4-13-2016 12:00 AM

A Pure Representationalist Account of Belief and Desire

Stephen Pearce
The University of Western Ontario

Supervisor
Angela Mendelovici
The University of Western Ontario

Graduate Program in Philosophy
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

1 Abstract

According to the traditional view, beliefs and desires are mental representations that play particular functional roles. A belief that P is state which represents P and plays the belief-role, while a desire that P is a state that represents that P and plays the desire-role. In this dissertation I argue that the traditional view has trouble accounting for (a) the role that belief and desire play in the causal and rational explanation of behaviour and (b) our knowledge of our own conscious, occurrent beliefs and desires. In its place I argue for *Pure Attitude Representationalism (PAR)*, which holds that beliefs and desires are not to be distinguished from one another by their functional role, but instead entirely by their representational content. According to PAR, desires are distinct from beliefs not because they do different things, but because desires represent things as being rewarding. Throughout the three papers that comprise this dissertation I argue that PAR provides a better explanation of (a) and (b).

Keywords

Attitudes, Belief, Desire, Representation, Functionalism, Self-Knowledge, Transparency

Contents

2 Introduction

We hold various attitudes towards the world. We believe that it is going to rain tomorrow; we desire that our favourite sports team wins the championship. This thesis is a philosophical investigation into the nature of such mental states, also known as "the attitudes". In particular, I am interested in what makes an attitude the type of attitude that it is (e.g. a belief rather than a desire).

According to the dominant view, attitudes have two components. The first component is their representational content. A belief that it is going to rain tomorrow represents that it is going to rain tomorrow. The second component is the attitude type: e.g. belief and desire. It is this second component that distinguishes a belief that it is going to rain tomorrow from a desire that it is going to rain tomorrow. Traditionally, philosophers have been primarily interested in the representational content of mental states. As a result, the second component of the attitudes has been neglected. In this thesis I investigate this second component. In particular, I am interested in the question of what it is for a given mental state to be the kind of attitude it is: what it is to be a belief rather than a desire.

The view that I oppose is that attitude types should be given a functionalist analysis: that what makes a given mental state a belief, rather than a desire, for instance, is how it is causally related to other mental states, sensory inputs and behavioural outputs (e.g. Fodor, 1975; Loar, 1981; Searle, 1983). Because of how little attention philosophers have paid to attitude type, this functionalist view is largely taken for granted. In my thesis I will develop an alternative view. According to my proposal, attitude types are representational in nature, just like attitude contents. For example, the difference between a belief and a desire is a difference in their representational content. What an attitude represents determines what kind of attitude it is. I call this view *Pure Attitude Representationalism*.

I will talk of mental states being "representational" or "intentional", or of mental states having representational or intentional properties. Understood in the broadest sense, representational properties are semantic properties. A mental state is said to be representational if and only if various semantic properties can be ascribed to it. Examples of semantic properties are: reference,

truth-conditions, truth-value. When we speak of a perceptual experience being illusory, for example, we are ascribing semantic properties to it, and are thus committed to perceptual experiences being representational.

Representational mental states take the world to be a certain way. The perceptual experience I am having right now takes there to be a laptop such-and-such a distance in front of me, a coffee cup next to it, and so on. How a representational mental state takes the world to be is called its 'content'. The content of a mental state (at least partially) determines the conditions under which it is true or veridical. My perceptual experience is veridical just in case there is such a laptop and coffee cup in front of me, and this is because it has the representational content that it does.

I intend talk of mental representation to be neutral on a host of theoretical issues. There are many different views about what kind of contents mental representations have. Some think that mental states have Russellian contents, which are structured entities composed of objects, properties and relations. Others think that mental states have Fregean contents, which are structured entities composed of modes of presentations of objects, properties and relations. Still others reject that contents are structured entities, and instead think that the content of a mental state is a set of possible worlds.

There are also many views about the metaphysics of representational content. In virtue of what does a given mental state have some representational content? Arguably the most popular view about this holds that a mental state M has content C just in case it carries information about C. The notion of 'information carrying' is then spelled out in terms of some naturalistically acceptable relation between M and C. For example, causal theories of content hold that M has content C just in case there is a nomological relation between C and tokenings of M (Fodor 1987). Teleological theories of content hold instead that M has content C just in case it is the evolutionary function of M to indicate C (Millikan 1984). Other theories deny that you can reduce mental representation to a naturalistic relation such as causation. The Phenomenal Intentionality Thesis instead tries to understand mental representation in terms of the phenomenal character of a mental state — the 'what it's like' of that mental state (Mendelovici and Bourget 2014).

Finally, some philosophers think that mental representations have a certain language-like structure. Fodor (1975), for example, holds that mental representations are sentences in a language of thought. These sentences, like those in natural languages, have compositional structure: the meaning of the sentence is a function of the meaning of the parts. The parts of these sentences are concepts.

There are thus many theoretical positions to take with respect to mental representation. Philosophers differ on what they take contents to be composed of; on the metaphysics of mental representation; and on whether mental representations have language-like structure. I mention the different theoretical positions that one can take with respect to mental representation in order to set them aside. I wish my account of the attitudes to be neutral with respect to these theories.

It is common to distinguish between dispositional and occurrent attitudes. An occurrent belief is one that you are currently tokening, one that is playing an active role in your mental life at the moment. In contrast, a dispositional belief does not entail the existence of any occurrent mental states. It is natural to say that you believed that 89 is larger than 5 before ever entertaining that proposition. Likewise, it natural to say that someone in graduate school for philosophy has a desire to become an academic philosopher, even when she is not actively thinking about philosophy. These are dispositional (or 'standing') attitudes. Unless stated otherwise, all talk of attitudes will exclusively be about occurrent attitudes, not dispositional attitudes. I will remain neutral with respect to the question of the relationship between occurrent and dispositional attitudes. I suspect that occurrent attitudes are more theoretically important because, in most cases, to ascribe to S dispositional belief that P, for example, is just to ascribe to S a disposition to have an occurrent belief that P. I am not committed to this claim, however, and nothing that I say herein will hinge on it.

This thesis consists in three papers. In the first, *What are Beliefs and Desires*, I argue that the traditional functionalist view of belief and desire is unable to account for their differing roles in the causal and rational explanation of behaviour. I argue that the best way to account for the fact that belief and desire play these different role is in terms of their having different representational contents. In the second, *Attitudes and Self-Knowledge,* I argue that while it is plausible that we have privileged access to our conscious, occurrent beliefs and desires, the

traditional functionalist picture makes it hard to account for this. I then argue that the best way to account for the privileged access that we have to our conscious, occurrent beliefs and desires is by rejecting this traditional picture, and instead adopt the view that belief and desire have different contents. In the last paper, *Reward and the Transparency of Desire*, I argue that desire is transparent in the following sense: that we can come to know what we desire by attending to the world. I argue that desires are transparent because we come to know what we desire by directing attention to the content of reward representations. These representations are posited by neuroscientific literature on goal-directed learning. I conclude by arguing that the simplest explanation of the phenomenon of desire transparency is that desires just are reward representations.

The central theme that runs throughout these three papers is the rejection of the traditional, functionalist picture of belief and desire. In each paper I raise an explanatory problem that the traditional view faces, and argue that PAR provides a better explanation in each case.

2.1 Bibliography

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

Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind, MIT Press.

Loar, B. (1981). Mind and Meaning, Cambridge University Press.

Mendelovici, A. and D. Bourget (2014). "Naturalizing Intentionality: Tracking Theories Versus Phenomenal Intentionality Theories." Philosophy Compass **9**(5): 325-337.

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

Searle, J. R. (1983). Intentionality: An Essay in the Philosophy of Mind, Cambridge University Press.

3 What are Beliefs and Desires?

3.1 Introduction

They are propositional attitudes, of course. They are attitudes that we bear to propositions. To believe that you are late for work is to bear the belief relation to the proposition <I am late for work>. To desire that you get the job is to bear the desire relation to the proposition <I get the job>. Call the proposition that you are related to the propositional content, and the type of relation you bear to it the attitude type.

Two questions arise:

1. What is it to be related to a proposition in general?
2. What is it to bear a certain type of attitudinal relation to a proposition?

I am going to assume that the answer to the first question is: to be in an inner state which represents it. To be related to the proposition <I am late for work> is to be in an inner state which represents that you are late for work. On this view, the propositional content of your attitude is the content of a mental representation. To answer the first question in this sort of way is to be a representationalist about propositional attitudes[1].

The representationalist tends to answer the second question by invoking functional roles. Different types of attitudinal relations correspond to the different sorts of functional roles of the state doing the representing. To believe that you are late for work is to be in a mental state which (a) represents that you are late for work and (b) has the functional role characteristic of belief ("plays the belief-role" or "is in the belief-box"). To desire that you get the job is to be in a mental state which (a) represents that you get the job and (b) has the functional role characteristic of desire ("plays the desire-role" or "is in the desire-box"). On this view, attitude types are functional roles. Call this view attitude functionalism (AF).

---

[1] Representationalists include Fodor (1987), Millikan (1984) and Dreske (1988)

The standard view thus understands attitudes to be internal mental representations that play particular functional roles. It understands attitude type and propositional content very differently. The propositional content of an attitude is what it mentally represents, while its attitude type is what it (typically) does.

In this paper I will accept representationalism, but reject attitude functionalism. In its place, I will argue for what I call pure attitude representationalism (PAR). On this view, attitude types, like propositional contents, are also mental representational properties. Beliefs and desires are distinguished from one another on the basis of how they represent things as being. My strategy in this paper will be to argue that PAR provides the best account of the attitudes' role in common sense explanation of human behaviour.

3.2 Folk Psychology

Folk psychology is a common sense theory used for explaining and predicting behaviour. You tell your friend to meet you in your office at noon the next day; unsurprisingly, the next day at noon your friend shows up at your office. Why? Because he believed that to meet you he must go to your office at noon, and he desired to meet you. Or he believed that you promised to be there, and you desired to not break your promise.

We all explain and predict behaviour in terms of attitudes such as belief and desire, and we do so with startling accuracy and ease. How do we do this? One view is that we possess a certain theory – folk psychology – that we employ in order to explain and predict behaviour. This theory contains certain generalizations that constitute its (*ceteris paribus*) laws. If someone desires P and believes that O leads to P, they will intend to O. If someone believes that a is F, they will believe that something is F. If you believe that P is dangerous you will fear P. These statements, and others like them, are the generalizations of common sense folk psychology.

As Fodor (1978/1987) argues, the explanatory power and predictive success of common sense folk psychology cannot be an accident. There must be something about how we work that explains the success of common sense folk psychology. The most natural way to do this is take such generalizations to be literally true. The entities that they quantify over – beliefs, desires, and other attitudes – are actual mental states we possess, and its generalizations express causal laws

that obtain between these mental states. This is a major motivation for thinking that the attitudes exist. We need to take folk psychology seriously, and to do this we must acknowledge that people really do have beliefs and desires.

Fodor argues that the best explanation of the success of folk psychology leads to a representationalist picture of the attitudes, according to which beliefs and desires involve mental representations. The first step is to note that many of the generalizations of common sense folk psychology exhibit systematic relationships between the propositions referred to by that-clauses: their "propositional contents". For example, if you believe that John and Mary are dating, then you will believe that Mary is dating someone. Here, "that John and Mary are dating" and "that Mary is dating someone" pick out propositions that are logically related: one entails the other. Or consider the following generalization: if you believe that if it rains then the ground will be wet, and you believe that it is raining, then you will believe that the ground is wet. Again, this generalization exhibits semantic relations between the propositional contents of the attitudes in question.

It is because of this, stresses Fodor, that describing the causal train of one's thoughts often looks like an argument. The causal flow from one attitude to another looks just like the flow of support from premises to the conclusion in an argument. So the first step towards an argument for representationalism is to note that a vast amount of the generalizations of common sense folk psychology exhibit semantic relations between propositional contents.

The second step is to claim that the best explanation of these generalizations is (1) that attitudes are mental representations, the contents of which just are their propositional contents, and (2) that the causal relations between the attitudes somehow mirrors the relations between their mental contents. Thus, the generalization above – that if you believe that John and Mary are dating, you will believe that Mary is dating someone – is not taken to be a brute fact about the attitudes. Rather, it is explained by the fact that attitudes have mental contents, and that the causal powers of the attitudes somehow respect relations between their contents. It is because your belief represents that John and Mary are dating that it tends to cause you to have a belief that represents that Mary is dating someone.

So, according to Fodor, we should be representationalists about attitudes because it provides a simple and compelling explanation of the predictive success of many generalizations of folk psychology. That is, the predictive success of these generalizations is due to the fact that they express causal relations between attitudes that obtain in virtue of logical or semantic relations between their propositional contents. This explains why some generalizations hold, but others don't. For example, compare:

> (L) If you believe that John and Mary are dating, then you will believe that Mary is dating someone.
>
> (L*) If you believe that John and Mary are dating, then you will believe that Mary enjoys reading Russian novels.

Representationalism explains why (L) holds but (L*) does not: the former exhibits logical relations between propositional contents, while the latter does not. Further, it explains a host of facts about the generalizations of folk psychology by appealing to a single phenomenon: logical relations between their propositional contents.

However, it does not explain everything. Compare the following generalizations:

> (BD) If you desire that you get the job, and believe that if you bribe the recruiter then you will get the job, then you will bribe the recruiter.
>
> (BB) If you believe that you will get the job, and believe that if you bribe the recruiter then you will get the job, then you will bribe the recruiter.

Likewise, compare (L) with the following:

> (LD) If you desire that John and Mary are dating, then you will believe that Mary is dating someone.

While (BD) and (L) hold, (BB) and (LD) do not, even though the attitudes exhibit the same logical relations between their propositional contents. The difference is in attitude type: (BD) contains a belief-desire pair with related propositional contents, while (BB) contains a belief-belief pair. Likewise, (L) goes from a belief to belief, while (LD) goes from desire to belief.

Mere representationalism does not have resources to explain why (BD) and (L) hold, but (BB) and (LD) do not. What is needed is an account of the difference between belief and desire that explains their different sorts of causal powers. Attitude Functionalism (AF) takes these causal differences between belief and desire to be part of the essential functional role of belief and desire. A desire is the sort of state which, among other things, makes the following schematic generalization true:

(D1) If you ___ that P, and believe that if you O then P, then you will O.

While a belief is the sort of state which, among other things, makes the following sorts of schematic generalizations true:

(B1) If you ___ that P, and ___ that P->Q, then you ___ that Q.

(B2) If you ___ that P, then you will tend to assert that P.

According to AF, then, we do not explain why certain generalizations hold but others do not in terms of the difference between beliefs and desires. Rather, we explain the difference between beliefs and desires in terms of which generalizations they participate in. But this is backwards. We want to explain behaviour by attributing attitudes to people. I explain your taking an umbrella with you by attributing to you a desire to not get wet. That is, I take your having a desire to not get wet to explain why you took an umbrella with you. There is something about desiring to not get wet (as opposed to believing that you are not wet) that explains your behaviour. The attitude functionalist cannot make sense of this. The only explanation she can give of my taking an umbrella is that I am in a state for which the following schematic generalization is true:

(U) If you ___ to not get wet, and you believe that an umbrella will prevent you from getting wet, the you will take an umbrella.

In other words, the attitude functionalist explains my umbrella taking behaviour by ascribing to me a state for which a generalization that covers my umbrella taking behaviour is true. And that

is all the attitude functionalist can say: they cannot say anything about why that generalization is true of the state that I am in[2].

For comparison, suppose that you ask me why a substance A dissolved in water, and I respond by attributing to A a state S for which the following generalization holds:

If a substance with state ___ is put into water, then it will dissolve.

In other words, I attempt to explain the fact that A dissolved in water by attributing to it a state for which a generalization that covers its dissolving behaviour is true. Intuitively, this would only be explanatory if I gave you a further explanation of why that generalization is true of S. But I have not done this. Likewise, the attitude functionalist does not offer an explanation of why (U) is true of desires.

Or, recall the commonly used example from Molière's *Le Malade Imaginaire*. When asked to explain why opium has caused someone to fall asleep, the philosopher responds: "Because of its dormative virtue!" This is a poor explanation, because to have dormative virtue is to just to be disposed to cause people to fall asleep. The opium cased someone to fall asleep because that is what it tends to do. But what we want is an explanation of *why* the opium tends to cause people to fall asleep.

The predictive power of the generalizations of folk psychology cries out for explanation. The best account of the attitudes it the one that best explains why certain generalizations hold but others do not. Representationalism goes some way towards this goal, but leaves many things unexplained. In particular, what is left unexplained is why certain generalizations are true of one attitude type but not others. According to attitude functionalism, these differences are taken to be constitutive of attitude types themselves. Beliefs and desires are different in virtue of the

---

[2] There are two related — yet distinct — problems with functionalist explanation that do not concern me here. The *causal exclusion* problem is that functional properties are epiphenomenal since they are realized in physical properties which are causally sufficient for their effects (Kim 1998, pp. 77-87). The problem of *metaphysically necessary effects* is that functional properties can't be said to cause the effects they are necessarily connected with, since causes must (according to a popular view of causation) be metaphysically separate from their effects. These problems do not concern me (Block, 1989). Let us grant that functional properties really do have causal powers. There is still a further question of whether they are *explanatory*.

different generalizations they appear in. But what we need is an explanation in the other direction. What is it about beliefs and desires that explains their differing roles in the generalizations of folk psychology?

3.3 What's wrong with functional explanations?

It was argued above that Attitude Functionalism is not able to provide an adequate explanation of certain of the generalizations of folk psychology. Attitude functionalism explains why people act in the way that they do by attributing to them states that are individuated partially in terms of their being disposed to cause people to act in those ways. However, one might take this to be a perfectly acceptable explanatory tactic. After all, we typically explain someone's wincing behaviour by saying that they are in pain, and isn't the property of being a pain partially constituted by it disposing you to engage in wincing behaviour?

Now, the charge against such explanations cannot be that they are empty. We learn many things about someone when their wincing is explained in terms of their being in pain (Bradley 2013, pg. 9). For example: that this is the typical effect of the state that they are in, which rules out the possibility that the wincing was a freak accident. We are also able to rule out the possibility that the wincing was caused by something external to the agent. Finally, pain is not characterized solely by its being disposed to cause wincing behaviour. It also tends to be caused by bodily damage, and tends to cause avoidance behaviour, and so on. Knowing this, we are able to predict that the wincing person has likely suffered bodily damage, and perhaps might engage in avoidance behaviour.

Explaining someone's wincing behaviour in terms of the attribution of a state that is partially characterized by its tendency to cause wincing behaviour, then, is not empty. And we can say similar things about the Attitude Functionalist's explanation of behaviour in terms of beliefs and desires. Take the example of explaining Jill's taking an umbrella with her in terms of her desire to not get wet. According to the Attitude Functionalist, this state is partially characterized by its

tendency to cause her to take an umbrella[3]. This explanation rules out the possibility that she took the umbrella because she mistook it for a sword, or that she was forced to take an umbrella by an evil scientist controlling her motor cortex. It also allows us to make certain predictions about other aspects of her behaviour, such as her potentially cancelling the picnic that she had planned for that afternoon.

However, while explanations like these are not empty, there is a clear sense in which something important is missing. A folk psychological explanation of someone's behaviour is supposed to *make sense* of that behaviour: that is, show how it was *rational* from the point of view of the agent. This is what Representationalism does for a certain subset of folk psychological generalizations. Why does Chuck believe that tomorrow is Wednesday? Well, it's because he believes that today is Tuesday, and that Wednesday comes after Tuesday. Such an explanation not only tells us what caused him to believe that tomorrow is Wednesday, but also why it is *rational* for him to have that belief. Similarly, when we explain someone's wincing behaviour in terms of their being in pain, we have a certain positive conception of the nature of pain, independent of its functional role, that makes sense of it causing wincing. Pain hurts, and that's why it makes us act the way it does.

Folk psychological explanations, then, have to play a dual role. Not only do they tell a story about what caused an agent to behave in a certain way or hold certain attitudes, but they also demonstrate why these effects are rational from the perspective of the agent. This is precisely what Representationalism does for certain of the causal generalizations of folk psychology. If you know of Jill that she believes that Mary loves John, then we can infer that she believes that Mary loves someone. The Representationalist accounts for the causal connection between these beliefs in terms of the semantic connection between their representational contents. As we saw in the section above, this allows for a compact account of many such causal generalizations. But it also allows us to make sense of why the former belief would lead Jill to have the latter belief.

---

[3] As long as she has the right sort of other attitudes, such as a belief that it is raining, a belief that taking an umbrella is an effective means of preventing herself from getting wet, etc.

Jill's belief that Mary loves someone is rendered rational in light of her belief that Mary loves John.

A theory of the attitudes must meet two desiderata. First, it must explain why certain causal generalizations hold, but not others. Second, it must do so in such a way as to rationalize them. The power of Representationalism is that it is able to meet these two desiderata for a certain subset of these causal generalizations, by appealing to a single phenomenon: the semantic content of the attitudes. But as we saw, mere Representationalism cannot meet these desiderata for those causal generalizations that crucially involve the differences between the attitude types.

What is needed is an account of attitude types that explains why they play different roles in the causal generalizations of folk psychology in such a way as to rationalize them as well. Let us take a closer look, then, at the roles that the different attitude types play in practical reason.

3.4 Practical Reason

Beliefs not only cause us to have other beliefs; they also give us reason to have other beliefs. And desires not only cause action; they also give us reason to act. In this section I argue that the different roles that belief and desire play in practical reason is best explained by holding that beliefs and desires have different representational contents.

We saw above that one of the generalizations of folk psychology is that if I believe that John is dating Mary, then I will believe that Mary is dating someone. This is a causal generalization. But not only does the former belief tend to cause me to form the latter belief, it also constitutes a reason to have the latter belief. And the explanation is the same: the propositional contents of the former beliefs entail the propositional content of the latter. Beliefs, then, give us reason to form beliefs with appropriately related propositional contents. Desires, in contrast, do not. My desire that John and Mary date each other is not a reason to believe that John is dating someone. My desire that I get the job is a reason for me to act in such a way as to make it more likely that I get the job. My desire for a slice of chocolate cake is a reason for me to eat a slice of chocolate cake.

The differences in the causal roles of beliefs and desires, then, is mirrored in their rational roles. It is the role of beliefs to both cause us to have and give reasons for other beliefs, while it is the

role of desires to both cause and give reasons for certain kinds of behaviour[4]. The Attitude Functionalist attempts to explain this by bundling these explanatory roles together and taking them to constitute the difference between belief and desire. But this explanation is incomplete. It does not help us to understand what it is about desiring a slice of chocolate cake that it could give me reason to eat a slice of chocolate cake. It merely says that to be in such a state it so be disposed to eat a slice of chocolate cake.

But what more to desire is there? In his 1987, Dennis Stampe considers the question of how desires provide reasons for action. In the case of beliefs, we reason from their objects — what we've been calling their propositional contents. According to Stampe, beliefs provide reasons in virtue of representing facts. My belief that John loves Mary represents the fact that John loves Mary, and it is this fact that constitutes my reason for believing that John loves someone. In contrast, when I desire to eat a slice of chocolate cake, I my eating a slice of chocolate cake is not a fact. So desires cannot give provide reasons in virtue of representing facts. Rather, according to Stampe, desires provide reasons merely in virtue of having them. I have a reason to eat a slice of chocolate cake simply because I desire it. This is what Stampe calls this the authority of desire. A desire that P gives us reason to act as to make it the case that P, merely by having the desire.

To sum up so far: According to Stampe, beliefs give us reasons because their propositional contents are facts. The authority of belief is thus derivative on the authority of facts. Desires, in contrast, do not give us reasons through their propositional contents. The authority of desire is not derivative. The act of desiring is itself a reason-giving force. But how? Stampe considers three such views.

3.4.1 Desires as Unpleasant Sensations

The first view that Stampe considers is that desires are like itches. An itch in your foot gives us reason to scratch it. Similarly, perhaps a desire to eat a slice of chocolate cake is an itch that only

[4] This is an oversimplification. Beliefs and desires, except in restricted cases, do not cause or give reasons for anything in isolation. Desires only cause/give reason for behaviour when combined with the appropriate beliefs.

eating a slice of chocolate cake will scratch. Desires get their reason-giving force because they are unpleasant, and since satisfying them is the only way to get rid of them, they give us reason to satisfy them.

Unfortunately, according to Stampe, the analogy between desires and unpleasant sensations breaks down. An unpleasant sensation gives us reason to get rid of the sensation in any way possible. Suppose you have a headache from dehydration. One way of getting rid of the headache is to drink a lot of water. Another is to take a lot of painkillers. While one of these methods is more advisable all things considered, the pain itself gives you equal reason to do either of them. All that matters is that you no longer have the pain, and both of those are effective means to achieve that.

In contrast, a desire on its own intuitively does not give you equal reason to do anything that would get rid it. Suppose that you desire to have a child, and that there is a pill that you can take that will get rid of your desire for a child. There are then two ways of getting rid of your desire for a child: having a child, or taking the pill. Your desire for a child alone, however, only gives you reason to do the first. You might have reason to do the second, but only as a result of further desires (such as the desire to not want to desire to have a child[5].) The reason-giving force of desires, according to Stampe, cannot be explained on the model of that of unpleasant sensations. What desires give you reason to do essentially involves their propositional content: that is, they give you reason to act as to make them true.

3.4.2 Desire Satisfaction as Intrinsically Good

The second view that Stampe considers is that desires give us to reason to satisfy them because satisfying desires is objectively good. My desire to eat cake makes it the case that my eating cake is an objectively good state of affairs. Thus the reason-giving force of desires is to be understood on the model of the reason-giving force of good things in general. If X is a good thing to do, then

[5] The apparent analogy between desires and pains, I think, is due to the fact that this is a common scenario: wanting P, but also wanting to not want P. In such a scenario, you do have reason to either get P or get rid of your desire for P. This is not the case in the scenario where you lack the relevant second order desire.

we have reason to X. And since satisfying our desires is a good thing to do, then we have reason to satisfy them.

The problem that with this account is that we can easily imagine someone desiring a state of affairs that is obviously not good. As Stampe points out, we can imagine a creature with desire malformed to the extent that "there would be nothing good about its desires being satisfied." This account gets the relation between goodness and desire the wrong way around. States of affairs aren't good because we desire them; we desire certain states of affairs (at least partially) because they are good (or at least we think so.)

Another problem that Stampe fails to notice is that this view is only able to account for the "objective" reasons that desires provide us with. An objective reason is a reason we have whether or not we realize it. I have reason to cancel my appointment with my hairdresser because my hairdresser is, unbeknownst to me, a serial killer. However, since I am not aware of this, I have no "subjective" reason, or reason "from the inside." From my point of view, I have no reason to cancel my hair appointment; from the view from nowhere, I do.

The claim that desires give us reason to satisfy them because satisfying them is objectively good is a claim about objective reasons. According to this account, my desire to eat chocolate cake gives me reason to do so even if I am ignorant about why it gives me that reason — that is, if I am so unfortunate as to have never read the philosophical literature on desire. Whether or not desires provide us with such objective reasons, they certainly supply us with subjective reasons. My desire to eat chocolate cake provides me with a reason to eat chocolate cake, a reason that I can grasp.

3.4.3 Desires as Representations of Goodness

The previous account located the reason-giving force of desires in the objective goodness of the state of affairs that would obtain were the desire to be satisfied. But we can clearly have desires for states of affairs that (a) would in fact not be objectively good or (b) would be objectively good, but we are not aware of this. In case (a), according to the previous account, desires do not provide us with reasons at all, and in case (b) desires provide us with only objective reasons. But clearly we have subjective reasons in both cases (a) and (b). A desire for an intrinsically bad state

of affairs — for example, the diabetic's desire for sweets — still provide us with subjective reasons to satisfy them. As do desires for states of affairs that we have been led to believe are objectively bad.

So, in virtue of what do desires provide us with subjective reasons to act as to satisfy them? If objective reasons involve how things are, then subjective reasons involve how things appear to us. I might have an objective reason to believe that my partner is angry at me because of their facial expressions; but if I do not notice these facial expressions for what they are, then I do not have a subjective reason to believe that my partner is angry at me. Likewise, if a desire is to provide us with subjective reasons to act as to satisfy their propositional content, then they must make this content appear to us in a certain way. The final view that Stampe considers, then, is that what is distinctive about desires is the way they make their propositional contents appear to us.

Desires provide us with reasons even when the states of affairs that would obtain if they were to be satisfied are not good, because desires present such states of affairs as being good. Desires present their objects to us as being good, and it is in virtue of this that they provide us with reasons to satisfy them. On this view, a desire that P is simply a representation that P is good. Call this the Desires as Representations of Goodness view (DRG).

In sum, if we are to accept that desires provide us with subjective reasons to act as to make their propositional contents true, then desires must present these contents to us as being a certain way, a way that renders them reasonable to pursue. The lesson of this section, then, is that the best way to account for the distinctive rational roles desire plays is in terms of the distinctive way that desires present their propositional contents to us: as being *good*.

3.5 Desires and Beliefs about Goodness

Can't one believe that something is good, and yet fail to desire it? Or desire something, and fail to believe that it is good? If a desire that P is simply a representation that P is good, then what distinguishes it from a belief that P is good? Clearly, DRG as stated is incomplete. Something must be added to the account in order to distinguish desires from beliefs about goodness. In this section, I will outline three ways that DRG can be fleshed out.

3.5.1 Beliefs Have Extra Content

The first way to distinguish a mere representation of goodness from a belief about goodness is to hold that beliefs about goodness do not merely represent that something is good. Rather, on this view, beliefs have their own distinctive way that they present their propositional contents. One possibility is that beliefs present their propositional contents as being *facts*.

Recall that, according to Stampe, the reason-giving force of beliefs is derivative on the reason-giving force of facts. My belief that it is going to rain tomorrow gives me a reason to bring an umbrella because the *fact* that it is going to rain tomorrow provides such a reason. However, beliefs provide us with reasons even when they are false, that is, even when their propositional contents are not facts. And when a belief does represent a fact, it provides us with reasons even when we are not aware of this. My belief that it is going to rain tomorrow gives me a subjective reason to bring an umbrella even when I do not know that this is a fact.

So, *pace* Stampe, the reason-giving force of beliefs is not derivative on the reason-giving force of facts. Rather, one might argue that if beliefs provide us with subjective reasons to act *as if* their propositional contents are facts, this can only be because belief *present* their contents *as being* facts. To believe something is not to merely represent it; it is to represent it as being a fact. It is this representational content that is able to provide us with reasons for action, whether or not the propositional content is really a fact. Thus, we locate the reason-giving force of a belief not in its propositional content, but in what it says about that propositional content. My belief that John loves Mary tells me that it is a fact that John loves Mary. Whether or not this is true is irrelevant to its reason-giving force.

A potential worry here is that the content <John loves Mary> and the content <it is a fact that John loves Mary> are identical. This is true on the view that contents are just sets of possible worlds: the set of worlds where John loves Mary is identical to the set of worlds where it is a fact that John loves Mary. But these contents are not identical according to more fine-grained views of content. For example, if contents are structured entities that are composed of objects, properties and relations, then <it is a fact that John loves Mary> will have constituents that <John loves Mary> does not.

One might object that, even on fine-grained views of content, <it is a fact that John loves Mary> and <John loves Mary> have the same constituents if one is *deflationist* about fact-talk. On this view, when I say 'it is a fact that P', I have said nothing more than simply 'P': 'it is a fact that' does not contribute anything to the meaning of the utterance. However, what concerns us is the content of *mental representations*, not the content of linguistic utterances. It is one thing to be a deflationist about fact-talk; it is quite another to be a deflationist about fact-*thought*. Indeed, if the content of my belief that P really is <it is a fact that P>, and if assertions express beliefs, then this might *explain* why it is redundant to add 'it is a fact that' to my assertion that 'P'.

Another objection goes as follows: if a belief that P is just a representation that P is a fact, then it seems that I am free to believe whatever I like. All I have to do is think to myself: it is a fact that the moon is made of cheese, and I come to believe this. But, the objection continues, I don't believe this, even when having that thought. So, beliefs must not just be representations of facts.

One response to this objection is to accept the premise that one is able to think of things being facts without believing them, and take this to mean that being a fact is not the content distinctive of beliefs. It might be that the content distinctive of beliefs is not expressible, and that "being a fact" is the closest approximation to it. Further, it might be that one does not have an ability to form thoughts involving any possible content. It seems plausible to me that some pieces of mental content are tied to methods of production. For example, it's plausible that visual perception is able to represent fine-grained colour properties that thoughts are unable to. We can think about colour determinables, but not fully determinate colour properties. In this case, it seems that these determinate colour contents are tied to the perceptual method of content production. Similarly, it might be that the content distinctive of belief is tied to the inferential method of content production. We can only come to have beliefs through a process of inference because only that process is able to produce contents distinctive of belief.

Still, one might be skeptical of the proposal that beliefs have any extra content. The extra content has to have something to do with P's being a *fact*, or P's *truth*, or P's *believability*. One might reasonably be suspicious whether one gets genuinely distinct mental contents by adding these properties – if they even are properties. In light of these worries, let's consider the next option for fleshing out the view that desires are representations of goodness.

3.5.2 Desires as Perceptions of Goodness

The view that Stampe himself eventually settles on is that desires are perceptions of goodness. Take the classic visual illusion of a straight stick partially submerged in water. Here, the stick appears to you to be bent. But if you are aware that this is merely an illusion, you will not believe that it is bent. Same with desires. When you desire a slice of chocolate cake, the cake appears to you as being good, even though you might not believe it to be good. According to Stampe:

> Desire is a kind of perception. One who wants it to be the case that p perceives something that makes it seem to that person as if it would be good were it to be the case that p, and seem so in a way that is characteristic of perception. (p. 359)

On this view, the distinction between beliefs about goodness and desires is derivative on the distinction between belief and perception in general. Of course, this move stands or falls on whether this distinction can do the explanatory work that we require of it. In particular, whatever the difference between perception and belief amounts to, it must be such as to account for the difference in *rational role* between beliefs about goodness and desires. The problem that we were initially trying to solve is that of explaining what is special about desires that explains their reason-giving force. But clearly desires and beliefs about goodness play very different rational roles. If the only difference between the two is that the former are *perceptions* about goodness while the latter are *beliefs* about goodness, then whatever *this* difference amounts to must be able to do the relevant explanatory work. If, for example, the only difference is that perceptions play the perception-role, while beliefs play the belief-role, then we are left with our original problem of explaining *why* they play those roles. What is it about *perceiving* that something is good that constitutes the unique reason-giving force of desire?

While the analogy to perception is helpful, it cannot be the end of the story. What it suggests, however, is that desires involve the way that things *appear*. The stick *appears* bent in water, even though it is not. But what is it for something to appear good?

3.5.3 Desires as Representations of Goodness-Appearance

Suppose that you are at the movie theatre, trying to decide which movie to see. You look over the movie posters. Suddenly you see a movie that interests you. “That looks good”, you say. By

this you might have several things in mind. Perhaps it has actors that you enjoy, or is directed by your favorite director. Or perhaps you like the aesthetic of the movie poster. In any case, what you mean when you say that the movie "looks good" is that the movie has certain features that you can use to predict whether the movie will be good. These features are not themselves 'goodness'. Rather, they *mark the presence* of goodness to you.

I think that this is just so with desires. Desires do not represent that something is *good*, but that something has a certain *marker* for goodness. Speaking loosely, to desire something is for our mind to tell us: "hey check this out, it looks good". The final strategy for fleshing out DRG, then, holds that desires do not represent *goodness* per se, but some property that can serve as a marker for goodness.

This proposal is developed in the final chapter of the present thesis. There I argue that there is good neuroscientific reason to believe that the brain keeps track of the value of *rewards*, and that such representations are vital to the operation of goal-directed learning. If this is right, then we have a prime candidate for a marker of goodness. Further, because of the role that these reward representations play in the production of behaviour, they are ideally suited to account for the distinctive causal and rational role of desire. On this view, then, to desire P is to represent that P is rewarding (to some degree), and to believe that P is good is simply to represent that P is good.

Earlier, when considering the view that beliefs might have extra contents, we considered the following objection. If a belief that P is just a representation that P is a fact, then aren't I able to come to believe anything, by merely forming the thought that *P is a fact*? It seems a similar objection can be made against the view that desires have extra contents. If a desire that P is just a representation that P is a reward, then aren't I able to come to desire anything, by merely forming the thought that *P is a reward*?

I think this objection can be dealt with in the same way. First: it might be that the content distinctive of desires is not expressible, and that "being a reward" is the closest approximation to it. So, when we take ourselves to be thinking that P is a reward, we aren't really invoking the distinctive content of desire. Further, it might be that one does not have an ability to form thoughts involving any possible content. So, just like how certain perceptible properties can only be represented in all their fine-grained glory by perception, it might be that the content

distinctive of desire is tied to a special method of content production that crucially involves the goal-directed learning systems.

In sum, we considered three ways of distinguishing desires from beliefs about goodness. According to the first, beliefs have extra content: perhaps they present their propositional contents as being *facts*. The problem with this view is that it is just not clear whether this really *is* extra content: whether there really is any difference between an inner state that represents that P and an inner state that represents that P is a fact. According to the second proposal, desires are perceptions of goodness. The problem here is that it is just not clear what the distinction between perceptions and beliefs amounts to. The final proposal, which I think is the right way to go, holds that desires do not represent that something is *good*, but that something is a *reward*. The property of reward serves as a *marker* for goodness, and the representation of it is better suited to account for the distinctive causal and rational role of desires.

3.6 Pure Attitude Representationalism

Let us take stock. Propositional attitudes are usefully understood to have two components: propositional content and attitude type. The job of a theory of propositional attitudes is to tell us what these components are. The standard view holds that propositional attitudes involve mental representations, and identifies their propositional content with the content of these mental representations. In addition, the standard view identifies the attitude type of a propositional attitude with the functional role of the mental representation.

The question we were left with at the end of section 3 was: why do propositional attitudes play the roles that they do? When we explain Joan's umbrella taking behaviour in terms of her belief that it is raining and her desire to not get wet, what is it about this kind of state that we find so explanatory? It is not enough to simply build in dispositions to cause such behaviour into the essence of propositional attitudes. This is because folk psychological explanations allow us to make *sense* of the behaviour of others. These explanations are not merely causal, but rational as well. Joan's behaviour becomes rationalized in light of these attitudes. What mere functionalism about attitude type is missing is a positive conception of the nature of the attitudes that explains why they play the functional roles that they do.

The suggestion of the previous section is that this positive conception of the attitudes involves the distinct way that they present their propositional contents to us. Desires present their propositional contents as being rewards, while beliefs just present their propositional contents. This is what was needed to explain the power of the attitudes to provide us with subjective reasons; but it can also serve as an explanation of why the attitudes play the causal roles that they do. A desire —but not a belief — that I get the job that I am interviewing for tends to cause me to work extra hard on my preparation because only a desire presents the proposition that <I get the job that I am interviewing for> as a reward[6].

According to the present account, then, the two components of the propositional attitudes - propositional content and attitude type - are simply aspects of the same phenomenon: mental representation. On this view, for me to desire that I get the job that I am interviewing for is simply for me to be in a mental state that represents that the proposition <I get the job that I am interviewing for> is a reward. For me to believe that it is raining is simply for me to be in a mental state that simply represents that the proposition <It is raining>. I call this view Pure Attitude Representationalism (PAR).

According to PAR, the relationship between what we call the 'propositional content' of an attitude and what is mentally represented is not as straight-forward as the standard view assumed. While the standard view identifies the propositional content of an attitude with the content of a mental representation, PAR holds that what we call the 'propositional content' of an attitude is only an aspect of what is mentally represented. The mistake of the standard view was

---

[6] One might object by pointing out that an account is owed of how such states are capable of playing the causal role that they do. Why should a state that presents a proposition as being good tend to cause me to act as to make it true? How do our behaviour producing mechanisms hook up with the representational content of mental states? I agree that these are important and interesting questions to ask, but I don't think that my view requires an answer to them. My view is presented as an alternative to the traditional functionalist picture of attitude individuation. According to this picture, a desire that P is a mental state that represents P and plays the desire-role. So, the traditional picture is also committed to their being content-bearing states that somehow hook up with our behaviour producing mechanisms. Worse, though, is that the traditional picture gives no account of what it is in about these states in virtue of which they have these effects. It simply says that desires are those states that play the desire-role, however they end up doing it. What my view adds is that these states must possess a content capable of playing the rationalizing role that they do, and it is in virtue of possessing this content that they play the desire-role.

to slide too quickly from 'being related to a proposition' to 'mentally representing that proposition'. Once this move has been made, it is difficult to make room for the different ways of being related to propositions. So, the standard view makes use of a typical explanatory tactic of modern philosophy of mind, and turns what cannot be explained into 'functional roles'. But as we have seen, this fails to capture the fact that attitude types — different ways of being related to propositions — play explanatory roles very similar to that of propositional contents. Crucially, it fails to capture their role in providing us with reasons. The solution, according to PAR, is to recognize that when we speak of being related to propositions in different ways, we are speaking of different ways that those propositions are presented to us, which is fundamentally a matter of mental representation.

Notice that, according to PAR, there is no metaphysical requirement that a mental state plays the desire-role in order to count as a desire. According to the traditional picture, subjects can be said to have desires only if they are disposed to act in ways appropriate to those desires, under suitable conditions. The "suitable conditions" qualifier is important. One can be said to desire a slice of chocolate cake, according the traditional picture, even if one is not engaging in chocolate cake acquiring behaviour. One may not know how to acquire chocolate cake, or one may not be able to engage in the behaviour necessary to acquire it — for example, if one is paralyzed[7]. Nevertheless, the traditional picture maintains that if these "suitable conditions" are met, then an agent who desires a slice of chocolate cake necessarily engages in chocolate cake seeking behaviour.

In Chapter 9 of his book Mental Reality, Galen Strawson (1994) presents a counter-example to this claim: a case of subjects that ought to be described as having desires, despite not being disposed to act in the appropriate ways towards the objects of their desires, even though the relevant "suitable conditions" are met. He asks us to imagine a group of individuals called "weather watchers", who, despite not having bodies and thus are incapable of movement, passively watch the weather going on around them. Strawson argues that it is natural to describe such creatures as having desires about the weather, despite their inability to engage in the

[7] Suitable conditions also plausibly involve not having competing desires, knowing whether something is a slice of chocolate cake, etc.

behaviour appropriate to such desires. Further, this lack of behaviour would not be due to lacking the appropriate connecting beliefs, or their bodies not functioning properly. So, for these creatures, "suitable conditions" are being met even when they are not disposed to engage in any suitable behaviour.

According to Strawson, desire is like oxygen:

> Oxygen is apt, in certain circumstances and when combined with certain other things, for quenching thirst in human beings, who ingest it. Similarly, it is apt, when combined with certain other things, to kill human beings who ingest it… Beliefs and desires are like this in their relation to action and behaviour. Given their already existing and independently graspable nature, they can enter into combinations in which they are said to constitute dispositions to action and behaviour. (pp. 277)

Strawson's claim, then, is that what we call "desire" is a state with a graspable nature that is independent of its behavioural dispositions. It is states with this nature that we attribute to weather watchers. For creatures like us, these states end up causing action. But this is not essential to desire.

I think this is exactly right. Beliefs and desires can only play the explanatory role that they do if they have a graspable inner nature. Further, because explanations in terms of beliefs and desires not only tell us what caused behaviour, but also why it was rational, this inner nature must be the sort of thing that can be rationalizing. The inner nature of a desire for chocolate cake must be such that it allows us to make sense of why one would act as to acquire a slice of chocolate cake: how does chocolate cake appear such that my acting as to acquire it is rational? Only mental contents are able to play this role, because only mental contents involve the way things appear to is.

Nevertheless, some philosophers might insist on only describing someone as believing or desiring something if they are disposed to act in the right sort of way under suitable conditions. At the end of the day, this may be a terminological issue. So be it. What I hope to have shown is that, if this is how you choose to use the terms "belief" and "desire", then it is not these terms

that figure in everyday folk psychological explanation, because they do not have the right explanatory power.

3.6.1 Beliefs and Entertainings

As we've seen, PAR holds that all attitudinal difference is a representational difference. The difference between a belief that P and a desire that P is that the latter represents P as a reward. It is representational content, not functional role, that is individuative of attitude type. However, one might object that this leaves no room for mere *entertainings*. It seems like I am able to merely entertain the proposition that <the moon is made of cheese>, without thereby believing it. But if beliefs are merely representations of their propositional contents, then it seems that, according to PAR, I *do* come to believe that the moon is made of cheese whenever I entertain that proposition.

There are two lines of response open to the proponent of PAR. First, one could deny the existence of mere entertainings. Mandelbaum (2013), for example, argues for a model of belief fixation where assenting or dissenting to a proposition is an automatic process. So, when I decide to entertain the proposition that <the moon is made of cheese>, I automatically assent or dissent, and thus either believe that the moon is made of cheese or believe that the moon is not made of cheese.

However, it is not clear whether Mandelbaum thinks that there is no such thing as mere entertainings, or that they just automatically become beliefs. If it's the latter, then PAR still faces a challenge: what exactly is it that changes when entertainings automatically become beliefs?

Another line of response is to hold onto the existence of mere entertainings (which I think is more phenomenologically plausible), but *deny* that entertainings and beliefs have the same content. I think the most plausible way of fleshing this out is to hold that when I merely entertain the proposition that <the moon is made of cheese>, I represent something like: <it is possible that the moon is made of cheese>. After all, mere entertainings do not commit one to the truth of the proposition. This is part of the rational role of entertainings. So, its representational content should be able to account for this. Adding modal qualifiers satisfies this desiderata: thinking that

it is possible that the moon is made of cheese does not commit one to thinking the moon *is* made of cheese.

Of course, now there is the problem of distinguishing between entertainings and beliefs about *possibility*. However, it seems plausible to me that these might end up being the same thing. To entertain P is *ipso facto* to believe that P is possible (under a fairly broad conception of possibility.)

To sum up: propositional attitudes have two components: propositional content and attitude type. The standard view identifies the former with the content of a mental representation, and the latter with the functional roles of these representations. Because of this, the standard view is unable to explain why beliefs and desires play the causal and rational roles that they do. In contrast, PAR understands both components of the attitudes to be aspects of the same phenomenon: mental representation. According to PAR, beliefs represent their propositional contents, while desires represent their propositional contents as being a reward. This allows for an explanation of why beliefs and desires play the causal and rational roles that they do, and thus vindicates their role in the common sense explanation of human behaviour.

3.7 Lewis Against Desire as Belief

In a series of classic papers (1988; 1996), David Lewis mounts an attack on the anti-Humean position that desires are - or are necessarily conjoined with - beliefs about goodness. According to this *desire as belief* thesis (DAB), whenever an agent desires that P, she necessarily also believes that P is good. Further, the strength of her desire that P is identified with the credence of her belief that P is good. Thus, understanding V to be a function that assigns values to propositions, and C to be a function that assigns credences to propositions:

(1) $V(P) = C(P \text{ is good})$

As Lewis argues, DAB faces difficulties when combined with the fairly plausible assumption that the strength of one's desire for P should not change upon learning that P is true. So:

(2) $V(P \mid P) = V(P)$

Since (1) is supposed to hold under any assignment of credences, we also have:

(3) V(P | P) = C(P is good | P)

But now once we combine (1), (2) and (3), we get the following

(3) C(P is good | P) = V(P | P) = V(P) = C(P is good)

We thus have the result that one's belief credence for the proposition that P is good cannot change upon learning that P. In other words, according to Lewis, the desire as belief thesis leads to the conclusion that P is good and P are probabilistically independent. But, there are probability distributions where they are not. For example, if we learn that ¬(P and P is good), then P and P is good will no longer be probabilistically independent. Therefore, the desire as belief thesis is false[8].

To evaluate Lewis' argument, note that it relies on the following two assumptions:

(1) The strength of your desire that P does not change upon learning that P.

(2) The strength of your desire that P corresponds with the credence of your belief that P is good.

(1) is fairly plausible. Suppose that you have a job interview, and desire that you get the job. Now suppose that you get a phone call a week later, informing you that you got the job. Intuitively, you still want the job just as much, even after learning that you got it.

(2), however, strikes me as quite implausible. The strength of your desire that P concerns how much you want P. You might really want to get that job, perhaps because it's ideally suited for you, it's in a nice city, pays well, etc. Or you might want it only a little bit, perhaps because it is just one potential job among many. However, the credence of your belief that something is good is something quite different. Suppose that there are two jobs that you are applying to. The first job is with company A, and is nothing special. It pays decently and is not too far away. In addition, you have a lot of friends working similar jobs at company A who all attest that it is

[8] The presentation of Lewis' argument owes a lot to Bradley and List (2009).

pretty decent. You have visited the company, and have seen that the working conditions are above average. In this case, we might say that you have a belief, with a very high credence, that you getting that job would be pretty decent. The second job is with company B. You have heard rumours that working at company B is very pleasant. Apparently, their offices have a fully stocked cafeteria with free, healthy and delicious food! You have also heard rumours that they pay much more than company B. You thus believe that you getting a job at company B would be really good. However, since you learned this information through rumours, and have never actually spoken with someone who works there, your credence is fairly low.

As we can see, the amount of goodness you believe something to have, on the one hand, and your credence in that belief, on the other, are quite distinct. The proponent of the desire as belief thesis, then, has a choice to make: which of these two phenomena should we identify with desire strength? Recall that the root of Lewis' problem was the plausible assumption that the strength of one's desire for P should not change upon learning that P. If we identify desire strength with belief credence, then we must say that the credence in your belief that P is good can also never change upon learning that P. If we instead identify desire strength with the degree to which you believe that something is good, then we must say that the degree to which you believe that P is good can never change upon learning that P. Unfortunately, there are cases where both of these are violated.

Suppose that, while you desperately want the job at company B, you believe that you are completely unqualified for it. You think that if they hired you, they would be making a very poor decision. You believe this so strongly, in fact, that you conclude that if they hire you, then it must not actually a good place to work after all. Now, suppose that you acquire some evidence in support of your getting the job. Perhaps a friend that works there tell you that she overheard her boss saying that they plan on hiring you. So now you must update the credence of your belief that you will get the job. But, because you believe that that the company would be making a poor decision if they hired you, you now also have reason to both lower your credence in your belief that the job is good, and lower the degree to which you believe that the job is good. So, it looks like no matter what the proponent of the desire as belief thesis identifies with the strength of one's desire to get the job, she must say the strength of one's desire to get the job is lowered upon learning that one will get the job.

However, I think that the initial assumption that the strength of one's desire that P cannot change upon learning that P is, despite initial intuitive appearances, just plain false. Reflect on the case under consideration: does it not strike you as quite implausible that your desire to get the job at company B would be wholly unaffected upon learning what you learned? Perhaps your insecurity about your qualifications are not reasonable. Nevertheless, it might be perfectly reasonable given these insecurities to desire the job less upon learning that you are more likely to get it.

The assumption that a desire's strength is wholly independent of our beliefs, such that no matter what we learn we will never be given reason to alter our desires, paints a grim picture of our mental lives. It paints us as slaves to our desires: we just can't help but to desire what we do, and nothing we learn about the world will ever change that. Admittedly, some of our desires are like that. Those unfortunate to find themselves in states of addiction can attest to this. And more mundane cases of weakness of the will are more commonplace. We often find ourselves knowing that what we are doing is wrong; we just can't help ourselves. We know that we should be working (on our article, for example) yet can't help our strong desire to spend time with our friends. Yet as common as such cases are, they are the minority. The vast majority of our desires are responsive to reason. We learn things about the world, and come to desire things in response to the things that we learn.

3.8 Conclusion

Beliefs and desires are centre stage in our mental lives. Our understanding of them is essential to the everyday practice of explaining each other's behaviour. Traditionally, philosophers have understood beliefs and desires to be mental representations, distinguished by their functional roles - what they do, or cause us to do. As we have seen, such a view is unable to account for the explanatory roles that beliefs and desires play in our lives. We use beliefs and desires to understand why people do what they do, in a way that renders their behaviour reasonable. I have argued that the best way to account for the distinctive rational roles that beliefs and desires play is in terms of the distinctive ways that beliefs and desires present their propositional contents. On this view, which I call Pure Attitude Representationalism, the difference between belief and desire is not functional role, but mental representational content. Beliefs merely represent their propositional contents, while desires represent their propositional contents as being a reward.

3.9 Bibliography

Block, N. (1989). Can the mind change the world? Meaning and Method: Essays in Honor of Hilary Putnam. G. S. Boolos, Cambridge University Press**:** 137--170.

Bradley, D. (2013). "Functionalism and The Independence Problems." Noûs **47**(1): 545-557.

Bradley, R. and C. List (2009). "Desire-as-belief revisited." Analysis **69**(1): 31-37.

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

Fodor, J. A. (1978). "Propositional attitudes." The Monist **61**(October): 501-523.

Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind, MIT Press.

Kim, J. (1998). Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation, MIT Press.

Lewis, D. (1988). "Desire as belief." Mind **97**(418): 323-332.

Lewis, D. (1996). "Desire as belief II." Mind **105**(418): 303-313.

Mandelbaum, E. (2013). "Thinking is Believing." Inquiry **57**(1): 55-96.

Mendelovici, A. and D. Bourget (2014). "Naturalizing Intentionality: Tracking Theories Versus Phenomenal Intentionality Theories." Philosophy Compass **9**(5): 325-337.

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

Stampe, D. W. (1987). "The authority of desire." Philosophical Review **96**(July): 335-381.

Strawson, G. (1994). Mental Reality, Mit Press.

4 Attitudes and Self-Knowledge

4.1 Introduction

It is commonly thought that we have special epistemic access to certain of our mental states, which affords especially well justified beliefs about them. Usually, privileged access is extended to phenomenal and representational properties: what experiences are like, and what they represent. However, the question of whether we have privileged access to *attitudinal* properties has not been given as much attention. These sorts of properties characterize different sorts of ways of being related to a representational content. In this paper, I focus on the attitudinal properties of being a belief or desire. When you believe that P, you bear the belief relation to P, whereas when you desire that P, you bear the desire relation to P. Do we have privileged access to our beliefs and desires?[9]

In the first part of this paper, I argue that there is good reason to believe that we do in fact have privileged access to attitudinal properties, and that such access is roughly on par with that to phenomenal and representational properties. We are never in a position where we have privileged knowledge that we have some occurrent attitude towards P, but are not sure whether it is a belief that P or a desire that P.

In the second part of this paper, I argue that our privileged access to attitudinal properties, when combined with a certain traditional picture of their nature, raises issues roughly analogous to the issues surrounding the compatibility of privileged access to representational properties and the view that mental contents are *broad*. Unfortunately, I will show that the standard sorts of responses proponents of broad contents have made are not as compelling when applied to attitudinal properties.

I will conclude by arguing, then, that the best explanation of our privileged access to attitudinal properties requires the *rejection* of the traditional picture of their nature. In its place I defend

[9] Henceforth, by 'attitudinal properties' I specifically mean the properties of being a belief or a desire.

Pure Attitude Representationalism, which holds that attitudinal properties are essentially representational properties.

4.2 Privileged Access

Self-knowledge[10], it is often said, is *special* in some way. Our knowledge of our own mental states is in some sense superior to our knowledge of the mental states of others. This is the claim of privileged access. This expression has famously been used by philosophers in a variety of distinct ways[11], according to what one takes to be special about self-knowledge. I think it's best to understand privileged access as consisting in a version of one or both of the following claims:

1. Self-knowledge is epistemically better than our knowledge of the mental states of others.
2. Self-knowledge is accomplished using a method that is different from how we come to know the mental states of others.

(1) concerns the epistemic status that beliefs about our own mental states enjoy. This can be understood in a variety of ways. It might mean that self-beliefs have an especially high degree of *justification,* the extreme being the Cartesian view that self-beliefs are *certain*. Or, it might mean that self-beliefs are much more likely to be *true,* the extreme again being the Cartesian view that they are *infallible.* We would never doubt someone's sincere assertion that they are in pain or prefer tea to coffee, but we would always be open to being corrected about our judgments about another person's mental states

(2) concerns *how* we come to have beliefs about our own mental states. It claims that we possess a special method that can only be used to form beliefs about our own mental states, and not the mental states of others. For example, it is plausible that we must rely on inference from observed behaviour in order to come to know the mental states of others. I might come to know that you are angry because of your red face and furrowed brow, or I might come to know that you want to

[10] By 'self-knowledge' I will mean knowledge of our own mental states; 'self-beliefs' will refer to the corresponding beliefs.

[11] Alston (1971) is an excellent review of the variety of ways "privileged access" has been used in the literature.

leave the party because you tell me so. This is not, however, how I come to know that I am angry or that I want to leave the party. To know these things about myself, I do not need to infer from my behaviour, linguistic or otherwise. Likewise, no one else but me can come to know these things about myself in the same way. I have, rather, a special method for coming to know of my own mental states. I will use the term 'introspection' to refer to this special method, whatever it happens to be[12].

While (1) and (2) are distinct, they can be related in a variety of ways. (2) might *explain* (1): self-beliefs are epistemically better because they are formed by a special process. Or (1) might provide *evidence* for (2): self-beliefs being epistemically better is evidence that they are formed by a special process because they are. Relatedly, the most defensible version of (1) might necessarily contain reference to (2). It is not plausible that *all* self-beliefs are epistemically better, since I might come to believe that I believe P just because someone told me that I do. Rather, it is plausible that self-beliefs are only epistemically better when they are formed in a special way: namely, the method referred to in (2). In what follows, I will understand the claim that we have privileged access to our mental states to mean both (1) and (2).

As stated, (1) and (2) concern our mental states in general. They are very broad claims about the asymmetry between our access to our own mental states, and our access to the mental states of others. Whether this broad reading of (1) and (2) is plausible depends on what you take to be included in the domain of mental states. For example, contemporary vision science posits all sorts of states that are operative in early visual processing, such as edge detecting. Clearly, (1) and (2) are not true of these sorts of states. Neither are they true of traits such as friendliness or stubbornness. It is not plausible, then, that we have privileged access to all of our mental states. It is also not plausible that we have privileged access to all of the *properties* of our mental states.

---

[12] It is important to note that (1) and (2) do not entail one another. It might be that self-beliefs are epistemically much better than beliefs about the mental states of others, even though they are formed through the same method. For example, perhaps we form both on the basis of inference from observed behaviour, the difference just being that we know a lot more about our own behaviour. It also might be that self-beliefs are formed by a special method despite being epistemically on par with beliefs about the mental states of others. For example, there might be a special mechanism in the brain which functions exclusively to detect the presence of certain mental states in us, but which is nevertheless no more reliable than the method that we use to know the mental states of others.

I want to get clearer on what sorts of mental properties we intuitively have privileged access to. I will do this by exploring perhaps the most obvious candidate for privileged access: perceptual experiences.

As I look around my surroundings, I see the redness of the coffee cup in front of me. On the basis of this, I come to know that I am currently having a perceptual experience of redness. In contrast, I might come to believe that you are having a perceptual experience of redness because I notice that your eyes are directed towards the very same coffee cup that mine are. Thus, my belief that you are having a perceptual experience of redness is formed on the basis of inference from observed behaviour, while my belief about my own experience is not. This difference in belief forming method is reflected in a difference in the epistemic status of the beliefs. I might very well be wrong that you are currently having a red experience. Perhaps you are blind, or perhaps your visual system works slightly differently from mine. I am much less certain of my belief about your perceptual experiences than I am in my belief about my own. In my own case, I find it very hard to entertain the possibility that I am wrong that I am currently having a perceptual experience of redness.

Granting that we have privileged access to perceptual experiences, it is not plausible that we have privileged access to all of the *properties* of a perceptual experience. Suppose that I am not hallucinating: there really is a red coffee cup in front of me. In this case, my perceptual experience of the red coffee cup has the property of being veridical. But this is certainly not something that I have privileged access to: I can very easily entertain the possibility that I am currently hallucinating. And I also don't have privileged access to the fact that the sort of perceptual experience I am now having *disposes* me, in certain circumstances, to go get another cup of coffee.

These properties of perceptual experiences that we do not have privileged access to have something in common: they are *relational*. They are characterized by how my perceptual experience relates to things distinct from it. It is no surprise, then, that we would not have privileged access to such properties. Intuitively, when I introspect my perceptual experience, I am confronted with aspects of its *intrinsic nature*.

In general, it seems plausible that we have privileged access to the *phenomenal properties* of our perceptual experiences. These are those properties that characterize what it is like to be us, at some moment. The redness of my perceptual experience is one such property: part of what it is like to be me is characterized by this redness. The way that I know what phenomenal properties are instantiated in my experience is unlike how you know what phenomenal properties are instantiated in my experience, and this difference is reflected in the better epistemic status of my self-belief.

In addition to phenomenal properties, it is also intuitive that we have privileged access to the *representational properties* of our perceptual experiences. The perceptual experience that I have when I look at the red coffee cup presents the coffee cup to me *as being* red. It tells me that the world is a certain way: namely, that there is a red coffee cup in front of me. Thus, it has the property of representing that there is a red coffee cup in front of me. When I see you looking at the same coffee cup, I must infer that you too are probably having a perceptual experience with this very same representational property, but I could be wrong[13].

What about states other than perceptual experiences? At this moment, in addition to occasionally glancing at my red coffee cup, I am also working on writing the present paper. In order to do so, I must entertain all sorts of thoughts. Right at this very moment, in fact, I am entertaining the thought that phenomenal properties are mysterious. How do I know that I am currently entertaining that thought? It seems quite natural to say that, however it is that I know this, I know it in a quite different way than how I know what you are thinking. I might infer form the fact that you are currently reading this paper that you are also having that thought. In this I could very well be wrong - perhaps you just skipped over that line, or stopped reading as soon as 'phenomenal' was mentioned at all. But it seems that I cannot make a similar mistake about my own case. I find it very hard to entertain the possibility that I am not, and never did, in fact, entertain the thought that phenomenal properties are mysterious. Thus it seems that present episodes of conscious thought are among the mental states that we have privileged access to.

---

[13] On many views, representational properties are relational. So there is a tension between the intuition that we cannot have privileged access to relational properties of our mental states, and the intuition that we have privileged access to representational properties. This tension is explored in the later sections.

But, as before, I do not have privileged access to all of the properties of this thought. Some philosophers believe that there are proprietary phenomenal properties for episodes of conscious thought. For people in this camp - the proponents of *cognitive phenomenology* - there is something that it is like to be consciously thinking that *P,* something over and above any associated sensory phenomenology[14]. If you think that there is cognitive phenomenology, then you probably also think that we have privileged access to such phenomenal properties of conscious thought.

Less controversially, thoughts also have representational contents. The representational properties of a thought are often the primary means of individuating them. The example thought above exemplifies this: it was introduced as the thought that phenomenal properties are mysterious*:* that is, as a thought which represents that *phenomenal properties are mysterious*. It is quite intuitive that we have privileged access to the representational properties of our thoughts. It seems that we are always in a position to know *what* we are consciously thinking of, and that we can come to know what we are thinking of in a way that no one else can[15].

What about attitudinal properties: that is, the property of being a belief or a desire? Do we have privileged access to these properties? Consider the plausibility of the following case. You are working at your desk and then, all of a sudden, the content *I have a meeting with a student this afternoon* enters your introspective awareness. You have privileged access to a mental state that represents that *I have a meeting with a student this afternoon.* However, you are just not quite sure whether you are now believing this content, desiring this content, or perhaps merely thinking or entertaining it. In order to make this further determination, you must rely on the same sorts of method you use to attribute beliefs and desires to other people. You observe your own behaviour, and consult your folk theory of the attitudes. Perhaps you notice yourself preparing your office to accommodate the student. Ah ha! you think. You must *believe* that you have a

---

[14] Thus, you might think that there is something it is like to consciously think that P, but only because consciously thinking that P causes you to engage in some sort of sensory imagination.

[15] In fact, this is often taken to be an argument for the existence of cognitive phenomenology. Pitt (2004), for example, argues that "it would not be possible … to identify one's conscious thoughts unless each type of conscious thought had a proprietary, distinctive, individuative phenomenology."

meeting with a student this afternoon. If you instead *desired* this, then you would not be preparing for their visit.

Such a case seems to be obviously implausible. It is not that we merely have the content *I have a meeting with a student this afternoon* in our introspective awareness, floating free from the attitude that we bear towards it. We do not, in the normal case, need to rely on inference from our behaviour to figure out just whether we *believe* or *desire* this content. Rather, apprehension of the content appears to come along with apprehension of the attitudinal relational we bear to this content.

This is reflected in our how we treat a sincere self-attribution of beliefs and desires. Consider the following scenario. You and your friend Jane are on your way to have lunch at your favourite restaurant. This restaurant is a little peculiar in that it serves different dishes every day, with no predictable pattern. You say to your friend: "I really hope that they are serving lasagna today!" Jane responds: "Are you sure that you *hope* that they are serving lasagna today? Perhaps you instead *believe* that they are serving lasagna today?" Clearly, Jane's response is inappropriate. It does not make sense to question someone's sincere assertion that they are having an occurrent hope (or desire) that P and not a belief that P. This is just as inappropriate as her saying: "Are you sure that you hope that they are s*erving lasagna*? Perhaps instead you hope that they are *closed*?" This is an inappropriate thing to say because there is no one in a better position than yourself to know that the content of your attitude is. In the same way, there is no one in a better position to know what kind of attitude you are now holding towards the proposition that they are serving lasagna today.

Just as with perceptual experience, it seems that we do not have privileged access to various *relational* properties that involve our attitudes. While I know that I desire that they are serving lasagna, I may not be aware of what *caused* me to have this desire. There is, in fact, a lot of empirical support of the claim that we are, in general, rather poor judges of the causes of our attitudes. In a widely cited paper, Nisbett and Wilson (1977) present evidence that subjects are unable to report on a change in their attitudes as being a result of experimental manipulation (Bem and McConnell 1970), or as a result of convincing argument (Goethals and Reckman 1973). They also report on studies that indicate that subjects misidentify the causes of their

physiological and emotional states (Storms and Nisbett 1970). Nisbett and Wilson themselves conduct a series of experiments investigating a subject's ability to accurately report on external influences on their attitudes. They found that subjects would quite often, under a variety of circumstances, confabulate explanations as to why they feel a certain way about something, or why they believe what they do.

To sum up: we have privileged access to our perceptual experiences, and occurrent thoughts and other attitudes such as belief and desire. We have a special method for coming to know about the phenomenal and representational properties of these states. In addition, and most importantly for the purposes of this paper, we saw that we have privileged access to the properties of being a belief or a desire. Our epistemic access to these attitude properties seems to be on par with our epistemic access to phenomenal and representational properties. We never seem to be in a position where we know that we having *some* sort of attitude towards the proposition P, but are not sure whether we *believe* or *desire* it. Thus, we have privileged access to attitudinal properties. However, it seems implausible that we privileged access to their *relational* properties. Studies show that we are often very poor judges of the causes of our attitudes

4.3 What are Attitudinal Properties?

On the traditional picture, being a belief or desire are functional properties of mental states[16]. Beliefs and desires are those things that play a certain role in inference and the production of behaviour. For example, a belief that P is thought to be that thing which, for example: tends to be caused by a perception that P; tends to be extinguished by a perception that not P; tends to cause, when combined with a belief that P > Q, a belief that Q; tends to cause one to assert that P, and so on. A desire that P, on the other hand, is thought to be that thing which, for example: tends to cause one to act as to make it the case that P; tends to cause, when combined with a belief that O leads to P, one to O; tends to cause one to assert that one wants P, and so on.

[16] The traditional picture can be found in many works, either explicitly or implicitly. For example, Fodor (1987) defends the view that a belief that P is a representation that P in the "belief-box", *mutatis mutandis* for the other attitudes. Dretske (1988) understands beliefs and desires in terms of the causal roles that they play in the production of behaviour. Cummins (1996) distinguishes between beliefs and desires according to their "cognitive function".

Let the above conditions for belief be denoted by 'belief-role', and the above conditions for desire be denoted by 'desire-role'. The thought is, then, that a belief that P is a representation that P which plays the belief-role, while a desire that P is a representation that P which plays the desire-role.

On this view, whether a given mental state counts as a belief or a desire is not simply a matter of its intrinsic nature. One could have two intrinsically identical mental states that both represent P, but one would be a belief and the other a desire in virtue of how these states are *disposed* to behave[17]. Notice that the criteria that make up the belief and desire roles are all qualified by "tends to." This is because a mental state counts as a belief or a desire only if it exhibits the criterial behaviour under certain circumstances, given appropriate background conditions. Thus, a belief that P only results in an assertion that P when one does not wish to deceive. A perception that P might not cause one to believe that P because one has further reason to think that their perceptual faculties are misleading them (such as when a stick partially submerged in water appears to be bent). Likewise, a desire that P will only cause one to make it the case that P in the absence of any conflicting motivations or anything that might prevent one from being able to make it the case that P.

These issues speak to the difficulties of specifying exactly what the dispositions relevant to a state playing the belief- or desire-role are. But they also make it clear just how complex attitudinal relations are on the traditional view. One might be counted as believing that P even if one is in a state which never *actually* does anything belief-like. It is enough that one is in a state that *would* do belief-like things *if* the background circumstances allowed it.

Already there is cause to be concerned. We saw in the previous section that while it is plausible that we have privileged access to certain properties of our perceptual experiences — for example, their representational properties — it is quite implausible that we have privileged

---

[17] It is important to distinguish this view from the view according to which beliefs and desires *themselves* are dispositional properties. On this latter view, the property of believing that P — a property of *agents* — is identical to being disposed to behave in certain ways, for example. However, the traditional picture understands beliefs — at least occurrent beliefs — to be non-dispositional properties of an agent. What is dispositional is the property of *being* a belief, which is a property that mental states have. An occurrent mental state counts as a belief if *it* has certain dispositional properties.

access to many of their *dispositional properties.* For example, we found it to be implausible that we have privileged access to my perceptual experience of a cup of coffee being disposed to cause me to go get another cup of coffee. We also found it implausible to suppose that we have privileged access even to my perceptual experience *now* causing me to want a cup of coffee. We saw that there is very good empirical evidence that we are very poor judges of what the causes of our mental states are. In general, we do not have privileged access to the causal relations between our mental states.

But it seems that this is just the way that the traditional picture understands attitudinal properties. The attitude that we bear to P is a matter, among other things, of the causal dispositions of the state that represents P. But if we are in general poor judges of the causal relations between our mental states, then it seems like we cannot have privileged access to attitudinal properties. For how could I have privileged access to my now *believing* that P, if my believing that P was partly a matter of its dispositional causal relations to other mental goings on?

In the remainder of this paper, I hope to flesh out this intuitive tension. I begin by exploring a similar issue: the apparent impossibility of privileged access to *broad* contents. I show that similar issues can be raised for attitudinal properties. Worse, I show that the sorts of replies that can be given on behalf of the possibility of privileged to broad contents do not transfer to the case of attitudinal properties.

4.4 Content Externalism and Observation

Content externalism is the view that what an individual's mental state represents is not simply a matter of the intrinsic properties of that individual. Rather, mental contents are in part determined by the relations that the individual bears to her external environment. Such contents are called "broad".

There are two motivations for the view that mental contents are broad. The first arises through reflection on the conditions under which we intuitively would ascribe a certain propositional attitude to a thinker. As Putnam (1975) argued, we can only ascribe the thought that 'water is wet' to a thinker whose environment predominantly contains H2O. And as Burge (1979) argued, whether we can ascribe a thought containing concepts such as *arthritis* depends on certain facts

about the thinker's linguistic community. In both cases, what a thinker thinks is not fixed by what the thinker is like: reference must be made to facts external to the thinker.

The second motivation for content externalism comes from the desire to naturalize representation. To naturalize some phenomenon is, roughly, to show how it can be given a place within a physicalist, scientific worldview. This usually takes the form of showing how some class of facts reduces to purely physical facts. In the case of representation, many such attempts have been made: some reduced representational facts to causal facts, some to teleological facts. What became clear was that the most plausible naturalistic account of representational is externalist. On the causal theory, for instance, a mental state represents some property P just in case it is reliably caused by all and only instances of P, in certain conditions[18]. Whether or not a given thinker has thoughts about P, then, will depend on whether the thinker's environment is such that there are Ps that bear this causal relation to the thinker.

We are now in a position to appreciate the intuitive tension between content externalism and our having privileged access to representational properties. Suppose I am now having a belief that I would express with the sentence "water is wet". Consider my twin on Twin Earth (where the "watery stuff" is XYZ, not H2O). My twin is also having a belief that he would express with the same sentence. But, as we learned from Putnam, while my belief is about H2O, my twin's belief is about XYZ, due to differences in our external environment. This is despite the fact that, *from the inside*, our beliefs are identical. How, then, could I have privileged access to my belief being about H2O?

The intuitive difficulty in accounting for privileged access to broad contents, I think, is that while broad contents are partially constituted by relations that we bear to aspects of our external environment, introspection has no access to these relations. In his 1989, Boghossian, argues that, in general, one is not able to tell, by merely inspecting an object, what *relational* properties that

---

[18] Causal views must solve the "disjunction problem". Mental states are clearly able to mis-represent. Sometimes my horse detector fires when confronted with something that merely looks like a horse. But then on what basis can we say that it represents horses, and not the disjunctive content *horse-or-horse-looking-thing*? Proponents of causal theories must provide qualifications to rule out intuitive cases of misrepresentation from determining the content of the state in question. See, for example, Fodor's (1987).

object has. His argument for this has two premises. First: that knowledge merely of an objects *intrinsic* properties cannot give you direct knowledge of its relational properties. And second: that inspecting an object can only give you direct knowledge of its intrinsic properties.

Of course, there is a sense in which I can inspect certain relational properties of objects. For example, it seems that we are able to tell through mere inspection what the monetary value of a coin is, which is clearly a relational property. Boghossian argues, however, that we are able to do this only because coins have certain intrinsic properties that covary with their monetary value: shape, size, etchings, etc. Inspection of a coin's monetary value is therefore indirect: it is mediated through direct inspection of the intrinsic properties that covary with its monetary value.

The same can be said for other relational properties that we are able to inspect. On some views of colour, colour is a relational property: the colour of an object is a matter of what wavelengths of light it is reflecting. So we are able to inspect the colour of an object indirectly through direct inspection of the light reflecting off of it. It also seems like we are able tell whether something is bitter or sweet through inspection, even though these are relational properties. But again, in these cases, it is plausible that these relational properties are indirectly inspected through direct inspection of some intrinsic property.

There are other sorts of putative counter examples to Boghossian's claim that only intrinsic properties can be inspected. Suppose I am facing two individuals, Stacy and Paul, who are talking to one another. It seems natural to say that I am able to see that Stacy is looking at Paul, or that her body is oriented towards Paul. But orientation and eye gaze are clearly relational properties. In these sorts of cases, however, I am not able to tell Stacy's orientation with respect to Paul by *merely* inspecting Stacy. I must also inspect Paul.

So, we can flesh out Boghossian's argument by adding the following two qualifications. First: mere inspection of an object can give you *indirect* knowledge of those relational properties that have an intrinsic correlate in that object. This covers the coin case: I can come to know the value of the coin through inspection because its value has an intrinsic correlate. Second: knowledge of relations through inspection of objects is possible if I am able to inspect both relata. This covers the orientation case: I can come to know about the orientation of objects through inspection because I am able to inspect both relata of the orientation.

If this is right, then we have a straightforward argument against the possibility of privileged access of broad contents. All that is needed is the premise that privileged access is accomplished by a kind of inspection. This is to think of the special method of privileged access as a kind of inner observation. If this premise is granted, then privileged access to broad contents would be impossible.

If the proponent of broad content wants to hold onto privileged access, and the idea that privileged access is accomplished by a kind of inspection, then she must demonstrate how broad contents are covered by one of the two qualifications made above. The first qualification requires that content-bearing states have intrinsic correlates of their broad contents, which is precisely what proponents of broad contents deny[19]. And even if such intrinsic correlates could be found, introspection could only result *indirect* knowledge of thought contents. The second qualification is also a non-starter: the relata of broad contents involves aspects of our external environment that we are *not* able to inspect.

4.5 Constitution and Content Embedding

As long as the special method of privileged access is understood to be a kind of inner observation, then it seems that we can only have direct knowledge of the intrinsic properties of mental states, which broad contents are not. The way forward for the content externalist, then, must be to deny that introspection is a kind of inner observation.

What is meant by this? We have been thinking of introspection as a process that *detects* certain features of our mental states, and then *produces* the relevant self-belief. Thus, the introspection of my perceptual experience as of a red cup in front of me involves the detection of certain phenomenal and representational properties instantiated in my experience, and the subsequent production of a belief about them. This is to think of introspection as roughly analogous to sensory perception[20]. My perception of the red cup in front of me involves the detection of certain colour, shape and size properties (among others), and the subsequent production of a

[19] A central tenant of content externalism is that intrinsic duplicates can have thoughts with different contents.

[20] Such a view can be found in Nichols and Stich (2002; 2003), Armstrong (1968) and Lycan (1996).

perceptual state which represents them. The difficulty that we ran into is that the sorts of properties that we are able to introspect are unlike the sorts of properties that we are able to detect in a manner analogous to perception. Perceptual processes are only able to afford direct knowledge of the intrinsic properties of objects, while its seems that we are able to introspect relational and dispositional properties, such as mental contents. Thus, it looks like if we are to explain the introspection of such relational properties, we must find an alternative to the observational model.

Suppose that someone writes a series of numbered sentences on a blackboard. Unfortunately, the sentences are written in some strange language that you have never heard of before. You are given a dictionary that translates all the sentences in the strange language into English. You are then asked to write, on the blackboard, sentences about those sentences, of the form: "Sentence X means that Y". Call this the meta-sentence. How are you to solve this problem?

The obvious solution is to look up the meanings of all the words in the numbered sentences, figure out what they mean, and then write out the meta sentence. This will take a long time, of course, but presumably it could be done. A much easier solution, however, is to simply append "Sentence 1 means that" in front of sentence 1, "Sentence 2 means that" in front of sentence 2, and so on. You have thus produced the right meta-sentences that contain the original sentences as constituents. This method is much faster, does not require the use of the translation book, and is error proof.

The first method can be understood to be roughly analogous to an observation approach to introspection. The meaning of the sentence is detected through a complex mechanism that involves consulting the translation book, and then a meta-sentence is produced that is about the interpreted meaning of the first sentence. Likewise, according to observation approaches, introspection of our thoughts consists in detecting that they have a certain content, and then producing a belief about the detected thought. The analog of the translation book would be a mapping from the intrinsic properties of the detected thought to its mental content. But according to content externalism, such a mapping depends on external factors that we do not have privileged access to.

An alternative to observation approaches understands introspection to be roughly analogous to the second method of producing the right meta-sentences. The idea is that the content of our thoughts does not need to be detected; rather, it simply needs to be redeployed or embedded directly into the content of the self-belief[21]. Thus, the content of my self-beliefs is partially constituted by the content of the first-order state of which it is about. Call this the *constitution* approach to introspection[22].

Take my belief that *water is wet*, and my twin's corresponding belief that *twater is wet*. What needs to be explained is this: how I am able to form the self-belief with content *I am believing that water is wet,* while my twin is able to form the self-belief with content *I am believing that twater is wet,* even though these beliefs are intrinsically identical? The solution, according to the constitution approach, is to simply note that while the content of my self-belief is partially constituted by the content of *my* first-order belief, the content of my twin's self-belief is partially constituted by the content of *his* first-order belief. There is no need to detect or identify the differences between these contents, since the differences reappear in the content of the self-belief for free.

To sum up: as long as we adopt an observational approach to privileged self-knowledge of mental contents, then such knowledge is incompatible with mental contents being broad. Observation of an object can only provide direct knowledge of its intrinsic properties, not its relational properties. The solution is to reject that introspection of mental contents is observational. The contents of first-order states do not need to be detected before producing the relevant self-belief about them. Rather, these contents can be simply embedded into the content of the self-belief. This is how to make privileged self-knowledge of mental contents compatible with content externalism.

---

[21] On Burge's (1988) proposal, self-beliefs are self-verifying because in the very act of entertaining a thought with the content *I am thinking that water is wet*, one is thereby entertaining a thought with the content *water is wet*.

[22] Such accounts can be found in Burge (1988), Heil (1988) and Gertler (2000).

4.6 Attitudinal Properties and Privileged Access

Recall that, according to the traditional picture, belief and desire are dispositional/functional in nature. A given mental state which represents that P counts as a belief or a desire partly in virtue of the causal relations it tends to bear to sensory input, behavioural output and other mental states in certain counterfactual scenarios. For example, a mental state which represents P counts as a desire that P only if it (tends to, given the right background conditions) causes one to act as to make it the case that P.

In this section, I will show that the considerations used to argue against the possibility of privileged access to broad contents raise similar problems for the possibility of privileged access to attitudinal properties, when such properties are understood according to the traditional picture. What is worse, however, is that the sorts of replies that can be given on behalf of the content externalist are not as compelling when given on behalf of the proponent of the traditional picture of attitudinal properties.

4.6.1 Observation of Attitudinal Properties

Recall Boghossian's claim that mere inspection (or observation) of an object can only give you knowledge of that object's intrinsic properties, not its relational properties. This was used in the previous section to argue against the possibility of privileged access to broad contents, because broad contents are relational properties. According to the traditional picture, attitudinal properties are also relational. Whether or not a given mental state M which represents P counts as a desire or a belief that P depends on how M is related to sensory input, behavioural output and other mental states. Worse, since attitudinal properties are dispositional, they involve the relations that mental states bear to sensory input, behavioural output and other mental states in merely *counterfactual* situations. If introspection is observational, then it seems like we cannot have privileged access to attitudinal properties.

We saw in section 4 that we can learn about relational properties such as *orientation* through mere inspection. For example, we are able to see that Stacy is standing to the right of Paul. However, knowledge of such relations is not accomplished merely through inspecting Stacy, but by also inspecting Paul. That is: knowledge of such relations is only possible if one is able to inspect all of the relata. However, the properties of belief and desire, according to the traditional

view, constitutively involve relations to things such as *possible future behaviour*. It seems quite implausible that these sorts of things can be detected.

We also saw in section 3 that there are some relational properties that we are able to *indirectly* inspect, in virtue of directly inspecting intrinsic properties that covary with them, such as the value of a coin. This sort of reply was not able to save observation of mental contents, because content externalism rejects that are such intrinsic properties that covary with broad contents. However, perhaps such a reply could be made to save observation of attitudinal properties. Plausible candidates are neurophysiological properties.

In order for such a response to succeed, neurophysiological properties that are highly – if not perfectly – correlated with the functional roles distinctive of the various attitude types must be identified. As far as I know, this has not been done. Very little work has been done investigating the neurophysiological differences between the propositional attitudes. One might object that there must be such neurophysiological properties, on pain of denying physicalism. However, there are no such neurophysiological properties for mental contents if content externalism is true, and no one, as far as I know, thinks that content externalism is a threat to physicalism.

It is not enough that the functional roles characteristic of belief and desire are neurophysiologically realized — no one denies that. Rather, it must be possible for there to be a "belief detector" that is sensitive to all and only those neurophysiological properties that realize the functional role of belief, *mutatis mutandis* for desire and the other attitudes. But practically any neurological property can realize that functional role, and the very same neurological property can realize the belief-role in one case and the desire-role in another.

However, even *if* such properties could be found, a functionalist about attitudinal properties can at best hold that attitudinal properties are *indirectly* known through detection. We first detect some intrinsic property of the state in question, and then, through some process, come to know which attitudinal property the state instantiates. This results in an asymmetry between our knowledge of attitudinal properties and our knowledge of representational properties[23]. Thus,

[23] That is, if we adopt a constitution account of knowledge of representational properties.

when we have a conscious belief that *I have a meeting with a student this afternoon*, and come to form a self-belief about it, the process by which we come to know what this belief is *about* is distinct from the process by which we come to know that it is a *belief*, and not a desire. To me, this seems quite phenomenologically implausible. When I reflect on what it is like to come to know my beliefs, its representational content and its being a belief come to me at once. This is no knock-down argument against the possibility of observational views of self-knowledge of attitudinal properties. It is, however, cause to be suspicious of them, and perhaps search for a better alternative.

4.6.2 Non-Observational Introspection of Attitudinal Properties

Let us explore, then, the possibility of a non-observational account of the introspection of attitudinal properties. In particular, let us consider a constitution account, according to which I come to know my occurrent beliefs and desires not by detecting the presence of certain attitudinal properties, but by embedding them into the content of the self-belief.

First, it must be noted that such an account will not be as straightforward as a constitution account of the introspection of mental contents. In that case, what needed to be explained was how, when I believe that *water is wet*, I come to form the self-belief that *I believe that water is wet* and not the self-belief that *I believe that twater is wet*. The answer was that part of the content of the self-belief — the *water is wet* part — is simply taken from the content of the first-order belief. In effect, such an account *identifies* a representational property of one state with a representational property of another. My self-belief represents that *water is wet* because that's what my first-order belief represents.

Things are quite different in the case of attitudinal properties. Here, what needs to be explained is how, when I believe that *water is wet*, I come to form the self-belief that *I believe that water is wet* and not the self-belief that *I desire that water is wet*. The answer cannot be that the relevant part of the content of the self-belief — the *I believe* part — is simply taken from the content of the first-order belief, because it doesn't *have* that content. Thus, we cannot just simply identify a representational property of my first-order belief with a representational property of my self-belief. Rather, we must tell a story about how the property of being a belief *itself* — and attitudinal properties in general — come to be embedded into the content of the self-belief.

I think that there are two ways such a story might go. On the *demonstrative approach*, the self-belief is understood to be demonstrative in nature, and points directly at the relevant properties of the first-order state. Perhaps my self-belief represents something like: *I am in* ____, where the blank is filled in by my first-order state. On the *substantive approach*, the self-belief is not demonstrative in nature, but rather affords substantive knowledge of the state in question. Perhaps my self-belief represents something like: *I am in A*, where A is a concept which is (partially) constituted by the first-order state itself.

Both of these approaches have their own difficulties. Suppose I am occurrently desiring that it will not rain tomorrow. According to the demonstrative approach, my self-belief about this desire would have content of the form: *I __ that it will not rain tomorrow*, where the blank is filled in by the attitudinal property of the first-order state that I am directing introspective attention to: which, in this case, is *desire*. Compare this to a case where I am instead occurrently believing that it will not rain tomorrow. Again, my self-belief about this belief would have content of the form: *I __ that it will not rain tomorrow*, except that in this case, the blank is filled in with the attitudinal property of *belief*. Thus, I come to have distinct self-beliefs in these two cases. The content of the self-belief that I form depends on the nature of the state that I direct introspective attention to.

Demonstrative concepts get their reference fixed through an act of directing attention. When I point at the table and say "that color", I can end up referring to the color of the table, and not any of its other properties, at least partially in virtue of the fact that I am *attending* to the table's colour. If I was not able to attend to the table's color, then nothing about my act of pointing would be able to single out that property among all of the other properties of the table. So, if the attitudinal component of the content of my self-beliefs is the result of an act of inner demonstration, then I must be able to attend to the attitudinal properties of my first-order states.

However, it is hard to see how I could be able to attend to such attitudinal properties, as long as they are understood to be relational properties. Earlier we saw that it is implausible that relational properties could be *detected.* It seems that the same could be said of *attention.* Suppose you are having a veridical experience of a coffee cup on your desk. What properties of the experience are you able to attend to? Clearly, you are able to attend to its phenomenal properties. Perhaps you

are also able to attend to its representational properties. But you cannot attend to its *veridicality*. You might contemplate or doubt its veridicality, but you cannot direct *perceptual attention* to it. Suppose that seeing coffee cups disposes you to refill your cup of coffee. You might be aware of this or you might not. After seeing the cup, you might then become aware of a desire to refill it. But you cannot attend to the *disposition* that the experience of the cup has to make you refill the cup. So, it is hard to see how we could attend to — and thus form demonstrative concepts of — attitudinal properties, as long as they are understood according to the traditional view.

The other approach to non-observational introspection of attitudinal properties fares a little better. On the *substantive* approach, we possess concepts of attitudinal properties that afford substantive knowledge of them because these concepts are partially constituted by the properties themselves. On this view, our knowledge of the attitudinal properties of our mental states is similar to our knowledge of their phenomenal properties. According to Chalmers (2003), for example, we possess "direct phenomenal concepts" that refer directly to certain phenomenal properties in virtue of those properties being contained in the concepts themselves. The direct manner in which we are able to think about phenomenal properties results in explanatory gap between the physical and phenomenal. These direct phenomenal concepts are completely isolated from our physical concepts. As a result, no amount of knowledge that makes use of physical concepts will ever suffice for knowledge that makes use of direct phenomenal concepts. Mary, from her black and white room, will not be able to figure out "what it's like to see red".

In this way, the account of the nature of phenomenal concepts is intimately related to the epistemology of phenomenal properties. The substantive approach attempts to extend this account to include concepts of attitudinal properties. However, the analogy breaks down: there doesn't seem to be an explanatory gap between the physical and the attitudinal. But this is precisely what we should expect if we were able to think about attitudinal properties directly, by invoking concepts that contain the properties themselves.

One might object that we should not expect to find an explanatory gap *wherever* such direct concepts are used. It is only *sometimes* that direct concepts result in an explanatory gap. They do in the case of phenomenal properties, but not in the case of attitudinal properties. This move, however, undermines the phenomenal concept strategy of providing a physicalist explanation of

the explanatory gap between the physical and the phenomenal. The phenomenal concept strategy says that an explanatory gap exists not because phenomenal properties are non-physical, but because of the direct way in which we are able to think about them (Chalmers 2003). The hope of the phenomenal concept strategy is to then give a purely physical explanation of these direct phenomenal concepts. However, if direct concepts do not *in general* result in an explanatory gap, but only in the case of phenomenal properties, then the phenomenal concept strategist's explanation of the explanatory gap is not complete. We are owed an additional explanation of why direct concepts result in an explanatory gap in the phenomenal case, but not in the attitudinal case. The worry is that any such explanation must make reference to non-physical features of phenomenal properties, thus undermining the original motivation.

This is not a knock-down objection to the claim that we possess "direct attitudinal concepts". However, it does raise worries that proponents of such a view have to deal with. If there are such concepts that allow us to think about attitudes directly, then why isn't there an epistemic gap between the attitudinal and the physical? Or perhaps proponents of such a view wish to hold that there *is* such an epistemic gap. Is this because attitudinal properties are in fact *phenomenological* properties? Of course, one need not follow the phenomenal concept strategy in holding that the epistemic gap is the result of direct concepts. Nevertheless, it is worth noting the explanatory cost one must pay in order to make a non-observational approach to the introspection of attitudinal properties work.

To sum up: while it is clear that we have privileged access to the attitudinal properties of our mental states, it is not clear how to account for this, as long as we continue to understand them according to the traditional view. If we adopt a detection account, then we must hold that self-knowledge of attitudinal properties is indirect, and thus very different from our knowledge of representational contents. We cannot use a content-embedding account, because, according to the traditional view, attitudinal properties are non-representational. And if we hold that attitudinal properties themselves are contained in the content of the self-belief, then we must either hold that we are able to direct introspective attention at the sorts of dispositional properties that are traditionally taken to be constitutive of attitudinal properties, or posit an explanatory gap between the attitudinal and the physical. It seems, then, that any account of privileged access to

attitudinal properties will face serious explanatory challenges, as long as we hold onto the traditional picture of their nature.

4.7 Pure Attitude Representationalism

According to the traditional picture, attitudinal properties — such as the property of being a belief or a desire — are complex dispositional/functional properties. Whether a given mental state M which represents P counts as a belief that P or a desire that P depends on the causal relations that M tends bear to sensory input, behavioural output and other mental states in various counterfactual circumstances. For example, M is a desire that P only if it tends to cause me to act as to make it the case that P, provided that I have no conflicting desires that are stronger, and that I am able to so act, etc.

In this closing section, I explore what an account of privileged access to attitudinal properties looks like that *rejects* this traditional picture. I suggest a picture according to which attitudinal properties — such as the property of being a belief or a desire — are actually representational properties. On this *Pure Attitude Representationalism* (PAR), a belief that P and a desire that P both represent P, but that's not all they represent. They have further contents in virtue of which one is a belief and one is a desire. I will argue, then, that the best explanation of our privileged access to attitudinal properties involves adopting PAR over the traditional functionalist picture.

To see how this works, consider the difference between the sense modalities. You might think that what makes a given perceptual representation belong to the modality it does is not a matter of what it represents. We can *see* that the coin is circular, and we can also *feel* that the coin is circular. These states seem to have the same content: something like *the coin is circular*. But then in virtue of what is one a state of *seeing* while the other is a state of *feeling*? Here, something like the traditional view of attitudinal properties comes to the rescue: seeing and feeling are just different functional/dispositional properties.

Such a view of the difference between sense modalities will face epistemic problems not unlike those faces by the properties of belief and desire. Presumably, we have privileged access to whether we are seeing or feeling the shape of the coin. But how could this be, if it being a state

of seeing or feeling is the sort of property that we can neither observe nor embed into the content of the self-belief?

Thus there is good reason to look for a better view of the difference between sense modalities. And such a view readily presents itself: the sense modalities just have different contents. There are certain properties — "unique sensibles" — that can only be represented by a single modality. You can only see colours; you can only hear sounds; you can only feel textures; you can only smell odours; you can only taste flavours.

Our original question was how to distinguish between a state of seeing that the coin is round and a state of feeling that the coin is round. On the unique sensibles approach, the answer begins by noting that the descriptions of these states under-specifies their contents. You do not see just the shape of the coin, but also its colour. You do not just feel the shape of the coin, but also its texture. It is this latter difference in their content that makes it the case that one is a state of seeing and the other a state of feeling.

According to PAR, the situation is the same for belief and desire. In addition to P, a desire that P also has other contents that only desires can have, likewise for belief. On the traditional picture, a desire that P is a state which represents P and plays the desire -role. On PAR, a desire that P is a state which represents that P and also has some distinctive desire-content. *Mutatis mutandis* for belief.

Such a view of desire is actually quite common in the ethics — if not the philosophy of mind — literature. According to an influential family of views, desires are *appearances of value* (e.g. Stampe 1987, Oddie 2005). To desire a slice of chocolate cake is for a slice of chocolate cake to appear valuable to you. Here, the distinctive desire-content is something about an object being valuable in some way. Any state that has this content is *ipso facto* a desire.

Of course, we can believe that things are valuable. So in order for such a view to work, we must find a distinctive sense of 'value'. That is: there must be some property $\text{value}_{d,}$ such that any state that represents that property is *ipso facto* a desire. I make the case for the existence of such a property in Chapter 3 of this thesis. In the remainder of this section I want to show that, if we

adopt PAR, then privileged access to attitudinal properties can be straightforwardly be explained as a simple case of content embedding.

Suppose that you are currently having a desire that the Blue Jays win the World Series. According to the traditional view, this is a state which represents that the Blue Jays win the world Series, and which plays the desire-role. In contrast, according to PAR, this is a state which represents something like: it would be good if the Blue Jays win the World Series. Suppose you direct introspective attention towards this state and attempt to form a self-belief about it. According to a content-embedding account of self-knowledge, this is done by taking the content of your desire and redeploying it into the content of the self-belief. According to the traditional view of desires, the content of your desire is simply that the Blue Jays win the world series. But then how are able to know that you *desire* this, rather than *believe* this? This was the mystery we were left with at the end of the previous section.

Since, according to PAR, the property of being a desire *just is* the property of representing that something is good, redeploying the content of the desire into the self-belief is all that is needed. There is no further question of how one is able to tell that one has a *desire*, rather than a *belief,* since one knows the content of the first-orders state, and it is this content that determines whether the state is a belief or a desire.

In addition to its simplicity, this view of our self-knowledge of attitudinal properties is, I think, much more phenomenologically plausible than any view that holds onto the traditional functionalist picture. Suppose you are about to interview for your dream job. As you sit in the waiting area, you have a conscious desiring that you get the job. The proponent of the traditional view seems forced to concede that the process that affords me privileged access to the *content* of this desire is quite distinct from the process that affords us privileged access to it being a *desire*. This just does not seem to be how things go. When I notice my desire that I get the job, I notice its content and its 'desireness' *at once*. Only PAR is able to accommodate this fact: its 'desireness' *just is* part of its content.

One might object that this view can only afford self-knowledge of a state that happens to be a desire — in virtue of the content that it has — but not self-knowledge *that* the state is a desire. After all, someone might just not be aware that states which represent that something is good are

desires. All that PAR and content-embedding guarantees is that I come to believe that I am in a state which represents that it would be good if the Blue Jays win the World Series*, not* that I come to believe that I *desire* that the Blue Jays win the World Series. However, if PAR is the right view of what belief and desires *are*, I think it plausible that we have at least an implicit conceptual grasp on the connection between being a desire and representing that something is good. Given such a grasp, my recognition that my mental state represents that something is good will thereby be a recognition that my mental state is a desire. This is just like how we have an implicit — and sometimes explicit — conceptual grasp on the connection between being a *visual* perceptual experience and representing colours, or between being an *auditory* perceptual experience and representing sounds.

PAR, then, provides a simple and compelling picture of how we able to have privileged access to our attitudes. Beliefs and desires are just representational states, and their two components - attitude type and what they are about - are both aspects of their representational content. When we direct our introspective attention at our occurrent beliefs and desires, we form a self-belief that redeploys their representational content, thus affording an especially direct form of self-knowledge.

4.8 Conclusion

We began with the intuition that we have privileged access to the phenomenal, representational, and attitudinal properties of our mental states, but not their dispositional and relational properties. This presented a puzzle: how could we have privileged access to attitudinal properties - such as belief and desire - if, as the traditional view holds, these properties are dispositional in nature? This is similar to the puzzle about privileged access to broad contents. If what your mental states represent is partially determined by what factors that can only be known *a posteriori*, then how can we have privileged access to such contents? We saw that the culprit is thinking about self-knowledge of contents as a form of *detection*. As Boghossian argues, only intrinsic properties of a mental state can be detected. Detection accounts of self-knowledge are incompatible with mental contents being broad. The solution — in the case of mental contents, was thus to abandon detection accounts of self-knowledge, and instead hold that we come to know about the contents of our first-order mental states by simply redeploying them into the content of our self-beliefs.

Unfortunately, we saw that such an account cannot solve the puzzle of privileged access to attitudinal properties, as long as such properties are understood, as the traditional view does, as being non-representational. The first take-away, then, is that the traditional view of the nature of attitudinal properties renders privileged access to such properties a mystery. The second take-away is that there is alternative to the traditional view of the nature of attitudinal properties that resolves the mystery. On this view — which I call Pure Attitude Representationalism — attitudinal properties are representational properties. The difference between a belief that P and a desire that P is not the functional role of the states, as the traditional view maintains, but simply their representational content. On such a view, the content embedding account of self-knowledge that was used to accommodate privileged access to broad contents can straightforwardly accommodate privileged access to attitudinal properties as well. More needs to be said, of course, in support of a view like PAR. What I have hoped to have shown in this paper, however, is that it is able to best make sense of how we are able to know, with that special authority, our beliefs and desires.

4.9 Bibliography

Alston, W. P. (1971). "Varieties of priveleged access." American Philosophical Quarterly **8**(July): 223-241.

Armstrong, D. M. (1968). A Materialist Theory of the Mind, Routledge.

Bem, D. J. and H. K. McConnell (1970). "Testing the self-perception explanation of dissonance phenomena: on the salience of premanipulation attitudes." Journal of Personality and Social Psychology **14**(1): 23.

Boghossian, P. (1989). "Content and self-knowledge." Philosophical Topics **17**(1): 5-26.

Burge, T. (1979). "Individualism and the mental." Midwest Studies in Philosophy **4**(1): 73-122.

Burge, T. (1988). "Individualism and self-knowledge." Journal of Philosophy **85**(November): 649-663.

Chalmers, D. J. (2003). The content and epistemology of phenomenal belief. Consciousness: New Philosophical Perspectives. Q. Smith and A. Jokic, Oxford University Press**:** 220--272.

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

Dretske, F. (1988). Explaining Behavior: Reasons in a World of Causes, MIT Press.

Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind, MIT Press.

Gertler, B. (2000). "The mechanics of self-knowledge." Philosophical Topics **28**(2): 125-146.

Goethals, G. R. and R. F. Reckman (1973). "The perception of consistency in attitudes." Journal of Experimental Social Psychology **9**(6): 491-501.

Heil, J. (1988). "Privileged access." Mind **98**(April): 238-251.

Lycan, W. G. (1996). Consciousness and Experience, Mit Press.

Nichols, S. and S. P. Stich (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds, Oxford University Press.

Nisbett, R. E. and T. D. Wilson (1977). "Telling more than we can know: Verbal reports on mental processes." Psychological Review **84**(3): 231-259.

Oddie, G. (2005). Value, Reality, and Desire, Clarendon Press.

Pitt, D. (2004). "The phenomenology of cognition, or, what is it like to think that _P_?" Philosophy and Phenomenological Research **69**(1): 1-36.

Putnam, H. (1975). "The meaning of 'meaning'." Minnesota Studies in the Philosophy of Science **7**: 131-193.

Stampe, D. W. (1987). "The authority of desire." Philosophical Review **96**(July): 335-381.

Stich, S. P. and S. Nichols (2002). Folk psychology. Encyclopedia of Cognitive Science. S. P. Stich and T. A. Warfield, Blackwell. **7:** 35-71.

Storms, M. D. and R. E. Nisbett (1970). "Insomnia and the attribution process." Journal of Personality and Social Psychology **16**(2): 319.

5 Reward and the Transparency of Desire

5.1 Introduction

According to some philosophers, belief is transparent in the following sense: I can come to know what I believe by attending outwards, not inwards. I can come to know whether I believe that P by attending to *P* itself.

This paper concerns the question of whether *desire* is transparent in the same sense. Can I come to know about my desires *transparently* – that is, can I come to know what I desire by attending outwards? Can I come to know whether I want to eat another slice of pizza by simply attending to the *pizza*? Or does desire introspection always involve some amount of inward attention?

A theory of desire transparency must offer an account of what it is about the world that we are attending to when we transparently know what we desire, and it must do so while meeting two desiderata. First, it must explain *weakness of the will* cases, where judgments of P's goodness come apart from self-ascriptions of a desire that P – for example, when the addicted smoker judges that smoking is not good, despite transparently knowing that they desire a cigarette. Second, it must be able to account for the *privileged access* that transparent desire introspection affords.

This paper will proceed as follows. In section 2, I give a brief overview of the transparency of belief, and highlight the role of *deliberation*. In section 3 I motivate the claim that desire introspection is transparent. In section 4 I discuss and critique extant accounts of desire transparency. I conclude that deliberation accounts of desire transparency are unable to explain *weakness of the will* cases, and instead propose an alternative according to which I come to know what I desire by directing introspective attention towards some representational content. In section 5 I argue that the best candidate for this representational content involves the property of *reward*, as understood by contemporary neuroscience of reinforcement learning. In section 6 I argue that the best way to account for the privileged access that transparent desire introspection affords is to hold that desires *just are* representations of reward. Finally, in section 7 I discuss a novel view of propositional attitudes that is suggested by the reward theory: that propositional

attitudes types are not individuated by functional role, as the standard view maintains, but merely by their representational content.

5.2 Transparency of Belief

Here is Gareth Evans with the classic presentation of the transparency of belief:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the questions whether p. (Evans 1982)

The transparency of belief is a certain *phenomenon* that we notice when we pay close attention to the introspection of our beliefs. What we notice is that, in many cases, we come to know that we believe P simply by attending to P itself. We do not look inwards, notice a mental state which we classify as a belief that P, and thus conclude that we believe that P. Rather, we do exactly what we do when we are trying to determine, not whether we *believe* P, but whether P is *true.* I will refer to a case of introspection that is transparent in this sense as one where we are introspecting using a *transparent method*.

There is a puzzle here: how can a transparent method result in the sort of *privileged* knowledge of our own beliefs that we take ourselves to have? Why should attending to the world give me any good reason to believe something about my mind?

The standard account holds that a transparent method can lead to privileged knowledge of whether we believe P because there is a transparent method which involves *deliberation* and *judgement.* When we are asked whether we believe that there will be a third world war, we deliberate over this question. And when we deliberate, we make it the case that we do or do not believe it. Thus, whatever we decide will correctly capture what we now believe. According to the standard picture, then, introspecting your belief that P using a transparent method works

because engaging in the transparent method - deliberating over or judging that P - *constitutes* whether or not you believe that P[24].

This leads to a general view: that transparent introspection of some mental state M involves engaging in a process of deliberation which makes it the case that I am or am not in M. Call this the deliberative account of transparency. The deliberative account is ideally suited to belief because the output of deliberation is belief. For this reason, as we will see, it is not suited to handle desires. But first: why think that desires are transparent at all?

5.3 Transparency of Desire and Weakness of the Will

Is knowledge of what I want transparent? Can I come to know what I want to eat for dinner by simply attending to the choices available? Or does self-knowledge of our desires always involve inward attention? Some authors have recently come out in favour of the view that desire introspection is transparent:

> ...often my eyes are ...‘‘directed outward – upon the world.’’ I can investigate my preferences by attending to the beer and the wine... (Byrne 2005)

> If asked whether I am happy or wishing that p, whether I prefer x to y, whether I am angry at or afraid of z, and so on, my attention would be directed at p, x and y, z, etc. (Bar-On 2004, p. 106)

According to *desire transparency*, there are cases where we know what we want transparently: sometimes, we know our desires by using a transparent method of attending to the world.

Suppose that you are sitting around, watching TV, when all of a sudden you feel the familiar pangs of hunger in your stomach, and thus realize that you are hungry. However, you are not yet sure what you want to eat. There are many options available. You know that you have the ingredients to make a tuna casserole or a chicken stir fry. Or, you could always just order pizza. How do you decide which option to choose?

---

[24] Such a view can certainly be found in Moran (2001), and arguably also in Byrne (2005)

Part of making this decision is determining what it is that you *want* to eat. What sort of food do you have the stronger desire for? This may not be the only consideration. You might be feeling lazy and thus not want to cook. Or you might be worried about money and thus not want to order out. Knowing what sort of food you desire is, however, in important part of deciding which option to choose. How do you go about doing this?

In my own experience, I find that what I do is simply think about the foods themselves. I will bring the pizza, the casserole and the stir fry before my minds eye, and consider their properties. I might think about their taste, or their texture, or their heartiness. Or I might simply think about them and see what 'strikes' me. In my experience, it is often the case that one of the potential objects of desire will just appear "bathed in a positive light". It will appear to me as being a certain way, a way that lets me know that I want it.

I am not sure whether my experience of introspecting my desires is typical. But if it is, then it seems that desire introspection is transparent after all. I come to know what I want to eat by attending to the potential objects of desire - and their putative properties - themselves, not inwards looking for some mental particular which I am able to classify as a desire.

Supposing that desire introspection is transparent, how should we explain it? What explains my ability to come to know what I want by simply attending to the world? Suppose that I eventually conclude that I desire pizza - which I do. According to the deliberative account, this is a result of engaging in a process of *deliberation* which *makes it the case* that I do or do not desire to eat pizza.

But what is it that I deliberate over? In the case of belief, I deliberated over whether the content of the belief was *true*. But this won't work for desires. For one, many desires do not have propositional contents. A desire for pizza simply has the content *pizza,* which is not capable of being true or false. Second, we often desire things that we know to be false - in fact, often *because* we know them to be false. I desire for there to be world peace, for instance. Finally, we need a way of introspectively distinguishing between beliefs and desires. If I judge that P is true, how do I know whether to conclude that I believe P or desire P?

Perhaps what we deliberate over is whether the potential object of the desire is *good*. When I want to know whether I desire pizza, I deliberate over whether *pizza is good*. I compare the strengths of my desire by comparing the degree to which I judge things to be good. On this view, then, I deliberate over what would be the *best* thing to eat: how *good* is it, all things considered, to eat pizza? Would it better than eating a casserole or a stir fry? Coming to the conclusion that it would be best to eat pizza, I conclude that I desire to eat pizza.

According to this account, we believe that we desire those things that we judge to be good. However, we often desire things that we do not think are good. Consider the *addicted smoker*, who desires a cigarette, despite knowing that smoking is bad. If she were to deliberate over the goodness of having another cigarette, she would conclude that it was bad, and thus come to falsely believe she does not desire a cigarette. Further, we often fail to desire things that we do think are good. Consider the *couch potato*, who knows that exercise is good, despite not desiring it. If he were to deliberate over the goodness of exercise, he would conclude that it was good, and thus come to falsely believe that he desires it. In both of these cases, deliberation over what is good is not a good guide to what one desires.

Sadly, our desires often do not line up with our beliefs about what is good. It seems that, in many cases, we cannot help what we desire. This is what caused Hume to say that it is "... not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." The idea here is that we cannot be held responsible for what we desire, because what we desire is not up to us. On the other hand, we can be held responsible for what we believe, because what we believe is up to us. In other words, while beliefs are responsive to our rational deliberation, our desires are not. And so there is no question of the reasonableness of our desires.

If this is right, then deliberation over what is good is not a very reliable guide to what we desire. Certainly, such deliberation does not *make it the case* that we now have that desire. Desires are fickle. They just aren't responsive to deliberation in the same way that beliefs are[25]. Call this the *weakness of the will* objection.

---

[25] Deliberation might result in, and thus afford self-knowledge of, a new *instrumental* desire. For example, if I desire world peace, and know that I do, and my deliberation leads me to conclude that the best way to achieve world peace

5.4 Accounts of Desire Transparency

5.4.1 Byrne's Desire Defeaters

Byrne (2011) defends a deliberative account of desire transparency that is supposed to deal with the weakness of the will objection. According to Byrne, transparent introspection of my desires involves attending to the 'desirability' of things, where something is desirable in the sense of having "the qualities which cause a thing to be desired: Pleasant, delectable, choice, excellent, goodly." In particular, Byrne holds that transparent desire introspection involves following an epistemic rule:

> (DES) If P is a desirable option, believe that I desire P.

As it stands, Byrne's view cannot yet handle the weakness of the will objection above. The couch potato knows that they do not want to exercise, despite also knowing that exercise is desirable.

To deal with these sorts of worries, Byrne introduces the idea of a 'desire defeater'. In most cases, DES is a good rule: it leads you to form true beliefs about what you desire. However, DES is a *defeatable* rule: there is possible evidence that can 'block' the inference from P being desirable to my desiring P. According to Byrne, knowledge of what I *intend* to do is what defeats such an inference. The couch potato may judge that exercise is more desirable than sitting on the couch all day. But he nevertheless does not conclude that he desires to exercise, because he knows that he *intends* to stay on the couch. More precisely, according to Byrne:

> Suppose one knows that φing is a desirable option, and considers the question of whether one wants to φ. One will not follow DES and conclude one wants to φ, if one believes (a) that one intends to ψ, (b) that ψing is incompatible with φing, and (c) that ψing is neither desirable nor better overall than φing. (pg. 26)

---

is to abolish capitalism, then I might come to desire the absolution of capitalism. But such deliberation only results in a new desire because I was already aware of my desire for world peace, and thus cannot account for awareness of this desire.

So, the couch potato does not follow DES and conclude that he desires to exercise because he believes that (a) he intends to stay on the couch and (b) staying on the couch is incompatible with exercising (right now) and (c) staying on the couch is neither desirable nor better overall than exercising (right now).

Byrne's view is thus a modified version of the deliberative account of transparency. According to Byrne, transparent introspection of my desire that P involves deliberating over the goodness (or 'desirability') of P. But once I conclude that P is good (or desirable), I do not always conclude that I desire P. Sometimes, additional evidence blocks this inference.

Ashwell (2013) raises two objections to Byrne's view. First: it seems that the couch potato need not judge that staying on the couch is neither desirable nor better overall than exercising. The couch potato may very well judge that staying on the couch is just as desirable as exercising. In such a case, condition (c) will not hold. What we have are two incompatible options, both of which are judged to be desirable. Byrne's account does not appear to have the resources to explain why we judge ourselves to desire one but not the other. To make matters worse, there are cases where we desire P, and not Q, even though we judge Q to be more desirable than P. The plight of the addicted smoker is precisely that she knows that quitting smoking is more desirable than another cigarette. Nevertheless, she desires - and knows that she desires - another cigarette.

Second, Ashwell notes that Byrne's account requires that self-knowledge of our intentions is *prior to* of self-knowledge of our desires. But this is not plausible, she argues, especially in the sort of case Byrne has in mind. Suppose that I judge that exercise is desirable, yet fail to conclude that I desire exercise because I know that I intend to stay on the couch. If I do not think that staying on the couch is desirable, then on what basis do I know that I intend to do so? And if I do think that staying on the couch is desirable (perhaps just as desirable as exercising), then why do I conclude that I intend to stay on the couch, and not exercise? If the addicted smoker judges quitting smoking to be more desirable than having another cigarette, then on what basis does she know that she intends to have another cigarette?

Byrne's epistemic rules account of desire transparency, even with the help of desire defeaters, does not seem to be able to adequately respond to the weakness of the will objections. Further, it does not seem that his account can be modified in order to do so. There just do not seem to be

any plausible desire defeaters that can be known independently of our desires. The deliberative account of transparency cannot be tweaked to handle desires, because desires just aren't responsive to deliberation in the right way.

5.4.2 An Alternative to Deliberation

The deliberative account of transparency is by far the most prevalent in the literature. On the deliberative account, transparent introspection always involves deliberation about how the world is. The deliberative account is ideally suited for beliefs, because beliefs are the outputs of deliberation. This is not so in the case of desire. What I desire and how I judge the world to be, sadly, often come apart. If desire introspection is indeed transparent, then it looks like we need a non-deliberative model of transparency.

Fortunately, the literature on phenomenal experience contains just such a model. It is often said that phenomenal experience is transparent, and this is meant in the following sense: that when we try to introspect the phenomenal properties of our experiences - what it is like to have them - we find that we are only able to attend to the putative properties of external objects (e.g. Tye 2002.) Like the transparency of belief and desire, the transparency of phenomenal experience is taken to be a datum that needs to be explained. How can it be that I am able to know what phenomenal properties are instantiated in my experience by simply attending to the world?

This is typically explained by a version of *representationalism* about phenomenal experience, the strongest version of which holds that phenomenal properties just are represented properties. On this view, the phenomenal redness of your perceptual experience of a red apple is just the red that your experiences represent the apple as having. I can come to know that my experience instantiates phenomenal redness because I am able to introspectively attend to the redness that the apple is represented as having. On this picture, transparent introspection of phenomenal properties does not involve deliberation, but directing introspective attention towards how the world is *represented* as being.

This suggests a non-deliberative account of desire transparency. I come to know what I desire by directing introspective attention towards an aspect of how my mind represents the world as being. On this view, for every desire, there is some represented content C such that when I

introspectively notice C, I self-attribute that desire. I come to know that I desire chocolate cake, not by deliberating over the features of chocolate cake, but looking to see how chocolate cake is represented as being. Call this the *content* account of desire transparency[26].

One might object that attending to representational contents, which are mental properties, is incompatible with transparency. Central to the idea of transparency is that our attention is focused outwards at the world, not inwards. Doesn't attending to representational properties necessarily involve "inwards" attention?

However, while representational properties are properties of mental states, they are also *directed* outwards at the world. To attend to the representational content of a mental state is to attend to an aspect of your mind that involves how it takes the world to be. Thus, attending to representational contents does involve directing your attention *outwards* in the relevant sense[27].

5.4.3 Ashwell's Appearances of Goodness

Ashwell (2013) proposes a view that, on my reading, is best understood as a content account of desire transparency. Ashwell distinguishes between *judgments* of goodness and *appearances* of goodness. According to her view, when we want to know whether we desire that P, we do not judge - or deliberate - whether P *is* good, but check to see whether P *appears* good. I come to know that I desire P when I notice that P appears good to me.

This deals with the weakness of the will cases in a simple and compelling way. While there are certainly cases where we desire P without judging that it *is* good, it's plausible that P still *appears* to be good to us in these cases. Take the case of the addicted smoker. She knows that

---

[26] In this paper I am only exploring the application of a content account to *desire* transparency. But there is a case to be made that a content account can help explain *belief* transparency as well. There are (at least) two reasons why one might favor a content account over a deliberative account of belief transparency. First, transparent belief introspection often does not seem *phenomenologically* to involve any sort of deliberation. Second, transparent introspection often affords self-knowledge of beliefs that I have no need to deliberate over (such as my belief that I live in Canada.)

[27] At the end of the day, it might be a terminological issue as to whether to consider such views 'transparent'. One might prefer to think of such views as showing how a process that appears to involve outward attention actually merely involves attention to mental representations: that is, as explaining away the appearance of transparency.

smoking is bad for her and is trying to quit. She also knows that quitting smoking is the more desirable option. But she can't help but crave a cigarette when she sees her co-workers go out for a smoke. When she craves a cigarette, she thinks about smoking, and the act of smoking is, in her mind, bathed in a positive light. Despite her knowing better, smoking still appears to be good to her. Likewise, the couch potato knows that exercise is good for him, and is much more desirable than sitting on the couch all day. Nevertheless, exercise just doesn't appear good to me, while sitting on the couch all day does.

Unfortunately, it is not quite clear exactly what the difference between judgments and appearances of goodness amounts to. Ashwell relies heavily on an analogy to perceptual experience. She says that perceptual experience and judgments about how things are around us come apart in two ways. First, we might judge that something is in front of us, even though we cannot see it (it might be behind a tree, for example.) Second, we might judge that things are not how we perceive them to be. Consider the perceptual illusion of partially submerging a straight stick on a glass of water. If you are familiar with the illusion (and therefore know it to be an illusion) you will not be fooled: you will believe that the stick is straight. Nevertheless, the stick will appear to be bent. It will appear to be as bent as it did when you were not aware that it was an illusion. Your beliefs about the true nature of the stick has no effect on how the stick appears to be.

It is clear, then, that Ashwell understands the distinction between a judgment and an appearance of goodness to be roughly on par with the distinction between judgment and perceptual experience. For something to appear good to you is for you to 'see' that it is good. However, it is not clear what *this* distinction amounts to. What is the difference between perception and judgment? What does it mean to 'see' that something is good?

This just raises the question again: what is it for P to appear to me to be good, as opposed to simply judging that P is good? What distinguishes an appearance of good from a judgment of good? As I see it, there are two approaches one might take. On the *content* approach, appearances and judgments of goodness differ in their representational content. My perception of the stick as being bent and a belief that the stick is bent have different contents. With regard to goodness, perhaps an appearance of goodness represents that *P appears to be good,* while a

judgment of goodness represents that *P is good*. The second approach holds that appearances and judgments of goodness both have the same content: *P is good*. What individuates them is some non-representational property of the state. The most plausible candidate is *functional role*. For example, it is sometimes said that what distinguishes perceptual representations from judgments is that the former are "informationally encapsulated". By this it is meant that higher-cognitive processes are unable to affect the content of perceptual representations. Intuitively, what we see is not up to us, but what we judge is. Perhaps something similar can be said for the difference between appearances and judgments of goodness. Let us call the functional role supposedly characteristic of appearances the 'appearance-role'. On the *functional role* approach, then, appearances of goodness are mental states which (a) represent that *P is good* and (b) play the appearance-role.

Let us first consider this functional role approach. The appearance view of desire transparency says that in order to figure out whether you desire P, you do not deliberate over P's goodness. Rather, you must consider whether P *appears* good. According to the functional role approach to appearances of goodness, P appears good to me just in case I am in a state which (a) represents that *P is good* and (b) plays the appearance-role.

The problem with this approach is that it abandons the motivation for the transparency of desire. The idea was that when we want to know what we desire, we turn our attention towards the world, not our minds. Transparency approaches to self-knowledge are supposed to be an alternative to the traditional picture where we engage in some sort of 'internal scanning' of our minds[28]. But, according to the functional role approach, transparent desire introspection requires that we direct our attention towards our minds in order to determine whether or not P appears good. We must not only attend to the content *pizza is good*, but also to the functional role of the state which represents this[29]. Why not just say that we determine whether we desire that P by considering whether we are in a mental state which represents P and has the functional role

---

[28] For example: Nichols and Stich (2002; 2003), Armstrong (1968) and Lycan (1996).

[29] One might object by pointing out that we *are* able to distinguish between perceptions and beliefs without inspecting the functional role of the state doing the representing. I think this is exactly right, and suggests that the functional role approach fails to account for the difference between perception and belief as well.

characteristic of desire? This is the traditional view of desire introspection that transparency is supposed to provide an alternative to. On the modular view of 'appears', the transparency of desire becomes nothing more than the traditional view in another guise.

If the right way of thinking about transparent desire introspection is in terms of appearances of goodness, then we require an account of what appearances of goodness are that is consistent with the central motivation of transparency views. We must be able to make sense of what it is to direct our attention towards P *appearing* to be good, despite our only attending outwards at the world. What the above suggests is that functional role approaches cannot accomplish this. Rather, it suggests some version of the *content* approach to appearances of goodness. When I direct my attention towards P appearing good, I am attending to a way that the world is: namely, it being such that P appears good. But this is just to say that an appearance of goodness *represents* that some part of the world appears good.

Let us return, then, to the suggestion that for P to appear good to me is for me to represent that *P appears good*. According to this approach, transparent self-knowledge of our desires is accomplished by introspecting *appearance* contents. Thus, I come to know that I want to eat pizza for dinner by directing introspective attention towards the content *Pizza appears good.*

But what sort of content is this? Suppose that you see that your friend has a furrowed brow and their face is red. It would be natural to say, in this case, that your friend appears to be angry. Here, to appear to be angry is to have certain perceptible features: a furrowed brow and a red face. We might say, then, that for someone to appear angry to me is for me to represent them as having these features. Further, it seems plausible that the reason we take these features of your friend to constitute their appearing angry is that these features *mark* the presence of anger. There is a causal connection between appearing angry and being angry.

I think this is the most plausible way of developing the content approach to appearances of goodness. On this view, for P to appear good to me is for me to represent that P has some *other* property which marks the presence of goodness. Then, I come to know that I desire P by directing introspective attention at a content that ascribes this marker of goodness to P. This is a version of the content account of desire transparency. Transparent desire introspection is accomplished by simply directing introspective attention to the right represented contents: in this

case, contents that involve a marker of goodness. This approach clearly avoids the weakness of the will counter-examples. Just as I might see that my friend has a furrowed brow and a red face but not believe him be angry, the addicted smoker might represent exercise as having a marker of goodness without believing it to be good.

In addition to avoiding the weakness of the will counter-examples, this proposal has independent motivations. It is clear that, if desire introspection is indeed transparent, then the relevant contents have to be somehow related to goodness. But it is implausible that goodness itself is directly involved in the content relevant to transparent desire introspection. GOODNESS is an abstract concept that enters into cognition only at the highest level. It seems divorced from low level, perceptual/motor engagement with the world. But this is precisely the level that we should expect to be relevant to desires. Desires are those states that are, ultimately, responsible for the causation of behaviour. They are what motivate us to act in the world. As such, any sort of cognitive agent tasked with avoiding danger and seeking food, shelter and sex must possess the capacity to desire. This is not to say that simpler cognitive agents are capable of self-knowledge of desires. Nevertheless, we should expect that the representational content relevant to such self-knowledge be the product of processes and capacities shared by simpler agents. Rats, for example, plausibly are not able to represent the abstract concept GOODNESS. But, as psychological and neuroscientific research suggests, they can represent the *value of a reward*. Further, the representation of the value of a reward can plausibly be thought of as marking the presence of something that is good for an organism. Food is rewarding for organisms because it is good for them - that is, food being good for an organism *explains why* they have evolved to treat food as a reward. What this suggests, then, is that transparent desire introspection more plausibly involves the search for what is rewarding, rather than what is good.

My proposal is that if we want to better understand what it is to represent something to *appear* good, then we should look to the growing literature on the neuroscience of learning. As we will see, there is a growing consensus that one of the primary functions of orbitofrontal cortex (OFC) is to keep track of the expected value, or reward, of particular outcomes.

5.5 Desire and Reward

5.5.1 The Neuroscience of Reward

Neuroscientists and psychologists distinguish between at least two different kinds of learning that facilitate making such decisions, and which have been shown to be functionally and neurologically distinct. The first is classical *Pavlovian conditioning*, or stimulus-response learning. Here, agents come to directly associate a previously motivationally inert stimulus with a particular action. In Pavlov's classic 1927 study, dogs are given food - which they salivate in response to - after hearing a tone. After repeated training, they begin to associate the tone with salivation behaviour, such that they salivate whenever they hear the tone, regardless of whether food was presented. Such learning develops automatic behaviour in response to stimuli. It is in this sense that stimulus-response conditioning develops habits, or mere reflexes.

The second kind of learning is *instrumental conditioning*, which involves altering an agent's behaviour by pairing a rewarding stimulus with spontaneous behaviour. Skinner (1938) placed hungry rats in a box with a lever that would release a food pellet. The rats would sometimes accidentally press the lever, which resulted in them receiving a food reward. Eventually, the rats associated lever-pressing behaviour with receiving the food reward, and thus learned that they could press the lever *in order to* receive a food reward.

This sort of learning involves more than the agent simply associating a stimulus with a particular action. Rather, the agent must make two associations: one between an action and an *outcome*, and the other between an outcome and a *reward*. Thus, the rat learns that pressing the lever will lead to food pellet dropping in the cage, and that the food pellet is rewarding.

'Reward' can be understood functionally to mean anything that can be used to reinforce voluntary behaviour. In this sense, 'reward' is to be understood in terms of the role that it plays in a theory of instrumental conditioning. The food pellet is considered rewarding for the rat because the rat alters its behavioural dispositions in order to get it. Further, we can speak of degrees of rewardingness in terms of how strong of an effect the reward has on altering an agent's behaviour. When given the choice between a food pellet and water, if the rat acts preferentially towards the food pellet, we can say that the food pellet is *more* rewarding - all else being equal - to the rat. Finally, the rewardingness of something depends on the internal state of

the agent. A satiated but thirsty rat will choose the water over the food pellet. Thus, satiety levels partially determine the rewardingness of the food pellet.

An important difference between Pavlovian and instrumental conditioning concerns the kinds of behaviours that are learned (Schroeder 2004, pg. 44-45). While Pavlovian conditioning involves reflexive or involuntary actions, instrumental conditions involves spontaneous or voluntary actions. The dogs in Pavlov's study couldn't help but salivate in response to hearing the tone, while the rats' lever pressing behaviour was spontaneous and goal-directed: they pressed the lever in order to get food. Thus, only instrumental conditioning involves the reinforcement of voluntary behaviour, which is to say that only instrumental conditioning involves rewards.

What this demonstrates is that simple Pavlovian conditioning and goal-directed learning are functionally distinct. In order to explain the former, we need only posit direct causal relationships between the apprehension of a certain stimulus (e.g. hearing the tone) and the performance of a certain behaviour (e.g. salivation). In order to explain goal-directed behaviour, in contrast, we must posit two distinct capacities. The first is that of associating a certain behaviour (e.g. pressing a level) with an expected outcome (e.g. acquiring food.) The second is that of associating an expected outcome (e.g. acquiring food) with the rewardingness of that outcome.

What arises, then, is a picture of goal-directed learning that crucially involves the ability to encode the rewardingness of certain outcomes. An agent must be able to categorize an outcome as rewarding in order for it to play the right role in the production of behaviour. Food motivates the rat to alter its voluntary behaviour *because* the rat's cognitive system is able to treat food as a reward.

Earlier we said that we can understand 'reward' functionally as anything that can be used to reinforce voluntary behaviour. However, now we see that 'reward' can also be understood as something that an agent treats an outcome as being. Agents must be able to keep track of whether, and to what degree, certain outcomes are rewarding. In my view, the best way of understanding these capacities is in terms of *representation.* The rat *represents* that the food pellet is rewarding, and it is this representation of rewardingness that explains the role that the food plays in altering the rat's voluntary behaviour. On this view, for O to be rewarding to

degree D for S is for S to represent that O is rewarding to degree D. This representation in turn plays a role in making S act preferentially towards seeking out O to degree D. Thus, 'rewarding for X' can be analyzed in terms of X representing rewardingness. But this should not be taken to be an analysis of the property of rewardingness itself. Rather, it is best to think of rewardingness as a *primitive* property that need not actually be instantiated. Rewardingness is simply a marker for goodness: it is a flag that organisms attach to objects in order to maximize their own well being.

Thus, the functional characterization of a reward – something that can be used to reinforce voluntary behaviour – is only part of the story. What is required is an explanation of *how* rewards understood in this way are able to reinforce behaviour. And the explanation being offered here (one that, as we will see shortly, is dominant in the neuroscience literature) is that certain outcomes reinforce behaviour *because* the agent is able to represents them as being rewarding. This representation in turn plays a role, under normal circumstances, in altering its behavioural dispositions in order to get the outcome.

This view has empirical consequences that can be tested. If it correct, then we should expect to find an area of the brain that carries information about the rewardingness of outcomes. Such an area would exhibit activation that tracks self-reports of rewardingness for particular outcomes. Changes in the activity in this area would track changes in, for example, taste aversion and satiety. Ideally, we should be able to predict which action an agent is to perform in a choice task by analyzing the activation in this area.

Contemporary research in the neuroscience of decision making strongly suggests that the orbitofrontal cortex (OFC) is such an area. In a review article on the function of the OFC, it is asserted that:

> Today, through a remarkable convergence of studies conducted in species ranging from rats to humans, OFC is widely conceived as a place where the 'value' of things is represented in the brain. (Mainen and Kepecs 2009)

This conclusion is arrived at through the analysis of a variety of experiments. Neurons in the OFC of monkeys respond to rewarding substances. The response of these neurons tracks changes

in satiety level and associative learning. For example, these neurons will gradually stop responding to certain foods as the monkey is satiated (Rolls 2000). In contrast, the representation of tastes in the primary taste cortex is not modulated by satiety level (Yaxley, Rolls et al. 1990). Interestingly, monkeys will work for electrical stimulation in the OFC, but only if they are hungry (Mora, Avrith et al. 1979).

In addition to taste, the OFC has also been found to respond to pleasant and painful touch sensations, more so than to affectively neutral touch sensations (Francis, Rolls et al. 1999). In another review article on the role of the OFC in the representation of value, the authors conclude:

> An implication of these findings is that the orbitofrontal cortex may contribute to decision-making by representing on a continuous scale the value of each reward with, as shown by the single neuron neurophysiology, different subsets of neurons for each different particular reward. It is of course essential to represent each reward separately, in order to make decisions about and between rewards, and separate representations … of different rewards are present in the orbitofrontal cortex. (Rolls and Grabenhorst 2008, pg. 234).

Additional support for the claim that reward values are represented in the brain comes from work on the dopamine system, primarily the ventral tegmental area (VTA) and the pars compacta of the substantial nigra (SNpc) (Schultz, Tremblay et al. 2000). The dopamine-releasing neurons in this area have a baseline level of activity. When an agent is unexpectedly presented with a rewarding stimulus (such as food), the neurons briefly fire more quickly. However, when a rewarding stimulus is presented predictably (e.g. if it is always presented after a flashing light after conditioning), there is no increase in the firing rate of these neurons. What this suggests is that the firing-rate of these neurons encode the difference between the reward value that is predicted and the reward value that is actually received. In order to compare the difference between predicted and actual reward value, the system must be able to *represent* both the predicted and actual reward value received.

To sum up: in order to maximize the benefit a cognitive agent gets from their environment, a cognitive agent must be able to keep track of the expected value, or rewardingness, of particular outcomes. There is a growing consensus among psychologists and neuroscientists that it is the

function of the OFC to keep track of this sort of information, and that the representation of rewardingness plays an essential role in goal-directed behaviour.

5.5.2 The Reward Content Account of Desire Transparency

Ashwell suggested that we come to know that we desire P by attending to Ps *appearing* to be good. I argued that the best way to make sense of an appearance of goodness is in terms of the representation of a marker of goodness. The task was, then, was to find a property the representation of which could plausibly be thought of as marking the presence of something good for the organism.

What the above discussion about the neuroscience of reinforcement learning demonstrates is that there is such a property. There is good reason to believe that the orbitofrontal cortex functions to keep track of the reward value of particular outcomes. The representation of this property plays an important role in the production of goal-directed behaviour. Plausibly, such a system evolved in order for an organism to keep track of and therefore maximize the expected utility of its actions. Thus, there is good reason to think that the representation of the reward value of an action tracks what is *good* for an organism: we find sugary foods rewarding *because* they are a good thing for us to seek out (or at least were in our evolutionary past.) Further, reward representation comes apart from thoughts about what is good or bad. We can't help but find sugary foods rewarding, even though we know that too much of them is bad for us.

In order to figure out whether we desire P, we simply check to see if - and to what degree - we represent that *P is rewarding*. When I am trying to figure out whether I should eat pizza or Indian food for dinner, I reflect on these objects, which activates states in the orbitofrontal cortex that carry information about the their rewardingness. Activation of these states makes their contents accessible to introspection. I am thus able to attend to the degree of rewardingness that I represent these objects as having. When I want to know what I occurrently desire, I simply check to see what is represented as being rewarding, and to what degree. Call this the *reward content account of desire transparency* (or simply the reward content account.)

This view should not be understood as an alternative to Ashwell's, but as a development of it. It is true that transparent desire introspection involves attending to appearances of goodness. I have

simply fleshed out what it is for something to appear good to you: it is for you to represent it as being rewarding.

A benefit of content accounts of desire transparency is that they allow contents to play a role in transparent introspection that may not be capable of being the content of judgements. On deliberative accounts, I come to know that, for example, I believe that P as a result of *judging* that P. To extend such a view to desires, we would need to say that I come to know that I desire P as a result of making some judgment about P. Thus, on these views, the only sorts of content that could be relevant to the transparent introspection of desire are those contents that are capable of being judged. The reward theory has no such requirement. It could turn out that the value of a reward is something that thoughts are not able to represent. In this sense, rewardingness might be more like the content of sensory representations. Perhaps we cannot have a thought involving the precise shade of red that we see; likewise, perhaps we cannot have a thought involving the degree to which something is rewarding.

This view does not suffer from the weakness of the will objection. The representation of reward is distinct from the representation of goodness. The addicted smoker judges that she desires a cigarette not because she notices the content *cigarettes are good*, but the content *cigarettes are rewarding*. As we have seen, reward is represented in the orbitofrontal cortex all sorts of cognitive agents, including monkeys and rats. In contrast, since goodness is an abstract concept, it is plausible that perhaps only humans are able to represent it. It is this abstract sense of good that generates the problem cases for desire transparency. The sense in which smoking is bad, and exercise is good, is much different from the sort of information that a rat's brain keeps track of in order to decide what lever to press. Thus, while the addicted smoker knows that smoking is bad, she nevertheless less represents them as being rewarding. And while the couch potato knows that exercise is good, he nevertheless fails to represent it as being rewarding.

There are a few possible objections to the reward content account. First: experimental evidence has been given for the claim that the OFC *carries information* or *keeps track* of the reward value of various outcomes. But why should we think that this means there are *representations* of rewardingness? The reason is that, according to most 'naturalistic' accounts of representation, representation is a kind of information carrying (e.g. Fodor 1987, Dretske 1981 and Millikan

1984). Thus, if an area of the brain carries information about the rewardingness of an outcome, then it is plausible, according to these views, that they are representing it.

In addition to agreeing with philosophical accounts of representation, introducing representations of rewards coheres with general cognitive scientific methodology. It is a common explanatory strategy to introduce representations as 'stand-ins' for aspects of an agent's environment that it must keep track of (Bechtel 1998). For example, Tolman (1948) explained a rat's ability to navigate a maze by introducing 'cognitive maps' - spatial representations of the maze that allow the rat to plan and coordinate its behaviour.

Second: one might think that we can only introspectively attend to the contents of *conscious* mental states. The reward content account requires that, at least some of the time, representation of reward in the orbitofrontal cortex are conscious. Why think that they are? The reason is empirical: it has been found in numerous studies that subjective, conscious ratings of the pleasantness of a stimuli are correlated with activation in the OFC (e.g. Rolls 2007). The best explanation of this correlation is that some reward representations in the OFC are consciously accessible. However, the reward content account does not require that *all* reward representations are conscious. It only requires that, when I engage in transparent introspection of my desires, the reward representations that correspond to those desires are accessible to introspection. Relatedly, there is no requirement that reward representations are capable of being conscious in all kinds of organisms. Rats, for example, are able to represent rewardingness, but such representations are plausibly not conscious. Further, there is no requirement that the *process* of comparing the rewardingness of various outcomes be conscious.

Third: according to the reward content account, transparent self-knowledge of what we want is the result of introspecting the content of reward representations. But it looks like reward representation is more closely related to perceptual stimuli: taste and touch, in particular. This might account for our self-knowledge of what we want to eat, but what about self-knowledge of my desire to join a band, or my desire for world-peace? Is it plausible that my brain carries information about the reward value of these more abstract objects of desire?

This is a powerful objection, but I think it can be dealt with by way of the common distinction between intrinsic and instrumental desires. I desire money, but only because I desire things that I

believe money can get me. In contrast, I desire a rare steak for its own sake. In general, S's desire for P is intrinsic if does not entail that S has any other desire. In contrast, S has an instrumental desire for P only when (a) S believes that P will lead to O and (b) S desires O (either instrumentally or intrinsically.)[30] [31]

Many of our desires are plausibly instrumental. I desire world peace because I don't want people to suffer, and I believe that world peace will reduce others' suffering. I might desire to join a band because I like playing music, and believe that being in a band is the best way to do this as much as possible. There are two possibilities for how we could come to know about such instrumental desires according to the reward theory. The first possibility is that only intrinsic desires, but not instrumental desires, correspond to what is represented as being rewarding in the OFC. In order to know what we intrinsically desire, we simply attend to the contents of our reward representations. Self-knowledge of instrumental desires is accomplished by a two-step process. After determining what we intrinsically desire, we then engage in transparent introspection of our means-end beliefs. For example, I might discover that I desire to join a band after discovering (a) that playing music is rewarding and (b) that I believe that joining a band is the best way to do this as much as possible[32].

A second possibility, more speculative than the first, is that some intuitively instrumental desires correspond to reward representations in the OFC. On this view, the transparent introspection of all desires is simply a matter of determining what we represent to be rewarding. There is some preliminary experimental work in support of such a view. In one set of studies, it was found that

---

[30] One can desire P both intrinsically and instrumentally. I desire a rare steak for its own sake; but I might also at the same time desire it instrumentally, because I desire the cessation of hunger and know that eating the steak will bring this about.

[31] There are further conditions on having an instrumental desire. For example, I don't instrumentally desire the destruction of the world if I don't want to go work tomorrow.

[32] A related possibility is that many of the instrumental desires we attribute to ourselves are really beliefs about what we desire, that may end up being false. Perhaps we have been told that we ought to desire professional success, but we don't *really* desire it: we can't trace *professional success* back to something we find rewarding through a chain of means-ends beliefs. Maybe if we really paid attention to what we find rewarding and our means-ends beliefs, we would realize that professional success is just not important to us. Whether such cases count as genuine desires or not may end up being terminological.

different parts of the OFC process different kinds of reward representations (Sescousse, Redouté et al. 2010). Erotic stimuli were processed by a phylogenetically older region (posterior lateral OFC), while monetary stimuli were processed by a phylogenetically recent region (anterior lateral OFC). What this suggests is that the OFC may be responsible for representing the reward value of all sorts of stimuli, even those as abstract as money. If this is right, then it is possible that self-knowledge of even the most intuitively instrumental desires to be accomplished by simply introspecting the contents of reward representations.

One might object that this view suffers from a close relative to the weakness of will problem. If desires are just representations that something is rewarding, then it seems that I am able to come to desire something by simply convincing myself that it is rewarding. Thus the couch potato can come to desire exercise by convincing himself that it is rewarding; the addicted smoker can cease their desire for a cigarette by telling herself that it is not rewarding. This objection, however, relies on the claim that the sort of content that is distinctive of desires, which I have been calling reward, can be deployed into the content of beliefs. It seems to be a common idea in the philosophy of mind that the

Whichever of these possibilities ends up being correct, I think that it is clear that the reward content account has the resources to account for transparent self-knowledge of all of our desires, even our most abstract, instrumental ones.

5.6 The Puzzle of Desire Transparency

Recall the puzzle of the transparency of belief. How could attending to the world possibly give me privileged access to what I believe? This puzzle was solved by the deliberative account of belief transparency, according to which I come to know that I believe P by deliberating over the truth of P. Since the output of deliberation is belief, I will *ipso facto* believe whatever I conclude. Thus, if I simply self-attribute a belief about whatever the outcome of my deliberation is, I am guaranteed to be right.

However, consider the corresponding puzzle of the transparency of desire. How could attending to the world possibly give me privileged access to what I want? According to the content account of desire transparency, I come to know that I desire P by directing introspective attention towards

some represented content. I defended the view that the relevant contents involve the primitive property of reward. But this does not yet solve the puzzle.

According to the standard picture, beliefs and desires have the same contents. I can believe that it is going to rain tomorrow, or I could desire this. My belief and my desire, on this picture, both represent that *it is going to rain tomorrow*. The difference between beliefs and desires is not what they represent, but the sort of relation we bear to this content. This relation, in turn, is typically understood to consist in the functional role of the state doing the representing. A belief that it is going to rain tomorrow is a state which represents *it is going to rain tomorrow* and plays the belief-role; mutatis mutandis for desire.

So, how could noticing the content *it raining tomorrow is rewarding* possibly give me privileged access to my desire that it rain tomorrow? The content I notice is not the content of the desire self-attributed; rather it is the content of a completely distinct mental state.

This creates something of a problem for the proponent of desire transparency. It seems like deliberative accounts of desire transparency are hopeless, because desires just aren't responsive to deliberation in the same way beliefs are. But then if we adopt a content account of desire transparency, it seems like we must give up on privileged access. After all, simply noticing the content *P* is not enough to self-attribute a desire that P, because it could be the content of any number of different attitudes. Any content that could plausibly give me reason to self-attribute a desire that P - such as *P is rewarding* - cannot, according to the standard picture, be the content of my desire itself. So, any plausible version of a content account must hold that we come to know what we desire by directing introspective attention at the content of a state which is not my desire itself. How could this possibly give me privileged access to what I desire?

Here, then, is an argument against the possibility of any theory according to which we can have privileged access to our desires transparently:

> P1. Attending to the content of my desire that P cannot give me privileged access to my desire that P.
>
> P2. Attending to a content that is not the content of my desire that P cannot give me privileged access to my desire that P.

C1. Therefore, there are no contents that I can attend to that would give me privileged access to my desire that P.

P3. But attending to contents is the only way to account for desire transparency

C2. Therefore, there can be no theory of desire transparency compatible with privileged access.

How might one go about responding to this argument? A natural suggestion is to deny P2. The motivation for premise 3 is the idea that attending to the content of some mental state M_1 cannot afford one privileged access to some metaphysically distinct mental state M_2. Attending to the content *P is rewarding* cannot afford me privileged access to some other state which represents *P*. In response, one can go reliabilist about privileged access. Privileged access to a token mental state M does not consist in having some sort of direct acquaintance with M. Rather, it consists merely in being able to form beliefs about M via some especially reliable process. Thus, one could hold that, while the state which represents that *P is rewarding* is metaphysically distinct from my desire that P, they are nevertheless very strongly causally connected, such that if you are in the former, you are almost certainly in the latter. Directing introspective attention towards the content *P is rewarding*, then, is an extremely reliable process for forming beliefs about what you desire.

In fact, one could go further and hold that it is part of the functional role characteristic of desire to bear certain causal connections to representations of rewardingness. Ashwell (2013) appears to support this sort of move:

> I suspect that if this account is fleshed out, the most plausible way to develop the metaphysics will involve accepting that there is a causal connection from your desires to value appearances. Wanting things makes you see them in a certain light, and this is how you introspectively know what you want. (pg. 255)

However, such a move must give up on the *directness* of self-knowledge of our desires. On this reliabilist picture, we do not come to know that we desire P because of our direct access to the desire itself, but rather because of our direct access to the non-desire state of finding P to be rewarding. This constitutes self-knowledge of our desire that P only because it is reliably

connected with the non-desire state. Whether one is happy with such a move depends on what one takes the demands of privileged access to be.

In my view, the simplest way out of the puzzle is to hold that my desire that P is *identical* to my representation that P is rewarding. On this proposal, a desire that P does not merely represent *P*, but *P is rewarding*. Desires are representations of rewardingness. The strength of my desire for P is the degree to which P is represented as being rewarding. Thus, directing introspective attention at the content *P is rewarding* affords us privileged access to our desires because that just is the content of my desires. If I notice the content *P is rewarding*, then I *ipso facto* desire P (to some degree)[33]. On this view, transparent desire introspection is direct.

The proponent of desire transparency, then, has two options. The first is to reject that transparent introspection affords us *direct* access to our desires. At best, it affords us indirect access to our desires states through states that are merely correlated with them. The second is to reject that desires and reward representations are distinct after all. On this view, since transparent introspection affords privileged access to reward representations, it *ipso facto* affords privileged access to our desires.

5.7 Pure Attitude Representationalism

I have argued that the transparent desire introspection is accomplished by attending to representations of reward. I then argued that the best way to explain how this could result in privileged access to our desires is by holding that desire just are representations of reward. On this view, my desire for it to rain tomorrow is just a mental state which represents that *it raining tomorrow is rewarding.*

As we've seen, the standard view of beliefs and desires holds that they both have the same content, and they are distinguished by their functional roles. On this view, my desire for it to rain tomorrow represents that *it will rain tomorrow*. It is distinguished from my belief that it will rain

[33] On my use of the term, one can be said to desire P to some degree even if, for example, one desires Q to a greater degree and believes that Q and P are incompatible. Or, one can be said to desire P even if one believes that P leads to Q and one desires ~Q much more than P. Statements of desire are thus not 'all things considered.'

tomorrow by what it *does*. My desire for it to rain tomorrow – but not my belief – will cause me to be happy if it does end up raining. In contrast, my belief that it rains tomorrow – but not my desire – might cause me to cancel my picnic.

The view of desire that I am defending here thus rejects the standard view. On my view, the difference between beliefs and desires is essentially a difference in their representational contents. Beliefs and desires do different things in virtue of their different contents, not because that is what *individuates* them. Call this view Pure Attitude Representationalism (PAR).

One advantage of PAR, as explored in this paper, is that it allows for a simple view of transparent desire introspection. We come to know that we have a particular desire by simply directing introspective attention at its content. Another advantage, as explored elsewhere, is that it explains the different *rational* roles of beliefs and desires.

Further, there does not seem to be any especially compelling reason to hold on to the view that the content of my desire that P is simply P. The best that can be done is to note that the *sentence* that we use to ascribe my desire that P has 'that P' as the syntactic complement of the verb 'desires'. But this does not require that the *representational* content of the state so ascribed by that sentence is itself simply P. There is a substantive question about the relationship *in general* between the propositional complements in attitude ascriptions and the representational contents of attitudes themselves, and the default view - that they are the same - may not be the correct one.

Accounting for transparent desire introspection, then, leads us to a novel view of the nature of beliefs and desires themselves. We can come to know what we desire by simply attending to the world because desires just are states that take the world to be a certain way.

5.8 Conclusion

We began with the observation that beliefs are sometimes transparent: we are able to come to know what we believe by looking outwards at the world. This generated a puzzle: how could looking at the world give us privileged access to what we believe? The solution was to understood "looking at the world" to consist in *deliberation*. Deliberation about what is the case affords us privileged access to our beliefs, because deliberation results in belief.

We then turned our attention towards desire, and saw that it is intuitively plausible that they are sometimes transparent as well. We saw that deliberation cannot afford us privileged access to our desires in the same way it can for our beliefs, because deliberation does *not* result in a desire. I proposed an alternative understanding of the transparent of desire: we come to know what we desire by directing introspective attention at the content of some representation. I argued that the most plausible candidate is a representation of the *rewardingness* of something. On this view, we come to know what we desire by looking to see what we represent as being rewarding. We saw that there is good reason to believe that the brain - in particular, the OFC - in fact represents rewardingness. Finally, we considered the puzzle of desire transparency: how could attending to the content of reward representations afford me privileged access to my desires, if desires are not themselves reward representations? My proposed solution was simple: desires just are representations of reward.

5.9 Bibliography

Armstrong, D. M. (1968). A Materialist Theory of the Mind, Routledge.

Ashwell, L. (2013). "Deep, dark…or transparent? Knowing our desires." Philosophical Studies **165**(1): 245-256.

Bar-On, D. (2004). Speaking My Mind: Expression and Self-Knowledge, Oxford University Press.

Bechtel, W. P. (1998). "Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science." Cognitive Science **22**(3): 295-317.

Byrne, A. (2005). "Introspection." Philosophical Topics **33**(1): 79-104.

Byrne, A. (2011). Knowing What I Want. Consciousness and the Self: New Essays. J. P. JeeLoo Liu. Cambridge, Cambridge University Press.

Dretske, F. (1981). Knowledge and the Flow of Information, MIT Press.

Evans, G. (1982). The Varieties of Reference, Oxford University Press.

Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind, MIT Press.

Francis, S., E. T. Rolls, R. Bowtell, F. McGlone, J. O'Doherty, A. Browning, S. Clare and E. Smith (1999). "The representation of pleasant touch in the brain and its relationship with taste and olfactory areas." NeuroReport **10**(3): 453-459.

Lycan, W. G. (1996). Consciousness and Experience, Mit Press.

Mainen, Z. F. and A. Kepecs (2009). "Neural representation of behavioral outcomes in the orbitofrontal cortex." Current Opinion in Neurobiology **19**(1): 84-91.

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

Mora, F., D. B. Avrith, A. G. Phillips and E. T. Rolls (1979). "Effects of satiety on self-stimulation of the orbitofrontal cortex in the rhesus monkey." Neuroscience Letters **13**(2): 141-145.

Moran, R. A. (2001). Authority and Estrangement: An Essay on Self-Knowledge, Princeton University Press.

Nichols, S. and S. P. Stich (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds, Oxford University Press.

Pavlov, I. P. (1927). Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex, Oxford University Press: Humphrey Milford.

Rolls, E. T. (2000). "The Orbitofrontal Cortex and Reward." Cerebral Cortex **10**(3): 284-294.

Rolls, E. T. (2007). "Understanding the mechanisms of food intake and obesity." Obesity Reviews **8**: 67-72.

Rolls, E. T. and F. Grabenhorst (2008). "The orbitofrontal cortex and beyond: From affect to decision-making." Progress in Neurobiology **86**(3): 216-244.

Schroeder, T. (2004). Three Faces of Desire, Oxford University Press.

Schultz, W., L. Tremblay and J. R. Hollerman (2000). "Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia." Cerebral Cortex **10**(3): 272-283.

Sescousse, G., J. Redouté and J.-C. Dreher (2010). "The architecture of reward value coding in the human orbitofrontal cortex." The Journal of Neuroscience **30**(39): 13095-13104.

Skinner, B. F. (1938). The behavior of organisms: an experimental analysis. Oxford, England, Appleton-Century.

Stich, S. P. and S. Nichols (2002). Folk psychology. Encyclopedia of Cognitive Science. S. P. Stich and T. A. Warfield, Blackwell. **7:** 35-71.

Tolman, E. C. (1948). "Cognitive maps in rats and men." Psychological review **55**(4): 189.

Tye, M. (2002). "Representationalism and the transparency of experience." Noûs **36**(1): 137-151.

Yaxley, S., E. T. Rolls and Z. J. Sienkiewicz (1990). "Gustatory responses of single neurons in the insula of the macaque monkey." Journal of Neurophysiology **63**(4): 689-700.

6 Conclusion

In this thesis I have defended a view that I call *Pure Attitude Representationalism (PAR)*. According to PAR, the nature of beliefs and desires is exhausted by their representational content. This is an alternative to the traditional view, according to which beliefs and desires are individuated by their functional role. According to the traditional view, the difference between a belief that P and a desire that P is that the state *does*, or is *disposed* to do. In contrast, according to PAR, the difference between a belief that P and a desire that P is not what it does, but what it represents. In particular, I have defended the view that while a belief that P simply represents *P*, a desire that P represents that *P is rewarding*. Being a desire *consists* in representing something as being rewarding.

This thesis consisted on three papers. In the first paper, *What are Beliefs and Desires*, I argued that the best account of what beliefs and desires are ought to be able to account for the role that they play in the explanation of behaviour. We saw that these explanations are both *causal* and *rational*. Appealing to the beliefs and desires of someone not only tells us what caused them to do or say what they did, but also why it was rational, from their point of view, to do so.

The traditional view of belief and desire is unable to make sense of this dual explanatory role. The traditional view is unable to explain certain of the causal generalizations of folk psychology: in particular, why some causal generalizations are true of one attitude type but not the other. The functionalist component of the traditional view takes these differences to be constitutive of attitude types themselves. But what we need is an explanation in the other direction: what it is about beliefs and desires that explains their differing roles in the causal generalizations of folk psychology? Further, because appeal to attitudes also rationalizes behaviour, our account of belief and desire must also be able to explain this.

I argued that the best way to account for the dual explanatory role of belief and desire is in terms of their representational content. The reason that belief and desire play different explanatory roles is because they have different contents: they present the world as being different ways, a way that renders different behaviour rational in their light.

In the second paper, *Attitudes and Self-Knowledge*, I argued that while it is intuitive that we have privileged access to our occurrent, conscious beliefs and desires, the traditional functionalist picture of their nature makes it is hard to make sense of this access. We are never in a position where we know that we have *some* occurrent, conscious attitude towards P, but are not sure whether we *believe* P or *desire* P. Rather, apprehension of the object of our attitude through introspection brings with it apprehension of the kind of attitude we bear towards the object. But since, according to the traditional view, the difference between a belief and desire concern differences in what these states are *disposed* to do under various *counterfactual* circumstances, it is hard to see how we could have privileged access to such properties.

I showed how this tension is similar to the tension between the idea that we have privileged access to the *contents* of our thoughts and the idea that contents partially constituted by how things stand in our external environments. According to this *content externalism*, mental contents are not intrinsic properties of mental states, but rather crucially involve relations to things external to us. We saw that the best way around this problem was to reject an *observational* view of self-knowledge of thought contents, according to which we come to know what we are thinking about by first *detecting* the content of the first-order state, and then generating the corresponding higher-order belief about it. Instead, we can hold that the content of the first-order state is simply *embedded* into the content of the higher-order belief, thus getting rid of any need to detect it. We then saw, however, that such a move is not as plausible in the cause of attitudinal properties like belief and desire, as long as such properties are understood to be functional and dispositional in nature.

I then argued that if we reject this view of the nature of attitudinal properties, and instead adopt PAR, then the content-embedding account of self-knowledge can straightforwardly be extended to account for privileged access to attitudinal properties. Thus, PAR allows for a simple and compelling account of our self-knowledge of attitudinal properties, while the traditional view requires significant explanatory costs.

In the last paper, *Reward and the Transparency of Desire*, I began by that desire is transparent in the following sense: that I can come to know what I desire by directing my attention outwards at the world. While the dominant explanation of belief transparency is that it is accomplished

through *deliberation*, I argued that desire transparency cannot be the result of deliberation. There is no question one can deliberate over that will reliably lead one to self-knowledge of one's desires. This is due to the *weakness of will* problem: that sometimes our desires and our judgments about what is good (or 'desirable') come apart. The couch potato knows that he doesn't desire exercise, even though he might judge that exercise is good; the addicted smoker knows that she does desire a cigarette, even though she knows that smoking is not good.

I then considered a recent attempt to account for desire transparency in terms of not judgment or deliberation but *appearance*: we come to know what we desire by attending to what appears good to us. I argued that this is best understood as what I call a *content account of transparency*: we come to know what we desire not by deliberating over some question, such as whether something is good, but through attending to some aspect of how our minds represent the world as being. But what content is relevant to desire transparency? I argued that the neuroscientific literature on goal-directed learning shows that there is good reason to think that part of the brain is responsible for keeping track of the *reward value* for various outcomes. These reward representations in turn play a crucial role in the production and motivation of goal-directed behaviour. I thus argued that this is a good candidate for a content relevant to desire transparency: we come to know what we desire by directing attention to what our mind represents as being rewarding. Finally, I argued that the best explanation of how such a procedure of attending to the content of reward representations could afford privileged knowledge of what we desire is that desires *just are* reward representations.

In sum, then, in this thesis I argued for a novel view of the difference between belief and desire. According to PAR, beliefs and desires are distinguished from one another solely on the basis of their representational content. I argued that this view is the best explanation of the explanatory roles that belief and desire play in folk psychology, as well as the best explanation of our privileged access to our beliefs and desires. I also showed that the neuroscience of goal-directed learning provides us with a strong candidate for the distinctive content of desires: reward.

Curriculum Vitae

Stephen Pearce

Education

Ph.D., Philosophy, Western University (Sept. 2011 – April 2016)； Supervisor: Angela Mendelovici (Philosophy); Advisory Committee: Rob Stainton (Philosophy), Chris Viger (Philosophy)

M.A., Philosophy, Western University (Sept. 2010 – August 2011)

B.Sc. (Honours), University of Toronto (Sept. 2004- April 2009); Major: Cognitive Science and Artificial Intelligence

Areas of Specialization

Philosophy of Mind, Philosophy of Cognitive Science

Areas of Competence

Metaphysics, Epistemology, Philosophy of Language, Kant (First Critique)

Presentations and Commentaries

Detection Theories of Self-Mindreading and the N ature of the Propositional Attitudes – International Association for Computation and Philosophy, 2013

On the Difference Between Beliefs and Desires - Western Graduate Student Colloquium, 2013 *R eward and the T ransparency of Desire* - Western Graduate Student Colloquium, 2014

Commentary on Luke Roelofs' *The Unity of Consciousness, Within Subjects and Between Subjects* - Canadian Philosophical Association Congress, 2014

Awards

Social Sciences and Humanities Research Council of Canada (2013-2015) - $40000

Ontario Graduate Scholarship (2011-2013) - $30000

Western Graduate Research Scholarship (2011-2015) - $34144

Western University Faculty of Arts and Humanities Entrance Scholarship (2011-2012) - $2000

Teaching Experience

Grader – Introduction to Philosophy (2010-2011)

Teaching Assistant – Advanced Introduction to Philosophy (2011-2012)

Teaching Assistant – Critical Thinking (2012-2013)

Teaching Assistant – Critical Thinking (2014-2016)

Professional Activities

Assistant General Editor, PhilPapers (2013-Current)

Co-organizer – Philosophy of Mind, Language and Cognition Graduate Conference (2011,2012)

Graduate Courses Taken

Topics in Philosophy of Mind: Representation (Angela Mendelovici)

Philosophy of Perception (John Nichols)

Externalism in Philosophy of Mind (Benjamin Hill)

PhD Proseminar (Tracy Isaacs)

History of Analytic Philosophy (Rob Stainton)

Kant's First Critique (Corey Dyck)

Kant and the Philosophy of Mind (Corey Dyck)

Survey of Philosophy of Science (Wayne Myrvold)

Empiricism in Philosophy of Science (Christopher Smeenk)

Free Will (John Thorp)