this view, if sentences, maps, and traffic signals have directions of fit it is not in the same sense that beliefs and desires do, since these *are* representations, not attitudes to representations. If this is the sense of direction of fit that Humeans have in mind, then I do not exactly dispute their claim that desires have the world-to-mind direction of fit, or that this is essential to goal-setting. Instead, I very much doubt a presupposition of this claim, which is that beliefs and desires are attitudes to representations rather than being representations themselves. However, I have not argued this point, and will not do so here. Again, if this is the right interpretation of Smith and Lenman, then some parts of the way they express their view are misleading, such as their calling beliefs and desires ‘representations’.

In this chapter, I have considered the implications of my claim that desires have only the mind-to-world direction of fit for a certain form of Humeanism about rational motivation. I have argued that this claim is consistent with the Humean Theory of Motivation itself and with some associated theses, such as the teleological account of rationalising explanation. However, assuming that Smith, like me, takes direction of fit to be a property that beliefs and desires share with non-mental representations, my claim is inconsistent with one of the premises in his argument for HTM, and also gives us good reason to deny another. As I understand their views, Smith and Lenman go wrong in thinking that a certain property of some representations – the world-to-mind direction of fit – is necessarily co-extensive among mental states with the property of setting goals for action. This idea is mistaken, because desires have the latter property and not the former.

Conclusion

It is a platitude, but nonetheless true, that the value of a project often lies as much in the journey as in the destination. My conclusion that desires have only the mind-to-world direction of fit is significant, as I explained in part III, but is also a corollary of two broader claims, which I defended in parts I and II respectively. The first is the thesis that what it is for something to be a desire is for it to be a member of the natural kind of psychological state that functions as an input to the goal-directed control system, with the role of tracking the reward values of outcomes. The second is the Discretion View: the claim that biological representations have imperative content if their consumers lack discretion, and have indicative content if their producers lack discretion. Each of these three claims – the corollary, and the two theses from which it follows – has its own implications, and raises its own questions.

One implication of my claim that desires have only the mind-to-world direction of fit, discussed in part III, is that Smith's very well-known argument for the Humean Theory of Motivation fails. However, the Humean Theory itself is not contradicted by any point I have defended. Somewhat more broadly, Smith's argument has popularised the idea that beliefs and desires are distinguished by their directions of fit, which are properties they share with other representations. In connection with this, it has also encouraged the idea that the directions of fit are necessarily co-extensive among mental states with certain roles in practical and theoretical reasoning. These are the roles of setting goals for action, performed by desires and perhaps also by other states, such as evaluative beliefs, and of keeping track of states of affairs which are relevant to determining how to promote our goals, performed most obviously by instrumental beliefs. These two closely-related ideas, which are not often questioned, are both false if my claim about the direction of fit of desire is correct. Whether or not my arguments succeed, I hope to have shown that it ought to be regarded as an open and substantive question how the properties of beliefs and desires as representations, and their properties as components of rational processes, relate to one another.

Meanwhile, the account of desire developed in part I is of interest independently of the issue of direction of fit, because desires are central to many philosophical theories, as well as having profound impacts on our lives. Having one's desires satisfied – or perhaps some privileged subset of them – is thought by some philosophers to be constitutive of well-being. Desires are also central to theories of free will, self-control and moral responsibility. It has recently been argued that what it is to be virtuous is to have and act on the correct desires (Arpaly & Schroeder 2014). And desires also figure prominently in theories in meta-ethics, particularly concerning the nature of practical reasons and of rational motivation. In addition to all this, as I noted in the introduction to this thesis, our desires both have great influence on our behaviour and experiences in the short term, and shape the projects and concerns that run through our entire lives. So knowing what desires are can help us to understand, and consequently to critique or defend, all of these philosophical theories, and can also help us to understand our own choices and feelings. The scientific evidence and theories that I present in part I should be the subject of much greater philosophical scrutiny than is currently the case; the issue concerning direction of fit that I have chosen to explore here is just one of many possible implications.

Finally, the Discretion View offers a new response to a foundational question about representation. I suspect that no one, simple formula can specify with complete generality what it takes for an entity to be a representation with a particular content and direction of fit. Instead, different theories (or theses within a grand theory) will be needed for representations of different kinds. However, even if this is correct, it is also plausible that there is some basic class of representations from which the others derive, in various ways. For example, representations that stand for individual objects or properties may be contentful, and have their content determined, in virtue of the roles they play in semantically complex representations with indicative or imperative content. Or sentences in natural languages may have their content and status as representations determined in part by the relationships they stand in to mental representations. If there is such a basic class of representations, then biological representations – that is, those that have directions of fit in virtue of their biological functions and those of their producers and consumers – are an excellent candidate. This is one of the deepest ideas of the teleosemantic movement. So biological representations are an important subject.

Also, as I discussed at several stages in chapters 6, 7 and 8, direction of fit in biological representations is intimately tied to their status as representations, and their content. In particular, what gives biological representations their status as representations is their having either the function that constitutes the mind-to-world direction of fit, or the function that constitutes the world-to-mind direction of fit, or both.

Bibliography

Adams, Christopher D. 1982. “Variations in the Sensitivity of Instrumental Responding to Reinforcer Devaluation.” *The Quarterly Journal of Experimental Psychology Section B* 34 (2): 77–98.

Adams, Christopher D., and Anthony Dickinson. 1981. “Instrumental Responding Following Reinforcer Devaluation.” *The Quarterly Journal of Experimental Psychology Section B* 33 (2): 109–21.

Alexander, G. E., M. R. DeLong, and P. L. Strick. 1986. “Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex.” *Annual Review of Neuroscience* 9: 357–81.

Altman, J. E. J. 1986. “The Legacy of Emotivism.” In *Fact, Science and Morality: Essays on A.J. Ayer’s Language, Truth and Logic*, edited by Graham Macdonald and Crispin Wright. Oxford: Basil Blackwell.

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.

Anscombe, G. E. M. 1957. *Intention*. Harvard University Press.

Arbuthnott, Gordon W., and Jeff Wickens. 2007. “Space, Time and Dopamine.” *Trends in Neurosciences* 30 (2): 62–69.

Arpaly, Nomy, and Timothy Schroeder. 2014. *In Praise of Desire*. Oxford University Press.

Artiga, Marc. 2013. “Teleosemantics and Pushmi-Pullyu Representations.” *Erkenntnis*, 1–22.

Balleine, Bernard W. 1992. “Instrumental Performance Following a Shift in Primary Motivation Depends on Incentive Learning.” *Journal of Experimental Psychology: Animal Behavior Processes* 18 (3): 236–50.

Balleine, Bernard W., Nathaniel D. Daw, and John P. O’Doherty. 2008. “Multiple Forms of Value Learning and the Function of Dopamine.” *Neuroeconomics: Decision Making and the Brain*, 367–85.

Balleine, Bernard W., A. Simon Killcross, and Anthony Dickinson. 2003. “The Effect of Lesions of the Basolateral Amygdala on Instrumental Conditioning.” *The Journal of Neuroscience* 23 (2): 666–75.

Balleine, Bernard W., and John P. O'Doherty. 2010. "Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action." *Neuropsychopharmacology* 35 (1): 48–69.

Balleine, Bernard W., and Anthony Dickinson. 1998. "Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates." *Neuropharmacology* 37 (4-5): 407–19.

Barto, Andrew. 2007. "Temporal Difference Learning." *Scholarpedia* 2 (11): 1604.

Baumeister, R. F., E. Bratslavsky, M. Muraven, and D. M. Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology* 74 (5): 1252–65.

Baumeister, Roy F., Todd F. Heatherton, and Dianne M. Tice. 1994. *Losing Control: How and Why People Fail at Self-Regulation.* Academic Press.

Bechara, A., H. Damasio, D. Tranel, and A. R. Damasio. 1997. "Deciding Advantageously before Knowing the Advantageous Strategy." *Science (New York, N.Y.)* 275 (5304): 1293–95.

Berridge, Kent C. 2007. "The Debate over Dopamine's Role in Reward: The Case for Incentive Salience." *Psychopharmacology* 191 (3): 391–431.

Berridge, Kent C., and Terry E. Robinson. 2011. "Drug Addiction as Incentive Sensitization." In *Addiction and Responsibility*, edited by J. Poland and G. Graham. Cambridge, MA: MIT Press.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information.

Bray, Signe, Shinsuke Shimojo, and John P. O'Doherty. 2010. "Human Medial Orbitofrontal Cortex Is Recruited during Experience of Imagined and Real Rewards." *Journal of Neurophysiology* 103 (5): 2506–12.

Bromberg-Martin, Ethan S., Masayuki Matsumoto, and Okihide Hikosaka. 2010. "Dopamine in Motivational Control: Rewarding, Aversive, and Alerting." *Neuron* 68 (5): 815–34.

Burge, Tyler. 2010. *Origins of Objectivity*. Oxford University Press.

Burke, Kathryn A., Theresa M. Franz, Danielle N. Miller, and Geoffrey Schoenbaum. 2007. "Conditioned Reinforcement Can Be Mediated by Either Outcome-Specific or General Affective Representations." *Frontiers in Integrative Neuroscience* 1: 2.

Cagniard, Barbara, Peter D. Balsam, Daniela Brunner, and Xiaoxi Zhuang. 2006. "Mice with Chronically Elevated Dopamine Exhibit Enhanced Motivation, but Not Learning, for a Food Reward." *Neuropsychopharmacology* 31 (7): 1362–70.

Cannon, Claire Matson, and Richard D. Palmiter. 2003. "Reward without Dopamine." *The Journal of Neuroscience* 23 (34): 10827–31.

Carroll, Lewis. 1895. "What the Tortoise Said to Achilles." *Mind* 4 (14): 278–80.

Cisek, Paul. 2007. "Cortical Mechanisms of Action Selection: The Affordance Competition Hypothesis." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1485): 1585–99.

Clark, Andy. 2013. "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36 (3): 181–204.

Clark, Andy, and David J. Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

Coccaro, Emil F., Michael S. McCloskey, Daniel A. Fitzgerald, and K. Luan Phan. 2007. "Amygdala and Orbitofrontal Reactivity to Social Threat in Individuals with Impulsive Aggression." *Biological Psychiatry* 62 (2): 168–78.

Corbit, Laura H., and Bernard W. Balleine. 2003. "The Role of Prelimbic Cortex in Instrumental Conditioning." *Behavioural Brain Research* 146 (1-2): 145–57.

Corbit, Laura H., J. L. Muir, and B. W. Balleine. 2001. "The Role of the Nucleus Accumbens in Instrumental Conditioning: Evidence of a Functional Dissociation between Accumbens Core and Shell." *The Journal of Neuroscience* 21 (9): 3251–60.

Cummins, Robert C. 1975. "Functional Analysis." *Journal of Philosophy* 72: 741–64.

Damasio, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Darwall, Stephen L. 1983. *Impartial Reason*. Cornell University Press.

Davidson, Donald. 1987. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60 (3): 441–58.

Davidson, R. J., K. M. Putnam, and C. L. Larson. 2000. "Dysfunction in the Neural Circuitry of Emotion Regulation–a Possible Prelude to Violence." *Science (New York, N.Y.)* 289 (5479): 591–94.

Davies, Martin. 2015. “Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects.” In *International Encyclopedia of the Social and Behavioral Sciences*, edited by J. D. Wright, Second Edition. Oxford: Elsevier.

Daw, Nathaniel D., Yael Niv, and Peter Dayan. 2005. “Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control.” *Nature Neuroscience* 8 (12): 1704–11.

Dennett, Daniel C. 1983. “Styles of Mental Representation.” *Proceedings of the Aristotelian Society* 83: 213–26.

Dickinson, A., B. Balleine, A. Watt, F. Gonzalez, and R. A. Boakes. 1995. “Motivational Control after Extended Instrumental Training.” *Animal Learning & Behavior* 23 (2): 197–206.

Dickinson, Anthony, and G. R. Dawson. 1988. “Motivational Control of Instrumental Performance: The Role of Prior Experience of the Reinforcer.” *The Quarterly Journal of Experimental Psychology Section B* 40 (2): 113–34.

Dickinson, Anthony, and D. J. Nicholas. 1983. “Irrelevant Incentive Learning during Training on Ratio and Interval Schedules.” *The Quarterly Journal of Experimental Psychology Section B* 35 (3): 235–47.

Dickinson, A., S. Squire, Z. Varga, and J. W. Smith. 1998. “Omission Learning after Instrumental Pretraining.” *The Quarterly Journal of Experimental Psychology Section B* 51 (3): 271–86.

Doolittle, W. Ford, Tyler D. P. Brunet, Stefan Linquist, and T. Ryan Gregory. 2014. “Distinguishing between ‘Function’ and ‘Effect’ in Genome Biology.” *Genome Biology and Evolution* 6 (5): 1234–37.

Döring, Sabine A. 2010. “What a Difference Emotions Make.” In *A Companion to the Philosophy of Action*, edited by Timothy O’Connor and Constantine Sandis. Wiley-Blackwell.

Dretske, Fred. 1988. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Ekman, Paul. 1992. “Are There Basic Emotions?” *Psychological Review* 99 (3): 550–53.

Field, Hartry. 1978. “Mental Representation.” *Erkenntnis* 13 (July): 9–61.

Foddy, Bennett, and Julian Savulescu. 2010. “A Liberal Account of Addiction.” *Philosophy, Psychiatry, and Psychology* 17 (1): 1–22.

Fodor, Jerry A. 1975. *The Language of Thought*. Harvard University Press.

———. 1990. “A Theory of Content I.” In *A Theory of Content*. MIT Press.

Glimcher, Paul W. 2011. “Understanding Dopamine and Reinforcement Learning: The Dopamine Reward Prediction Error Hypothesis.” *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 3: 15647–54.

Godfrey-Smith, Peter. 1992. “Indication and Adaptation.” *Synthese* 92 (2): 283–312.

———. 1996. *Complexity and the Function of Mind in Nature*. Cambridge University Press.

———. 2013. “Signals, Icons and Beliefs.” In *Millikan and Her Critics*, edited by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons.

Goode, R., and P. E. Griffiths. 1995. “The Misuse of Sober’s Selection for/Selection of Distinction.” *Biology and Philosophy* 10 (1): 99–108.

Gospic, Katarina, Erik Mohlin, Peter Fransson, Predrag Petrovic, Magnus Johannesson, and Martin Ingvar. 2011. “Limbic Justice–Amygdala Involvement in Immediate Rejection in the Ultimatum Game.” *PLoS Biology* 9 (5): e1001054.

Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B. R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. 2013. “On the Immortality of Television Sets: ‘Function’ in the Human Genome according to the Evolution-Free Gospel of ENCODE.” *Genome Biology and Evolution* 5 (3): 578–90.

Griffiths, Paul. 2009. “In What Sense Does ‘Nothing Make Sense Except in the Light of Evolution’?” *Acta Biotheoretica* 57 (1-2): 11–32.

Haber, Suzanne N., and Brian Knutson. 2010. “The Reward Circuit: Linking Primate Anatomy and Human Imaging.” *Neuropsychopharmacology* 35 (1): 4–26.

Hammond, Lynn J. 1980. “The Effect of Contingency upon the Appetitive Conditioning of Free-Operant Behavior.” *Journal of the Experimental Analysis of Behavior* 34 (3): 297–304.

Hare, Todd A., Colin F. Camerer, and Antonio Rangel. 2009. “Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System.” *Science (New York, N.Y.)* 324 (5927): 646–48.

Hare, Todd A., Jonathan Malmaud, and Antonio Rangel. 2011. “Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice.” *The Journal of Neuroscience* 31 (30): 11077–87.

Hawkins, Jennifer. 2008. “Desiring the Bad Under the Guise of the Good.” *Philosophical Quarterly* 58 (231): 244–64.

Hnasko, Thomas S., Bethany N. Sotak, and Richard D. Palmiter. 2005. "Morphine Reward in Dopamine-Deficient Mice." *Nature* 438 (7069): 854–57.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford University Press.

Holton, Richard, and Kent Berridge. 2014. "Addiction Between Compulsion and Choice." In *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*, edited by Neil Levy. Oxford University Press.

Horvitz, Jon C. 2009. "Stimulus-Response and Response-Outcome Learning Mechanisms in the Striatum." *Behavioural Brain Research* 199 (1): 129–40.

Hull, Clark L. 1943. *Principles of Behavior: An Introduction to Behavior Theory*. Appleton-Century Company, Incorporated.

Humberstone, I. L. 1992. "Direction of Fit." *Mind* 101 (401): 59–83.

Hursthouse, Rosalind. 1991. "Arational Actions." *Journal of Philosophy* 88 (2): 57–68.

Hyman, Steven E. 2005. "Addiction: A Disease of Learning and Memory." *The American Journal of Psychiatry* 162 (8): 1414–22.

Kaller, Christoph P., Benjamin Rahm, Joachim Spreer, Cornelius Weiller, and Josef M. Unterrainer. 2011. "Dissociable Contributions of Left and Right Dorsolateral Prefrontal Cortex in Planning." *Cerebral Cortex (New York, N.Y.: 1991)* 21 (2): 307–17.

Kennerley, Steven W., and Mark E. Walton. 2011. "Decision Making and Reward in Frontal Cortex: Complementary Evidence from Neurophysiological and Neuropsychological Studies." *Behavioral Neuroscience* 125 (3): 297–317.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray. 2011. "Speed/Accuracy Trade-off between the Habitual and the Goal-Directed Processes." *PLoS Computational Biology* 7 (5): e1002055.

Killcross, Simon, and Etienne Coutureau. 2003. "Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats." *Cerebral Cortex (New York, N.Y.: 1991)* 13 (4): 400–408.

Kirk, Ulrich, Martin Skov, Oliver Hulme, Mark S. Christensen, and Semir Zeki. 2009. "Modulation of Aesthetic Value by Semantic Context: An fMRI Study." *NeuroImage* 44 (3): 1125–32.

Klein, Colin. 2007. "An Imperative Theory of Pain." *Journal of Philosophy* 104 (10): 517–32.

Kripke, Saul A. 1980. *Naming and Necessity*. Harvard University Press.

Landreth, Andrew. 2009. “The Emerging Theory of Motivation.” In *The Oxford Handbook of Philosophy and Neuroscience*, edited by John Bickle. Oxford University Press.

Lenman, James. 1996. “Belief, Desire and Motivation: An Essay in Quasi-Hydraulics.” *American Philosophical Quarterly* 33 (3): 291–301.

Levy, Neil. 2014. *Addiction and Self-Control: Perspectives From Philosophy, Psychology, and Neuroscience*. Oxford University Press.

Lewis, David. 1969. *Convention: A Philosophical Study*. Harvard University Press.

———. 1980. “Mad Pain and Martian Pain.” In *Readings in the Philosophy of Psychology*, edited by Ned Block, 1: 216–22.

———. 1994. “Reduction of Mind.” In *Companion to the Philosophy of Mind*, edited by Samuel Guttenplan. Blackwell.

Lycan, William G. 1987. *Consciousness*. MIT Press.

McDowell, John. 1979. “Virtue and Reason.” *The Monist* 62 (3): 331–50.

Millikan, Ruth G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

———. 1990. “Compare and Contrast Dretske, Fodor, and Millikan on Teleosemantics.” *Philosophical Topics* 18 (2): 151–61.

———. 1995. “Pushmi-Pullyu Representations.” *Philosophical Perspectives* 9: 185–200.

———. 1996. “On Swampkinds.” *Mind and Language* 11 (1): 103-17.

———. 2002. “Biofunctions: Two Paradigms.” In *Functions*, edited by Andre Ariew. Oxford University Press.

———. 2004. *Varieties of Meaning: The 2002 Jean Nicod Lectures*. MIT Press.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford Clarendon Press.

Neal, David T., Wendy Wood, Mengju Wu, and David Kurlander. 2011. “The Pull of the Past: When Do Habits Persist despite Conflict with Motives?” *Personality & Social Psychology Bulletin* 37 (11): 1428–37.

Neander, Karen. 1995. “Misrepresenting and Malfunctioning.” *Philosophical Studies* 79 (2): 109–41.

———. 1996. “Swampman Meets Swampcow.” *Mind and Language* 11 (1): 118–29.

———. 2013. “Toward an Informational Teleosemantics.” In *Millikan and Her Critics*, edited by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons.

Nelson, Randy J., and Brian C. Trainor. 2007. “Neural Mechanisms of Aggression.” *Nature Reviews Neuroscience* 8 (7): 536–46.

Niv, Yael, Nathaniel D. Daw, Daphna Joel, and Peter Dayan. 2007. “Tonic Dopamine: Opportunity Costs and the Control of Response Vigor.” *Psychopharmacology* 191 (3): 507–20.

Niv, Yael, Peter Dayan, and Daphna Joel. 2006. “The Effects of Motivation on Extensively Trained Behaviour.” *Leibniz Technical Report*, Hebrew University 2006-6.

O’Doherty, John, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. 2004. “Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning.” *Science (New York, N.Y.)* 304 (5669): 452–54.

O’Doherty, J., J. Winston, H. Critchley, D. Perrett, D. M. Burt, and R. J. Dolan. 2003. “Beauty in a Smile: The Role of Medial Orbitofrontal Cortex in Facial Attractiveness.” *Neuropsychologia* 41 (2): 147–55.

Olds, J., and P. Milner. 1954. “Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain.” *Journal of Comparative and Physiological Psychology* 47 (6): 419–27.

Ostlund, Sean B., and Bernard W. Balleine. 2005. “Lesions of Medial Prefrontal Cortex Disrupt the Acquisition but Not the Expression of Goal-Directed Learning.” *The Journal of Neuroscience* 25 (34): 7763–70.

Padoa-Schioppa, Camillo. 2011. “Neurobiology of Economic Choice: A Good-Based Model.” *Annual Review of Neuroscience* 34: 333–59.

Papineau, David. 1984. “Representation and Explanation.” *Philosophy of Science* 51: 550–72.

———. 1993. *Philosophical Naturalism*. Blackwell.

———. 1998. “Teleosemantics and Indeterminacy.” *Australasian Journal of Philosophy* 76 (1): 1–14.

Papineau, David, and Patrick Butlin. forthcoming. “Normal and Addictive Desires.” In *Addiction and Choice: Rethinking the Relationship*, edited by N. Heather and G. Segal. Oxford University Press.

Parkinson, J. A., J. W. Dalley, R. N. Cardinal, A. Bamford, B. Fehnert, G. Lachenal, N. Rudarakanchana, K. M. Halkerston, T. W. Robbins, and B. J. Everitt. 2002. "Nucleus Accumbens Dopamine Depletion Impairs Both Acquisition and Performance of Appetitive Pavlovian Approach Behaviour: Implications for Mesoaccumbens Dopamine Function." *Behavioural Brain Research* 137 (1-2): 149–63.

Peacocke, Christopher. 1992. *A Study of Concepts*. MIT Press.

Pietroski, Paul M. 1992. "Intentionality and Teleological Error." *Pacific Philosophical Quarterly* 73 (3): 267–82.

Plassmann, Hilke, John O'Doherty, and Antonio Rangel. 2007. "Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions." *The Journal of Neuroscience* 27 (37): 9984–88.

Platts, Mark. 1979. *Ways of Meaning*. London: Routledge and Kegan Paul.

Price, Carolyn S. 1998. "Determinate Functions." *Noûs* 32 (1): 54–75.

———. 2001. *Functions in Mind: A Theory of Intentional Content*. Oxford University Press.

Raia, Pasquale, and Shai Meiri. 2006. "The Island Rule in Large Mammals: Paleontology Meets Ecology." *Evolution: International Journal of Organic Evolution* 60 (8): 1731–42.

Rangel, Antonio, and Todd Hare. 2010. "Neural Computations Associated with Goal-Directed Choice." *Current Opinion in Neurobiology* 20 (2): 262–70.

Redgrave, Peter. 2007. "Basal Ganglia." *Scholarpedia* 2 (6): 1825.

Redgrave, Peter, Kevin Gurney, and John Reynolds. 2008. "What Is Reinforced by Phasic Dopamine Signals?" *Brain Research Reviews* 58 (2): 322–39.

Redgrave, Peter, Manuel Rodriguez, Yoland Smith, Maria C. Rodriguez-Oroz, Stephane Lehericy, Hagai Bergman, Yves Agid, Mahlon R. DeLong, and Jose A. Obeso. 2010. "Goal-Directed and Habitual Control in the Basal Ganglia: Implications for Parkinson's Disease." *Nature Reviews Neuroscience* 11 (11): 760–72.

Redgrave, Peter, Tony J. Prescott, and Kevin Gurney. 1999. "The Basal Ganglia: A Vertebrate Solution to the Selection Problem?" *Neuroscience* 89 (4): 1009–23.

Rescorla, Michael. 2013. "Millikan on Honeybee Navigation and Communication." In *Millikan and Her Critics*, edited by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons.

Reynolds, John, and Jeff Wickens. 2002. “Dopamine-Dependent Plasticity of Corticostriatal Synapses.” *Neural Networks* 15 (4-6): 507-21.

Rolls, Edmund T., and Fabian Grabenhorst. 2008. “The Orbitofrontal Cortex and beyond: From Affect to Decision-Making.” *Progress in Neurobiology* 86 (3): 216–44.

Ruff, Christian C. 2011. “A Systems-Neuroscience View of Attention.” In *Attention: Philosophical and Psychological Essays*, edited by Christopher Mole, Declan Smithies, and Wayne Wu. Oxford University Press.

Schoenbaum, Geoffrey, Matthew R. Roesch, Thomas A. Stalnaker, and Yuji K. Takahashi. 2009. “A New Perspective on the Role of the Orbitofrontal Cortex in Adaptive Behaviour.” *Nature Reviews. Neuroscience* 10 (12): 885–92.

Schroeder, Timothy. 2004. *Three Faces of Desire*. Oxford University Press.

Schueler, G. F. 1995. *Desire: Its Role in Practical Reason and the Explanation of Action*. MIT Press.

Schulte, Peter. 2015. “Perceptual Representations: A Teleosemantic Answer to the Breadth-of-Application Problem.” *Biology and Philosophy* 30 (1): 119–36.

Schultz, W. 1998. “Predictive Reward Signal of Dopamine Neurons.” *Journal of Neurophysiology* 80 (1): 1–27.

Schultz, Wolfram. 2007. “Multiple Dopamine Functions at Different Time Courses.” *Annual Review of Neuroscience* 30: 259–88.

Schwitzgebel, Eric. 2002. “A Phenomenal, Dispositional Account of Belief.” *Noûs* 36 (2): 249–75.

Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Sehon, Scott R. 2005. *Teleological Realism: Mind, Agency, and Explanation*. MIT Press.

Sescousse, Guillaume, Jérôme Redouté, and Jean-Claude Dreher. 2010. “The Architecture of Reward Value Coding in the Human Orbitofrontal Cortex.” *The Journal of Neuroscience* 30 (39): 13095–104.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford University Press.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3): 379–423.

Sharot, Tali, Tamara Shiner, Annemarie C. Brown, Judy Fan, and Raymond J. Dolan. 2009. “Dopamine Enhances Expectation of Pleasure in Humans.” *Current Biology* 19 (24): 2077–80.

Shea, Nicholas. 2007. “Consumers Need Information: Supplementing Teleosemantics with an Input Condition.” *Philosophy and Phenomenological Research* 75 (2): 404–35.

———. 2013. “Millikan’s Isomorphism Requirement.” In *Millikan and Her Critics*, edited by Dan Ryder, Justine Kingsbury, and Kenneth Williford. Wiley-Blackwell.

———. 2014. “Reward Prediction Error Signals Are Meta-Representational.” *Noûs* 48 (2): 314–41.

———. 2015. “Distinguishing Top-Down From Bottom-Up Effects.” In *Perception and Its Modalities*, edited by D. Stokes, M. Matthen, and S. Biggs. Oxford University Press.

Sinhababu, Neil. 2009. “The Humean Theory of Motivation Reformulated and Defended.” *Philosophical Review* 118 (4): 465–500.

———. 2015. “Advantages of Propositionalism.” *Pacific Philosophical Quarterly* 96 (1): 165–80.

Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Smith, Michael. 1987. “The Humean Theory of Motivation.” *Mind* 96 (381): 36–61.

———. 1994. *The Moral Problem*. Blackwell.

———. 2011. “Humeanism about Motivation.” In *A Companion to the Philosophy of Action*, edited by Timothy O’Connor and Constantine Sandis. Wiley-Blackwell.

Sobel, David, and David Copp. 2001. “Against Direction of Fit Accounts of Belief and Desire.” *Analysis* 61 (1): 44–53.

Sober, Elliott. 1984. *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. University of Chicago Press.

Sterelny, Kim. 2003. *Thought in a Hostile World*. Blackwell.

Stocker, Michael. 1979. “Desiring the Bad: An Essay in Moral Psychology.” *Journal of Philosophy* 76 (12): 738–53.

Strawson, Galen. 1994. *Mental Reality*. MIT Press.

Strawson, Peter F. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48: 1–25.

Sulzer, James, Ranganatha Sitaram, Maria Laura Blefari, Spyros Kollias, Niels Birbaumer, Klaas Enno Stephan, Andreas Luft, and Roger Gassert. 2013. “Neurofeedback-Mediated Self-Regulation of the Dopaminergic Midbrain.” *NeuroImage* 83 (December): 817–25.

Sutton, Peter. 2013. *Vagueness, Communication and Semantic Information*. King’s College London PhD Thesis.

Sutton, Richard and Andrew Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Tanaka, Saori C., Bernard W. Balleine, and John P. O’Doherty. 2008. “Calculating Consequences: Brain Systems That Encode the Causal Effects of Actions.” *The Journal of Neuroscience* 28 (26): 6750–55.

Thorndike, Edward Lee. 1905. *The Elements of Psychology*. A. G. Seiler.

Tindell, Amy J., Kyle S. Smith, Kent C. Berridge, and J. Wayne Aldridge. 2009. “Dynamic Computation of Incentive Salience: ‘Wanting’ What Was Never ‘Liked.’” *The Journal of Neuroscience* 29 (39): 12220–28.

Tolman, E. C. 1949. “There Is More than One Kind of Learning.” *Psychological Review* 56 (3): 144–55.

Tricomi, Elizabeth, Bernard W. Balleine, and John P. O’Doherty. 2009. “A Specific Role for Posterior Dorsolateral Striatum in Human Habit Learning.” *The European Journal of Neuroscience* 29 (11): 2225–32.

van den Bos, Ruud, Susanne Koot, and Leonie de Visser. 2014. “A Rodent Version of the Iowa Gambling Task: 7 Years of Progress.” *Frontiers in Psychology* 5: 203.

Velleman, J. David. 1992. “The Guise of the Good.” *Noûs* 26 (1): 3–26.

Wassum, K. M., S. B. Ostlund, N. T. Maidment, and B. W. Balleine. 2009. “Distinct Opioid Circuits Determine the Palatability and the Desirability of Rewarding Events.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (30): 12512–17.

Williams, Bernard. 1979. “Internal and External Reasons.” In *Rational Action*, edited by Ross Harrison, 101–13. Cambridge University Press.

Wise, Roy A. 2004. “Dopamine, Learning and Motivation.” *Nature Reviews. Neuroscience* 5 (6): 483–94.

Wise, Steven P. 2008. “Forward Frontal Fields: Phylogeny and Fundamental Function.” *Trends in Neurosciences* 31 (12): 599–608.

Wunderlich, Klaus, Antonio Rangel, and John P. O’Doherty. 2010. “Economic Choices Can Be Made Using Only Stimulus Values.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (34): 15005–10.

Yin, Henry H., Barbara J. Knowlton, and Bernard W. Balleine. 2004. “Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning.” *The European Journal of Neuroscience* 19 (1): 181–89.

———. 2006. “Inactivation of Dorsolateral Striatum Enhances Sensitivity to Changes in the Action-Outcome Contingency in Instrumental Conditioning.” *Behavioural Brain Research* 166 (2): 189–96.

Yin, Henry H., Sean B. Ostlund, Barbara J. Knowlton, and Bernard W. Balleine. 2005. “The Role of the Dorsomedial Striatum in Instrumental Conditioning.” *The European Journal of Neuroscience* 22 (2): 513–23.

Yin, Henry H., Xiaoxi Zhuang, and Bernard W. Balleine. 2006. “Instrumental Learning in Hyperdopaminergic Mice.” *Neurobiology of Learning and Memory* 85 (3): 283–88.

Zhou, Zhonghe. 2014. “Dinosaur Evolution: Feathers up for Selection.” *Current Biology* 24 (16): R751–53.