UNDERSTANDING NORMATIVE PRACTICAL REASONS

by

MARCELLO ANTOSH

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Ruth Chang

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October 2016

ABSTRACT OF THE DISSERTATION

UNDERSTANDING NORMATIVE PRACTICAL REASONS

by MARCELLO ANTOSH

Dissertation Director: Ruth Chang

Here I defend a version of internalism about normative practical reasons, which I call *discriminative stimulus internalism*. Discriminative stimuli feature prominently in explanations of human and non-human animal learning and behavior. And according to discriminative stimulus internalism, the property of being a reason is the property of being a discriminative stimulus of a special kind.

To defend this theory of reasons I first attempt to resolve a much simpler question: what regulates the concept of a normative practical reason? This question can be answered by characterizing both the rule and property that regulate the concept. According to what I call the *guidance account*, the rule that regulates that concept is explained by a more basic rule which determines how a wide variety of entities—from animals to artificial forms of intelligence—can learn to respond to considerations in particular ways, which, to observers, may appear to be instrumentally rational. These more basic rules are captured by principles of classical and instrumental conditioning and reinforcement learning. Those more basic rules explain when a consideration may come to guide an entity's response. When it can do so it is a discriminative stimulus. According to the guidance account, the property of being a discriminative stimulus regulates the concept of a reason.

The guidance account supports discriminative stimulus internalism in two major ways. First, it poses the parsimony challenge to a competing theory of reasons. This chal-

lenge casts doubt on the claim that normative beliefs and practices provide evidence for the existence of reasons offered by this competing theory. But it allows that those beliefs and practices provide evidence of the existence of reasons which are discriminative stimuli of a special kind. Second, the guidance account undermines two important objections against discriminative stimulus internalism: the extension and normativity objections. This straightforward, preliminary defense of discriminative stimulus internalism suggests it is well-positioned to emerge as the correct theory of normative practical reasons.

Acknowledgments

> The great influence at present exercised by Kant's teaching makes it worth while to state briefly the arguments by which he attempts to establish the duty of promoting the happiness of others, and the reasons why I am unable to regard these arguments as cogent... In the first place, that every man in need wishes for the aid of others is an empirical proposition which Kant cannot know *a priori*. We can certainly conceive a man in whom the spirit of independence and the distaste for incurring obligations would be so strong that he would choose to endure any privations rather than receive aid from others.
> — Henry Sidgwick, *The Methods of Ethics,* 389n

Although I can conceive, envy, and aspire to be him, Sidgwick's self-sufficient man, I am not. I am, however, an incredibly fortunate man.

Many people have made that so. Without their exceedingly generous combination of love, patience, money, kindness, time, money, guidance, patience, trust, money, and patience, it would have been impossible for me to complete this dissertation, save, perhaps, in a debtors prison many years from now. For shielding me from that bleak scenario, I wish to thank them.

I thank the members of my dissertation committee, Ruth Chang, Larry Temkin, Andy Egan, and David Sobel. Thank you all for your work which has and will continue to inform my own in important ways. Ruth and Larry, I must add that the content and quality of your work inspired me to choose to attend Rutgers—thank you for helping me make that excellent choice. And thank you all for the helpful discussions, feedback, guidance, and assistance with this dissertation and its defense.

Kathryn, I am grateful to you for your love, curiosity, loyalty, sense of humor, independence, self-knowledge, honesty, and patience. Thank you, too, and perhaps especially, for your impatience. It has helped make me more rational and only you could have

provided that key ingredient in a way I could appreciate.

I thank my family. Mom, Dad, and Peter, in the most fundamental ways over the course of my life you made this work possible by giving me the best opportunities and sustaining them into the present—I cannot thank you enough. Joan and Mauro, thanks for Kat. And, together with Daniel, Joanna, Kristine, Lisa, Matt, and Sean, thanks also for providing a home away from home always full of just the right diversions.

For discussions about rationality over the years and for being very good friends I thank Diana Bates, Alan Caces, Kirin Castelino, Jason Paul, Arthur Schipper, Erica Shumener, and James Simmons. For helpful conversations and feedback on these ideas I also thank Kristoffer Ahlstrom-Vij, Nathan Ballantyne, Tim Campbell, Vikram Duvvuri, Gabriel Greenberg, Doug Husak, Alex Morgan, Derek Parfit, Holly Smith, and Evan Williams. Mercedes Diaz deserves special thanks—I thank you for the tremendous and unrelenting practical assistance you have provided during the course of this work.

Finally, I thank the many students I have been fortunate enough to get to know and teach. Your curiosity and enthusiasm for topics related to this work have lifted my spirits when they were low and inspired me to refine some of these ideas.

The ideas in this dissertation have benefited most from the special support of two people. Exceptional thanks go to Ruth—no, not the biblical figure known for her loyalty and selflessness, but the chair of my committee known for the same. It has and will continue to be my passion to try to properly resolve the central questions about the nature of rationality. In key ways, at vital junctures, and under challenging circumstances, you helped ignite, guide, focus, and sustain that passion. I am not sure how to thank you for something as basic as an identity. I hope the theoretical progress I have made is some progress, too, toward a proper thanks.

Exceptional thanks also go to my dad. Dad, you are the honorary fifth member of my committee. The hours of philosophical discussions we have had over the years have been incredibly helpful and rewarding. The love, patience, and curiosity of any other dad would not have sufficed. More than anyone else, you have helped me with the theoretical progress I've made here. This is to say nothing of your greatest accomplishment—getting Mom to be my mom.

Table of Contents

Introduction to the Dissertation

This dissertation is about normative practical reasons. Normative practical reasons are considerations that count in favor of, support, or help justify the responses of some agent. They guide him toward responses that are rational, reasonable, and justified, and away from responses that are irrational, unreasonable, and unjustified.

A realist theory of normative practical reasons is one according to which the concept of a reason is about some property, and that at least some consideration is a reason because it has or instantiates that property. Competing realist views offer different candidates for what that property might be.

Here I defend a realist theory. I think the property of being a reason is identical with, reducible to, or fundamentally explained by a special case of a property familiar from the natural sciences—a discriminative stimulus. Because this property is a natural one with which we are familiar from everyday experience, and which the natural sciences incorporate in their models of learning, choice, and behavior, the view I defend is a variety of naturalism. Because the property involves an agent's preferences, the view I defend is also a variety of internalism. I call this theory *discriminative stimulus internalism*.

My defense of this view has two parts. The first preliminary part occurs in chapters 1 and 2, where I propose and defend an account—the *guidance account*—about what regulates the concept of a normative practical reason in an important class of cases: those in which this concept occurs in thought, and from an actualized first-personal perspective. According to this account it is useful to identify this concept's regulator in two ways. One

way is as the property that regulates the concept—the property is a discriminative stimulus. The other way of characterizing the regulator is in terms of the rule that regulates the concept. The rule is a function from an initial "input" mental state to an "output" mental state. The input state is the vital one. It is the state that disposes someone who is competent with the concept to have a token thought that some consideration is a reason for responding in some way. Roughly, this initial mental state consists of a certain belief and preference, and an association involving contents of each—I call this input state an *instrumental association*.

After proposing and defending the guidance account I then turn to considering how the guidance account can have substantive philosophical implications. In chapter 3 I present and discuss discriminative stimulus internalism and contrast it to a generic version of non-naturalism. I observe that normative beliefs and practices seem to constitute evidence for realism over anti-realism. But this presumption for realism can be countered by the parsimony challenge. The challenge is that the guidance account may adequately explain normative beliefs and practices without presupposing the existence of any property offered by some realist theory of reasons as a candidate for what reasons might be. If so, that would call into question the presumption for particular realist theories or, in the extreme, for realism as a whole. As I argue in sketch form, it seems that while non-naturalism faces this challenge, discriminative stimulus internalism avoids it. If this sketch is correct, this is an important initial advantage for discriminative stimulus internalism over non-naturalism and, in principle, over other varieties of realism about reasons as well.

In the remaining chapters I consider key objections to discriminative stimulus internalism and argue that the guidance account undermines them. One of these is the *ex-*

tension objection, which faults discriminative stimulus internalism for holding that certain considerations are reasons when they are not. Another is the *normativity objection*, which faults this theory for holding that the property of being a reason is natural when it is not.

If either of these objections were successful, this would undermine the presumption in favor of discriminative stimulus internalism. As my arguments reveal, both of these objections rest on a similar kind of premise: because the property of being a reason seems to be such and such, that's good evidence that it is such and such. The guidance account undermines this premise in each case by explaining why these appearances are misleading.

It seems discriminative stimulus internalism remains unscathed due to its ability to put the guidance account to use. Of course the defense of the guidance account rests upon appearances about how brains and minds work, and those rely on appearances about the natural world more broadly. Yet these appearances are ones we accept as part of our naturalistic conception of reality. If we accept them, then since they provide evidence for reasons only if reasons are discriminative stimuli of a certain sort, we should believe that is the best candidate for what reasons might be. In this way, it seems we have a preliminary argument for the conclusion that only some version of discriminative stimulus internalism will do.

Chapter 1

The Guidance Account

Introduction

A normative practical reason is a consideration that counts in favor of, supports, or helps justify the response of some agent. For example, suppose you are walking by a shallow pond when you notice a lone child drowning within it.[1] Since no one else is nearby, you are the only person who is able to save this child's life. You are wearing a new outfit which will be ruined by the murky pond water. Yet there is no time to remove it since the situation is urgent. To many people, it may seem that there are considerations both for and against trying to save this child. The consideration that your outfit will be ruined seems to be a reason for you to refrain from wading into the pond to save the child. And the consideration that the only way this child will survive is for you to wade into the pond and pull him out seems to be a reason for you to try to do so.

Although there is agreement in many cases about which considerations are reasons and which reasons are more important than others, there are many cases in which there is no such agreement.

There is less agreement about what fundamentally explains why something is a reason. Resolving this foundational question is of great theoretical and practical interest. Obligations and values can be understood in terms of reasons.[2] As a result, understanding

---

1 The example stems from Singer (1972).

2 As Joseph Raz (2000, 34) writes "The normativity of all that is normative consists in the way it is, or

the nature of reasons would help resolve some of the oldest and most central questions of the human condition. What is rationality, morality, and the meaning of life? How should we live? Because these theoretical questions have practical implications, their answers are of great practical significance as well.

I believe we can make great progress toward definitively answering these questions by first resolving a simpler one. What regulates the concept of a reason?

In this chapter I will provide an answer to this question and in the next I will defend it. In section 1, I distinguish between two ways of answering this question. The first identifies the property that regulates the concept of a reason while the second identifies a rule that users of the concept follow insofar as they are competent with the concept. For simplicity I restrict my focus by aiming to identify the rule and property which regulate the concept of a reason in only a central class of cases.

In section 2 I present my initial answer to the question. The answer centers around an important feature of reasons—they guide the responses of agents. In section 3 I refine the answer by first describing what the rule is like and then how it works in more detail. This involves understanding the processes of instrumental and classical conditioning, and it involves considering whether and to what extent there may be a difference between the way different kinds of entities can be guided by considerations.

1.1 Rules, Properties, and a Limited Class of Cases

What regulates the concept of a normative practical reason? I assume that, as is the case

provides, or is otherwise related to reasons." For similar views see Parfit (2011) and Scanlon (1998, 2014). See also Broome (2000) for a defense of the distinct normativity of requirements.

with many other concepts, there may be at least two ways to answer this question. It is helpful to begin by stating the first answer in a trivial way. The property of being a normative practical reason regulates the concept of a normative practical reason. That is, under the right conditions, if one is a competent user of the concept of a reason, when one thinks or asserts that some consideration is a reason, this is because this consideration *is* a reason. Or, more precisely, this is because this consideration has or instantiates the property of *being a reason*.

Stated in this way, the answer seems true, but trivially so. Still, there's an important idea beneath the triviality. Under the right conditions, someone who is competent with some concept is disposed to use it correctly—by applying it to all and only what it refers to or is about.[3] Thus, the first way of answering the question about what regulates some concept is to identify what the concept refers to, represents, or is about.

There is a second way of answering this question. Instead of identifying what the property of being a reason is, we can identify some rule that someone follows insofar as he is competent with this concept.[4] By following this rule one demonstrates one's competence with this concept. It is helpful to think of this rule schematically in terms of a function from one set of mental states—the input—to another—the output. For instance, if the output set consists of beliefs about whether various considerations are reasons, then the input may consist of beliefs, perceptions, recollections, and so on, which are precursors to

---

[3] The right conditions are idealized constraints that would allow concept users to determine what their concepts are about. Chalmers and Jackson (2001), stress the importance of these conditions. See also Chalmers (2004, 2012) for his discussion of the scrutability of reference and, Peacocke (2004) for discussion of a related constraint.

[4] The relation between rule following and understanding a concept figures centrally in many accounts of content, meaning, or semantics. For some examples consider Peacocke (1998) and Wedgwood (2001). See Kripke (1982) for the classical skeptical problem about meaning or content based on considerations of rule-following, and see Boghossian (1989) for a reply.

those competently formed output states.

The two answers should ultimately be related, but that relation need not be straightforward. For instance, consider the concept of water. Suppose that the property of being water is the property of being H2O. Then, in line with the first kind of answer, H2O regulates the concept of water. Yet people were competent with the concept of water long before its chemical composition or structure were known. Thus, insofar as one is competent with that concept, one need not, for instance, be disposed to believe that some substance is water on the basis of some initial belief that this substance is H2O. It is much more likely that the rule is of a different form. Perhaps the rule was based on the superficial features of water, such as its being found in lakes or rivers, its being transparent and odorless, and its being potable. Thus, perhaps the rule is a function to the belief that some substance is water from the belief that this substance is potable, odorless, transparent, falls from the sky as rain, and is found in lakes and rivers.

On closer inspection it is clear how the two answers relate. A substance's being H2O explains why in the environment of the speaker, it is found in rivers and streams and falls as rain. It also explains why to the speaker, it is potable, colorless, and odorless.

Together these two kinds of answers may help to account for different aspects of a concept's meaning or content.[5] The property-based answer may help to account for the concept's reference, while the rule-based answer may help to account for how the concept

[5] And so it is not surprising that leading theories of content and meaning incorporate elements of either or both kinds of answer into their accounts. For example, conceptual role theories (e.g. Harman (1975)) treat the rules as being rules of inference. Causal role theories (e.g. Fodor (1990), Dretske (1981)) treat the regulator as the object or property that causes token instances of the concept. Two factor theories (e.g. Field (1977), Block (1986), and Greenberg and Harman (2007)) incorporate features of both conceptual role and causal role theories and so too the rule and property or object which regulates a concept.

relates to other concepts.

For simplicity, I restrict the task in three ways. Competent use of a concept can arise in many forms, like thought, speech, writing, or some more creative variety of expression. So, first, I limit my focus to competence with the concept of a reason as it occurs in truth-apt modes of thought, like belief, intuition, judgment, imagination, and supposition.

Second, I aim to identify what regulates first-personal varieties of the concept. This is the use of the concept that arises when you think about which considerations are reasons for and against particular responses of yours. For instance, your belief that you are running low on toothpaste is a reason for *you* to buy more toothpaste invokes this variety of the concept. But your belief that the fact that your neighbor's child is crying is a reason for *your neighbor* to attend to her does not.

Third, with respect to these first-personal uses of the concept, I will focus on those uses involving scenarios in which your identity remains more or less as it actually is. Thus the rule I aim to identify will account for your thoughts about what your reasons would be were you to suddenly find yourself, given who you are now, in the Antebellum South or on some distant planet. But it will not directly aim to account for your thoughts about what your reasons would be were you to be a different person, such as a bigoted slave owner in the Antebellum South, or a native of that distant planet who also possesses the wildly different beliefs and preferences typical of some alien civilization.

So, in what follows, my concern centers on *actualized first-person* uses of the concept of a reason, as it occurs in thought, and when there is little to no departure in your identity. While I am open to the possibility that radically different rules or properties

account for different varieties of the concept in uses beyond thought, I suspect that is not the case. While I will not argue for the claim here, there is good reason to suspect that actualized first-personal use of this concept as it occurs in thought is among the most basic forms of normative thought.[6] In any event, let's now turn to the task of identifying the rule and property that regulate the concept of a reason for those uses I have in mind.

1.2 The Guidance Account and an Initial Worry

My proposal about what regulates the concept of a reason centers on an important and widely-recognized feature of reasons, namely, their potential to guide agents to respond in particular ways.[7]

> *The guidance account (first pass)*
>
> The rule that regulates the concept of a reason is explained by another rule. This other rule explains how some consideration may be capable of guiding an agent to respond in a particular way.
>
> The property that regulates the concept of a reason is the property of being a consideration that is capable of guiding an agent to respond in that particular way.

One immediate worry about this account is that it may center on the wrong feature of reasons. For although the response-guiding feature of reasons is important, very plausibly their most salient feature is their normativity.[8] Considerations are normative practical reasons because they support, count for, or help justify the responses of agents. That "count-

---

6 What reason is that? As we will see from the guidance account shortly, this class of cases is explained by a more basic rule that is vital to the lives of many living entities. Insofar as other normative concepts rival actualized first-person uses of the concept of a reason, that is because they are also explained by more basic rules of this sort.

7 Falk (1963), Velleman (1992), Darwall (1983), Williams (1981), Smith (1994), Markovits (2014), Hieronymi (2011), Stevenson (1944), Way and Whiting (2016), Nagel (1970), Korsgaard (1986).

8 Parfit (2011, pt 1, ch 1, 31) for the primitive nature of this normative role. For a sustained discussion of normativity more generally see the articles in Dancy (2000).

ing for"—that normativity—is what all and only reasons have in common. It is the characteristic feature of reasons and any account of what regulates the concept should focus on that feature rather than any other.

More fully, as a competent user of the concept of a reason, it may seem to you that your use of the concept is regulated by a normative phenomenology—that is, a distinct experience of what it is like when, to you, it seems that some consideration counts for, supports, or helps to justify some particular response of yours. Your awareness of that experience, and not the fact that some consideration may be guiding you to respond in some way, is what regulates your thought that a consideration is a reason for responding in some way.

For instance, suppose you have been worried about a close friend who has not responded to your many recent phone calls. You decide to visit your friend's house to make sure she is okay. When you arrive, you notice that her door is ajar and badly damaged. And, as you peer inside, you notice a lamp and a small table that have been knocked on their sides. At once, as you attend to these considerations, you may have a distinct phenomenological experience. To you, they seem to count for various responses of yours, including alerting the authorities immediately, being very seriously concerned about your friend's welfare, and entering but only with tremendous caution. This experience may seem different from the affective and motivational states you also experience. And it may seem to you to that *this* phenomenological experience—and not any way in which some consideration may guide your response—is what regulates your concept of a reason.

I am open to the possibility that, when more fully developed, this worry may pose a challenge to the guidance account. But I suspect it won't.

One immediate problem with the worry is that it seems to rest on a confusion. The confusion is that there is a relevant distinction between a consideration's being normative and its being a reason. But in this context there is no such thing. To describe the normativity of reasons we use concepts like *count for*, *support*, and *help justify*. But these concepts purport to refer to the same property that the concept of a reason does. To think that some consideration is a reason for some response is, for our purposes, another way of thinking that this consideration counts for that response. Thus the normativity that a consideration may have when it counts for some response is the same thing as its being a reason for that response. Similarly, the experience of what it is like when it seems that some consideration counts for some response is the same experience of what it is like when it seems that that consideration is a reason for that response.

More importantly, this worry overlooks a key possibility. First, put in phenomenological terms, it may be that when some consideration is capable of guiding some agent to respond in some way, that explains why it seems to count for some response. Thus the distinctive phenomenology of what it is like for something to seem to be a reason for responding in some way might emerge from whatever makes it play this role.

As we move from phenomenology to metaphysics, a parallel point holds. Sometimes when a consideration seems to count for some response it really might. And perhaps it is a reason precisely because of the way it can guide this agent to respond in this way. Clearly the guidance account already shows a superficial similarity to varieties of metanormative internalism which respect something akin to Bernard Williams's explanatory constraint.[9] While this similarity is interesting and will be put to work later, the focus

[9] Williams (1981).

in this chapter and the next concerns the more straightforwardly empirical issue of what regulates the concept of a reason.

1.3 The Guidance Account and the Problem of the Normative

To best understand the guidance account, it is helpful to place it in the context of a worry about different kinds of rationality. The worry concerns whether the guidance account will be able to accommodate an important difference in the way that considerations may guide the responses of agents and the way considerations may guide the responses of other entities which we may call *subjects*.

Psychologically normal adult humans are paradigmatic examples of agents. In contrast, examples of subjects include human babies, some non-human animals, and future varieties of artificial intelligence. Although subjects are not agents, their responses may be guided by considerations. Lizards climb up tree branches where they perceive insects. Babies cry when they observe their caregivers leaving the room. Foraging rodents interrupt their activities and scramble to safety upon detecting nearby predators. These seem to be clear cases in which considerations guide subjects to respond in particular ways. But this kind of guidance may be fundamentally different from the way considerations can guide the responses of agents. In her famous passage about the problem of the normative, Christine Korsgaard may be endorsing the existence of such a difference.

> A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities but it is not conscious *of* them... But we humans animals turn our attention on to our perceptions and desires themselves...
>
> And this sets us a problem no other animal has. It is the problem of the normative. For our capacity to turn our attention on to our own mental activities is

> also a capacity to distance ourselves from them, and to call them into question... I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I act? Is this desire really a *reason* to act? The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward.[10]

Let's grant that there is some important difference between the way that considerations guide the responses of subjects and the responses of agents. Might this spell trouble for the guidance account?

Maybe. But there are good reasons for thinking it won't. First, the guidance account seeks to identify the rule and property which regulates the concept of a reason among those who possess it. It makes no assumption about whether agents are the only entities that possess the concept. So it won't obviously run into problems about which entities can be competent with the concept of a reason.

Second, the guidance account paves the way for answering a substantive theoretical question about what the property of being a reason is (or what makes or fundamentally explains why some consideration a reason).[11] But the account itself is not that substantive theory. So it won't obviously run into problems about which entities have reasons—that is something for the substantive theory to handle.

There is a more important reason for thinking that the guidance account won't run into any trouble on this score. For despite Korsgaard's elegant presentation of the problem of the normative, which may suggest there is such a difference, the capacities may

---

[10] Korsgaard (1996, 92-93). Italics in the original. See also Nagel (2002, 146) for a similar view. The latter poses an interesting contrast with the younger Nagel (1978, 3) who conceived "ethics as a branch of psychology."

[11] Or: what it is for some consideration to be a reason, what being a reason consists in, etc.

turn out to be much more alike than they appear to be at first.[12]

While subjects may not face the normative problem Korsgaard describes, they clearly face other problems. These are the problems of learning and deciding what to do. Roughly, the learning problem is the problem of learning how to respond to satisfy one's preferences and the deciding problem is problem of deciding what to do to sufficiently—or perhaps optimally—satisfy one's preferences.[13]

It is an interesting question why there would be entities—like subjects and agents—that have preferences and so face the learning and decision problems at all.[14] Careful consideration of that issue is well beyond the scope of this dissertation. Here I assume that some broadly Darwinian answer is correct: subjects and agents have preferences (and face the learning and decision problems) because it was adaptive for their ancestors to develop the capacity to have preferences.

Of course, preferences do not exist in a vacuum. They are part of a broader set of mechanisms involving non-normative perceptual capacities and in some cases even capacities for other truth-apt states like belief, imagination, and supposition. The basic Darwinian idea is intuitive. Capacities for perception and preference allow subjects greater flexibility with respect to how they might respond to a changing environment. Having a preference to avoid being crushed by a falling log and a perceptual mechanism that allowed a subject to detect falling logs put it in a better position to have descendents in

---

[12] I make no claim about whether Korsgaard intends this rather strong interpretation. The point is that it can easily be read in that light.

[13] See especially Parker and Smith (1990). This way of stating the problems makes a direct link to work in artificial intelligence on reinforcement learning and dynamic programming which treat learning and deciding in this context as a problem of optimization. See Sutton and Barto (1998) and Schultz (2015) for explicit discussion along these lines in the artificial and psychological cases.

[14] For discussion of this issue see Glickman and Schiff (1967), Schulz (2011, 2013), Schultz (2015), Feinberg and Mallat (2016), Sterelny (2003), Parker and Smith (1990), Öhman and Mineka (2001).

later generations, compared to a subject that lacked this combination of preference and perception. When resources vital for survival diminish in some location, subjects with preferences for these reasons may take note, search, and find abundant resources elsewhere. Subjects who can't or don't do that may, on the whole be much more likely to die out and not have descendents in successive generations.

Having preferences and perceptual capacities is not particularly useful without also having learning capacities. In an incredibly static environment full of falling logs, perhaps it's conceivable that there might evolve subjects with an innate preference against falling logs or to avoid falling logs. But the actual environment is incredibly dynamic so a much more flexible learning mechanism seems to have arisen. Subjects were disposed to develop preferences for certain ends and learning mechanisms that would help them learn what to do to satisfy these preferences and what to decide to best satisfy preferences, given the potential for conflict.

For example, consider a rodent with both a preference for eating insects and for escaping from unfamiliar approaching objects. The rodent's first problem is to learn how to satisfy these preferences. Because the world is dynamic there could not have been selective pressures which bestowed on the rodent knowledge of what to do to satisfy these preferences. Instead, selective pressures equipped the rodent with learning dispositions for *classical* and *instrumental conditioning*.[15]

---

[15] Pavlov (1927) and Skinner (1938) and Thorndike (1911) for some of the early canonical work related to these learning processes. See Lieberman (2011), Domjan (2010), Gray and Bjorklund (2014), for contemporary discussions. Sometimes instrumental conditioning is called *operant conditioning*—Skinner's term—or *reinforcement learning*. I restrict *reinforcement learning* to refer to the abstract characterization of the optimization learning processes developed in, especially, the field of artificial intelligence with roots in mathematics, engineering, and physics. See Sutton and Barto (1998) for a classic introduction to the field.

The fear response is a good example of classical conditioning. Suppose our rodent has a disposition to be afraid of unfamiliar, fast-approaching objects. The predatory wolves that live nearby can be precisely such objects. Our rodent isn't born with a fear of wolves. But through exposure to wolves it can learn to fear wolves. Imagine while just emerging from its burrow, the rodent suddenly perceives a fast-approaching object—the wolf. If our rodent happens to survive this episode and other ones in the future, it gradually learns what to fear: things that smell and appear a certain way—the way wolves smell and appear. Once it is familiar with how wolves smell and appear our rodent can also learn to fear locations where wolves frequent, like the shrubs at the base of the hills or the grasslands near the lake. And our rodent can become afraid whenever it detects the scent or appearance of a wolf, even if there is no wolf nearby, in what amounts to a precautionary measure. This process of first exhibiting an involuntary response to one stimulus and, through a kind of learning, coming to exhibit that involuntary response to another stimulus, is classical conditioning.

While classical conditioning concerns involuntary responses—like fear, hunger, thirst, surprise, arousal, etc.—instrumental conditioning concerns voluntary responses. Just as our rodent isn't born knowing what to fear, it isn't born knowing what to do to survive. Its involuntary hunger response is triggered when it is low on nutrients and energy and it may begin to exhibit relatively involuntary gnawing behavior. This may help it, somewhat by chance, discover something edible—perhaps nuts, fruit, or grain. Just as repeated exposure to fast approaching wolves can help our rodent learn what to fear, repeated consumption of food can help it learn what to eat. Nuts—with their particular shape, color, texture, and aroma—but not similar looking pebbles are edible for our ro-

dent. And while it may on occasion attempt to eat a pebble or a stone, over time it becomes better at differentiating pebbles and stones from nuts and attempting to eat only the latter.

How do these learning processes work? What rule is being implemented by the brains of these subjects? Since at least the ancient Greeks, and well into the twentieth century, it was thought that *contiguity*—nearness in time of the relevant sequence of events—was the defining feature of the rule. Because the rodent smells the wolf's particular scent *at or about the same time as* it detects a fast-approaching object, it can learn to associate the scent with that object. As the association is improved, the scent itself can elicit the fear response even when there is no wolf nearby. It is because the rodent's hunger dissipates *soon after* it eats fruit—which also tastes and smells sweet *while* being consumed—that it learns to eat them. And it is because the rodent has found fruits *when* in the bushes, but *not when* by the rocky riverbed that it forages for them by the bushes.

Only after the middle of the twentieth century did researchers discover that, while important, contiguity was not the factor that best explained the learning rule being implemented. Instead, *contingency* seemed to be vital.[16] This insight is a key element in the Rescorla-Wagner model of learning which "has proved to be one of the most remarkable and influential models in psychology."[17] The intuitive idea is that what's really important for subjects to learn is not so much sequences of events. Instead, the key thing to learn is

---

[16] The landmark papers are Rescorla (1966) and Rescorla (1967). See discussion of the full Rescorla-Wagner model and contingency in, for example, Lieberman (2012). The discovery was first made in the case of classical conditioning. Classic demonstrations in instrumental conditioning are in Hammond (1980), Hammond and Weinberg (1984). Dickinson and Charnock (1985). For further discussion see for example, Murphy and Lupfer (2014),

[17] Lieberman (2012, 106). The landmark Rescorla-Wagner model is presented in Rescorla and Wagner (1972). Contingency is also central to the more abstract models of reinforcement learning.

what *predicts* or *causes* what. A contingency between events captures this predictive or causal structure more precisely than contiguity. The fact that the learning rule implemented by subjects is sensitive to contiguity may have been a useful and somewhat heuristic adaptation, as contiguity is often an excellent proxy for contingency. But contingency is the key.

Let me now define this notion of a contingency. For our purposes it will be enough to focus on a contingency relevant for instrumental conditioning.

> *Contingency definition in an example*
> Suppose the probability that the rodent is sated given that it is near the bushes and it forages is greater than the probability that it is sated given that it is near the bushes and it does not forage. If so, the proposition that the rodent is sated is *contingent* upon the consideration that it is near the bushes and the rodent's response of foraging.

So, subjects are able to learn contingencies. They have a disposition which implements some learning rule that allows them to detect contingencies under certain conditions. What contingencies are they capable of detecting? For our purposes, it is enough to focus on contingencies between considerations and particular ways of responding, on the one hand, and things they might have preferences for or against on the other. Preferences are important functional states which specify a subject or an agent's goals or ends. Learning theorists call these goals or ends *rewards* and *punishments* and understand preferences in terms of them.[18] Intuitively, when a subject or agent has a preference for something, this object is a reward. And when a subject or agent has a preference against something, this something is a punishment. For simplicity I will often limit my focus to rewards for brevity.

---

[18] Schultz (2015), Lieberman (2011), Domjan (2010), Gray and Bjorklund (2014). Some theorists use the term *reinforcer* and *punisher* in place of *reward* and *punishment*.

In at least some subjects, and in all agents, preferences are complex functional states that involve at least five components which may be disassociated.[19] First, they involve a representation of the reward or punishment. Thus they have a "cognitive" or "intentional" component. For simplicity, I will assume rewards and punishments are propositions even if that is implausible in the case of non-humans. Second, when subjects fulfill preferences this may be accompanied by the occurrence of a positive hedonic experience—like tasting something sweet—or the cessation or diminution of a negative hedonic experience—like being relieved from pain. Third, aside from this subjective experience, having a preference for something involves having an affective or emotional attitude toward it. The rodent has a preference for being in its burrow when it rains—it dislikes being in the rain and likes being covered. Fourth, preferences involve a motivational component—subjects may be more or less motivated to respond in ways that would satisfy their preference. Finally, preferences are implicated in learning rules such as those of classical and instrumental conditioning.

Let's take stock. Subjects have preferences and perceptions. The capacity to have these states was an adaptation that involves classical and instrumental conditioning. These conditioning processes are kinds of learning rules implemented by the brains of subjects. The rules are sensitive to contingencies. The contingencies that are most relevant for our purposes are contingencies between propositions which are rewards for the subject, on the one hand, and considerations and responses of the subject on the other.

---

[19] For discussion of the functional components of preferences see Schultz (2015) and Berridge, Robinson, and Aldridge (2009), Berridge and Robinson (2003). Those discussions are framed in terms of the components of reward. I've framed the matter in terms of preferences here for convenience. See also Feinberg and Mallat (2016) for their discussion of the origin of consciousness which they tie closely to the development of capacities for instrumental conditioning.

What happens when subjects learn? The contemporary view paints a complex picture too intricate to detail here.[20] I will simplify the picture by limiting attention to two items. First, the consensus in both human and a wide array of non-human cases—particularly in mammals and birds—is that conditioning processes are mediated by mental representations. The intuitive idea is that behaviorally relevant associations are formed between and among mental representations of the reward, consideration, and response implicated in the learning process. For instance, when the rodent learns to forage for fruit near the bushes, this is because it has discovered a contingency between the rewarding proposition that *it is sated* and the consideration that *it is by the bushes* when it issues the response of *foraging*. Learning this is mediated by the mental representation of a preference that *it is sated*, and of the consideration that *it is by the bushes* and the response of *foraging*. And for our purposes the most relevant association is formed between the consideration that *it is by the river* and the response of *foraging*.

> The core of the learning rule can be summarized as:
>
> A function from:
> The mechanisms which detect a contingency between a proposition (which is a reward or punishment) that is or can be the object of some preference, and some consideration and response the subject can perform.
>
> To:
> An association that is formed between a representation of the consideration and response. Call this "output" association an *instrumental association*.

The considerations featured in these instrumental associations figure prominently in research on learning and have their own name: *discriminative stimuli* (and *discriminative*

[20] For the most basic introduction see the discussion of conditioning in Gray and Bjorklund (2014). For more through discussions see Leiberman (2012) and Domjan (2010). For the very latest advances see handbooks such as McSweeney and Murphy (2014) and Menzel (2008). For important cross disciplinary work bringing the psychology of learning with the advances in artificial intelligence see Schultz (2015) and Jozefowiez (2002).

stimulus for the singular). When a consideration is a discriminative stimulus for some subject, it is capable of guiding the subject to respond in a particular way. Which way? The way with which its representation is paired in an instrumental association.

Dispositions which implement learning rules allow subjects to solve one part of their problem—the problem of learning how to satisfy preferences at all. But preferences can conflict. When our rodent hasn't eaten in quite a while being sated would be a reward. When our hungry rodent happens upon a predatory wolf and is in danger, being in its burrow, where it is safe from the wolf, would also be a reward. Even when our rodent has learned how to solve each of these problems individually by knowing what to do when hungry and what to do when being stocked by a wolf, how does it manage to solve the more challenging problem of deciding what to do when it has conflicting preferences?

Korsgaard's hunch is that the rodent's preference is its will. In other words, from the outside, it seems as if the rodent simply responds to whichever preference is greatest at the time. I will grant that simplifying assumption. For our purposes, it seems accurate enough: when out foraging and our rodent stumbles upon a wolf, if it has learned to fear wolves, then it scurries to safety immediately. Of course, we see now that what may seem too hastily characterized as an instinct or impulse is a learned response to a consideration.

But how does this "decision" get made? Why can't the rodent "decide" to continue foraging? Again the Darwinian answer is relevant. In ancestral rodents certain kinds of decision problems were frequent enough that it was adaptive to be equipped to make the right decision. What matters, for the purposes of ancestral rodents, was not to be decisively committed to foraging when hungry. After all, predators may be nearby and hungry too. Rodents with a decisive preference for foraging even when equipped to be afraid of

predators and experienced with their predatory presence were selected against. The ancestral rodents that did better had a more efficacious fear response as a means of solving this decision problem of conflicting preferences. This fear response would suppress the rodent's appetite, more fully arouse its perceptual faculties, and direct them to the object of fear and heighten the affect and motivation associated with the preference to flee. This suite of capacities requires some learning but operates so seamlessly to appear from the outside to be "instinct" or "impulse". Even if those characterizations are accurate, it is important to recognize that such "impulses" now present in rodents are the result of subtle learning and decision rules implemented by the brains of rodents. As this simple example suggests, assuming the learning problem has been solved, it is not so much rodents that solve the decision problem of conflicting preferences, but selective pressures. For problems that were frequent enough for ancestral populations of rodents, subtle rules associated with learning and decision shift the priorities of subjects to guide them to respond to the current set of considerations in ways that solve those ancient problems.

We are finally in a position to see how the way considerations guide the responses of subjects is not so different from the way they guide the responses of agents. Korsgaard is right. We are self-aware—we can reflect on our impulses, desires, preferences, and attitudes and scrutinize them. But keep in mind, we share ancestors with rodents. And, of course, we descend from more advanced subjects. These more recent ancestors were highly social and had significant cognitive capacities despite not having the ability to reflect in quite the same way as we can now.

Because we descend from those subjects, we inherited the same capacities which enabled them to solve not only learning and decision problems in general, but also some

specific learning and decision problems which were frequent and important enough to require a solution. Many of the so-called *social* or *moral* emotions seem to be implementations of learning and decision rules that solved such problems.[21]

Guilt, embarrassment, retribution, shame—all these dispositions seem to implement rules which guide subjects—and even agents—to respond in particular ways that solve learning and decision problems that were common and important enough in the ancestral past to require solutions. After a long day I am tired and hungry, and thinking about getting a meal when suddenly I remember that I failed to return grandmother's phone call. It's too late to call her tonight, but I feel incredibly guilty—my plans of getting a meal and going to bed now turn to plans about how to call grandmother first thing in the morning, and when to pay her a visit. Of course we are disposed to implement the solutions to more ancient problems as well. Fear, anger, disgust—these and other responses, which operate by way of conditioning, helped our ancestors quickly learn and decide to respond in ways that, in effect and on balance, solved important, recurring problems.

When we introspect and scrutinize with our reflective capacities, what exactly are we scrutinizing? It is not too much of a simplification to suggest that we are often scrutinizing competing impulses or preferences and the considerations and responses that are associated with them. I am aware of the consideration that I am hungry. That brings to mind getting something to eat. I have on hand salad and cake. I am aware of the consideration that if I eat the cake I would very much enjoy it—it would be an experience I really like. I'm also aware of the less salient consideration that eating the salad is better for my

[21] See for instance Greene (2013) for a proposal along these lines.

health. Somewhat subconsciously the thoughts of the cake brought to mind various associations, one of which was grandma—and it is around this time when I remembered that she was expecting my call. I've learned to feel guilty when I upset the feelings of others, especially others I care about deeply. And so the consideration that I've upset grandmother brings to mind responses which might make her feel better—something I very much prefer.

As I reflect, it is clear that I am reflecting upon, among other things, considerations and the responses they bring to mind. The considerations themselves are discriminative stimuli and they can guide me to respond in ways to obtain rewards. These pairs of considerations and responses feature as content in the instrumental associations that my learning and decision rules have generated.

In Korsgaard's presentation of the problem of the normative, it is all too easy to get the sense that when we scrutinize whether some consideration really is a reason for the response it brings to mind, we might come to the conclusion that it is not a reason after all. But cases like that are the exception rather than the rule. For, typically, or so I submit, the considerations that come to mind seem even upon reflection to be some reason for the response they bring to mind.

For instance, while thinking about the cake or the salad, perhaps on some level I already sense there's better reason for eating the salad, or that eating the salad is what I should do. I can think that while also recognizing that I very much want to eat the cake precisely because I like the way it tastes. Perhaps I ultimately feel that I shouldn't eat the cake. Still, it doesn't seem true to reflective experience to suggest that I think there is no reason to eat the cake. It certainly seems to me that the consideration that eating the cake

would be an enjoyable experience or that it tastes good are precisely those reasons. It's just that, on reflection, it seems to me that those are not particularly good reasons overall —to me, my long term health matters more. Indeed, that is why it is so easy to rationalize, devise excuses, and delude ourselves. The considerations that come to mind rarely seem to lack normativity altogether. Instead, typically, they seem to have it to some degree.

Already then, notice that it seems our reflective capacity is reflecting upon, among other things, considerations and responses that come to mind because of instrumental associations. These are the same kinds of instrumental associations that guide subjects to respond in the apparently "impulsive" and "instinctive" ways they do. Our reflective consciousness, it seems, allows us an additional measure of control over which of these considerations will prompt us to make some response. It allows us to potentially influence the ultimate decision in a way that subjects simply cannot do, or cannot do to the same degree.[22]

Yet the candidate consideration-and-response pairs that compete to influence choice all stand in some contingency with some rewarding proposition that we prefer. Typically even on reflection all of them will seem to be at least some reason for the response they bring to mind. So, I suggest, subjects are already in the normative "ballpark". The considerations which on reflection seem to us to be reasons are a subset of the considerations that come to mind because of learning and deciding dispositions which we share with subjects.

---

[22] Again for interesting research about the nature and function of consciousness see Feinberg and Mallat (2016).

But what about the difference between the cake and the salad? Eating the cake would be much more enjoyable than eating the salad. But even without too much reflection it seems to me that the consideration that eating the cake would be much more enjoyable is not a particularly good reason for me to eat it. The consideration that the salad is more nutritious seems to be, in comparison, a much better reason to eat the salad. Where does that apparent difference in normativity come from?

Our reflective consciousness isn't the only thing that sets us apart from subjects. We have vastly more sophisticated cognitive capacities. Besides perceptual capacities, we have an incredible ability to store information as memories and beliefs. This information is not static. We can shift our attention to consider some memories and beliefs and then others. We can move beyond memories of actual events to suppositions about possibilities. We can think about consequences of events, predict new ones, and rule out others. Even if we do not willfully access this spectacular array of information, it may be active subconsciously. This may have been how thinking about the cake prompted me to remember that I failed to call grandmother—as I've known since childhood, she makes delicious cakes.

Finally, and in addition to these rich abilities, we have our stock of preferences and the learning and deciding dispositions that come with them. As we attend to the choice of eating the cake or the salad, our minds can already project ahead, anticipating the consequences of each choice and comparing them. Although in the immediate future the cake would be pleasing to eat, this automatic comparison between projections tells a different tale. Our awareness of how delicious the cake could be now was natural selection's solution to a common and important enough decision problem our ancestors faced:

the "value" of eating sweets—like all these available fruits—is high in the sense of being adaptive. There was no problem of obesity on the savannah. An adaptive strategy was to consume energy-packed, nutritious food when it was available. And part of that solution involved making certain foods very pleasing to eat.

That old solution may have been good enough for our ancestors but it can often lead us into trouble. Our awareness that the health benefits of eating the salad are a better or stronger reason for eating it over the tasty cake is, I suggest, our awareness of a new "value". Like the old value this one concerns what's adaptive—what is the best way to go about fulfilling the preferences we have, given how they are weighted in us? My suggestion is our normative awareness is not fundamentally different from our awareness of the tastiness of the cake or the fruit or the blandness of the salad. While that tastiness is the short-term "value" of eating the cake or salad, our awareness of the greater reason to eat the salad is an awareness of the long-term "value" of eating the salad compared to the cake. This "value" is just shorthand for talk of what would be an optimal way of fulfilling all the preferences we have given their weights.

So what should we make of the so-called "problem of the normative"? Insofar as there is a problem here, it is a *problem of riches*. Indeed, it may be best to characterize this problem as a *solution* to those *other* problems of learning and deciding, which agents and subjects share. Because we can introspect we are able to feel, intuit, judge, and have beliefs about the normative significance of considerations. We can then scrutinize those intuitions—subjecting them to hypothetical scenarios, testing whether in those cases the intuitions hold. This spectacular capacity allows us to find *better* ways of responding to the world to satisfy our preferences. This solution of the normative is a welcome one in-

deed.

If these ideas are broadly right, then the rules that guide the responses of subjects and those that guide the responses of agents are closely related. In both cases those rules are learning mechanisms which detect contingencies. The most relevant contingency for our purposes is the contingency between a proposition which is a reward for some agent (or subject) and so is or can be an object of this agent's preference, and a consideration and response this agent can perform. When this contingency has been detected, the rule produces an association between a representation of that consideration and that response. I call this association an *instrumental association* since it can guide both subjects and agents to respond to considerations in ways that, from the outside, at least, seem instrumentally rational.

This instrumental association may be the output of the rule that explains how considerations can guide agents to respond in particular ways. But it is the input of the rule that regulates the concept of a reason. The output of the latter rule is, of course, a belief, thought, intuition, judgment, etc. that this consideration is a reason for responding in that particular way. Sometimes it will be helpful to also characterize this rule, somewhat artificially, in terms of beliefs in preferences. Thus the full statement of the guidance account's rule and property which regulate the concept of a reason is this.

Guidance account (final)

Rule version
For some agent, the rule that regulates the concept of a reason is a function

> From: either (i) An instrumental association within the agent which holds between representations of some consideration and response upon which a proposition, which is a reward for the agent, is contingent, or (ii) An agent's preference for some proposition. And his "belief" that that proposi-

tion is contingent upon his responding in some way when some consideration obtains.

To: this agent's belief that the consideration is a reason for responding in that way.

Property version
For some agent, the property that regulates the concept of a reason is the property of being a discriminative stimulus for this agent.[23]

Conclusion

I've now presented the guidance account, which is a proposal about the rule and property that regulate the concept of a reason in actualized first-personal cases of that concept. The core idea of the guidance account is that, at bottom, the rule that regulates the concept is explained by another rule. This other rule, as we now see, is really a set of rules that solve learning and deciding problems for subjects and agents alike. These problems concern how the agent can learn and decide what to do so as to satisfy his preferences in a sufficiently optimal way. Some of these rules determine how classical and instrumental conditioning work. A key feature of these rules is that they are sensitive to contingencies between preferences on the one hand and considerations and responses on the other. The learning and deciding rules are adaptations that detect such contingencies and form instrumental associations between representations of the relevant considerations and responses. These considerations are discriminative stimuli and they are capable of guiding the responses of subjects and agents alike. They are also the property that regulates the concept of a reason. And the rule that does so is a function from instrumental associations to beliefs about reasons. I will now turn to defend this account.

---

[23] My discussion of discriminative stimuli came quickly. See above in this section shortly after I presented the "core" of the learning and deciding rule, if you missed it.

Chapter 2

Defending the Guidance Account

Introduction

In this chapter I argue that the guidance account is true—at least in its broad contours. If so, that would mean that the property that regulates the concept of a reason is a discriminative stimulus. It would also mean the rule that regulates that concept is a function from an instrumental association to a belief about a reason.

One set of evidence that favors the account, and which I present in section 1, is that it can explain several key features of the concept of a reason. In sections 2 and 3 I turn to consider important challenges to the account. These challenges are cases in which it seems the guidance account does not explain what regulates the concept of a reason. On closer inspection however I show how these cases are explained by subtle variations of the account. In this way they are exceptions that prove the rule and provide further evidence that the account is correct. Finally, in section 4, I compare the guidance account with its most obvious rival, which I call the *detection account*. Because the guidance account is much better supported by the evidence in the preceding sections, and because there are further independent reasons to reject the detection account, I conclude that the guidance account is true.

2.1 Evidence for the Guidance Account

The form of the argument in this section is simple. An account aiming to identify what regulates the concept of a reason should be capable of satisfying several key explanatory constraints associated with the concept of a reason. First, I will identify these constraints and then argue that the guidance account satisfies them.

As I will assume, to account for what regulates the concept of a reason, we need to identify some rule that a competent user of the concept follows and which enables him to have thoughts about reasons. What is this rule like?

The first constraint on the rule is this. It must identify some condition that holds between token representations of consideration and response, within some agent. When this condition is met, the agent is disposed to believe that this consideration is a reason for responding in this way. Thus the first constraint on this rule is that it maps token representations of consideration and response to token propositions that this consideration is a reason for responding in this way. It does so because the representations of consideration and response meet some condition. And as a result, the agent is disposed to believe this proposition.

For instance, consider some thoughts that Larry and some sadist, Sadie, have about reasons. When Larry is considering what to do, it may seem to him that the consideration that the Dodgers game will be enjoyable is a reason for attending. When Sadie attends to the fact that agony feels the way it does, that may seem to her to be a reason for subjecting others to agony. Of course, Larry may not share a belief like Sadie's. When he reflects on the consideration that agony feels the way it does, that does not seem to him to be a reason for subjecting others to it. And while he may disagree with Sadie about a wide variety of thoughts about reasons, Larry and Sadie both regard themselves to be

thinking and talking about reasons.

When Larry and Sadie have their thoughts involving the concept of a reason, they are both following the same rule (or sufficiently similar rules)—that is how they are able to understand each other as thinking and talking about the same kind of thing. In Larry, token representations of the consideration that the Dodgers game will be enjoyable and the response of attending satisfy some condition. And as a result, Larry believes that the consideration that the Dodgers game will be enjoyable is a reason for attending. In other words, Larry believes the concept applies to this consideration and response. But in Larry, that condition is not satisfied by the representation of the consideration that agony feels the way it does and the response of subjecting others to it. And so when he considers whether he shares some belief like Sadie's it seems to him that he does not.

More precisely then, the first constraint on any account of the rule that regulates the concept of a reason is that it should identify some condition that pairs of representations of considerations and responses satisfy. The rule should then map each pair to some particular proposition about a reasons that a competent user of the concept of a reason is disposed to believe.

The second constraint on the rule is that it must account for the fact that an agent's reason beliefs can be response-guiding. Thus the rule must explain how, when representations of some consideration and response meet some condition within some agent, he is thereby disposed to be guided to respond in the relevant way to this consideration.

For instance, Larry is disposed to believe that the fact that the Dodgers game will be enjoyable is a reason for attending. When Larry reflects on the fact that the game will be enjoyable, this may guide how Larry responds to his circumstances, perhaps prompt-

ing him to consider attending, or motivating him to attend. If Larry does not attend but recalls that it would have been enjoyable, this may prompt him to feel bad that he didn't attend, and to wish that he had. Thus, the rule should account for the fact that when an agent believes some consideration is a reason for responding in some way, insofar as the agent is instrumentally rational, this consideration is capable of guiding the agent to respond in this way.

The third constraint on the account concerns the condition that token representations of considerations and responses satisfy within the agent. This condition must identify some *property* that each pair of consideration and response satisfy. This property must be a viable candidate for what the concept of a reason is *of* or *about*. The constraint is modest and requires only that the account identify some property that has two key features which correspond to features of reasons themselves.

First, when some consideration is a reason it can relate to a response by being a reason for or against this response. Call this aspect of the way a reason relates to a response its *orientation*. Second, when some consideration is a reason it can relate to a response by being more or less strongly for or against it. Call this aspect of the way a reason relates to a response its *weight*. Thus, the rule should identify some property that holds between pairs of consideration and response and this property should have features that might correspond to the orientation and weight of reasons.

Before moving on, it is worth mentioning two concerns. The first concern is that this constraint seems to presuppose that the concept of a reason is about something. But perhaps it isn't. For example, some foundational views such as expressivism hold that the concept is *not* about anything and that, as a consequence, normative beliefs are not truth-

apt in the familiar way.[24] So in what sense is this third constraint a genuine constraint? Shouldn't the account leave open the possibility that the concept might not be about anything at all?

The second concern is related. It seems that this constraint requires that the rule specify some property *besides* the property of simply being a reason. But some non-naturalist views hold that there is no further property which the relation of being a reason might be, reduce to, or be explained by.[25] Does this proposed constraint presuppose that non-naturalist views are false as well?

The proponents of these kinds of views are right to be concerned. I think the account of what regulates the concept of a reason ultimately gives us some evidence against these views. But it does not do so simply because of this constraint.

First, to address the non-naturalist's concern notice that I've said nothing about what property must be specified by the rule. So it is entirely compatible with my account that the property identified by the rule that regulates the concept of a reason is simply the property of being a reason, and that this property is non-natural. This is clearly a viable candidate.

The expressivist's concern is more warranted. Why should it be a constraint on the account that it identify some property that the concept might be about? Strictly speaking, it shouldn't. Thus the constraint could be recast in a way that is friendly to the expressivist. The account should be capable of identifying some viable candidate property that might hold or that appears to hold between the consideration and response.

---

[24] See e.g. Gibbard (1990, 2003), Blackburn (1984, 1993, 1998).
[25] Scanlon (1998, 2014), Parfit (2011), Nagel (1976), Enoch (2011).

The reason for this constraint is simple. Even if there is no property that the concept of a reason is about, the concept itself has a relational structure. For instance, consider Larry's belief that the fact that the Dodgers game will be enjoyable is a reason for attending. Suppose reasons are an illusion of some kind since there isn't any genuine property or relation between that consideration and response that could plausibly be the property we are after. Still, it is possible for Larry to have a belief which *seems* to be about such a property since he has a concept that seems to be about such a property. Strictly speaking then, the constraint on the account is to explain why the concept of a reason *seems* to be about a property that holds between pairs of consideration and response, and which seems to have an orientation and weight. Still, and for simplicity, since my concern in later chapters is to explore the plausibility of a viable candidate, I will understand the constraint as demanding that the account identify one even if, technically it need only explain the appearance of such a candidate.

It would be good for the account about what regulates the concept of a reason to satisfy four further conditions.

First, it seems possible to learn about what reasons one has *a priori*. To discover his reasons for and against attending the Dodgers game, for instance, it may often be enough for Larry to reflect upon non-normative information about his circumstances and possible ways of responding to them. Doing so will give Larry an initial impression of the reasons for and against going. And when he scrutinizes these reasons, weighing some against others, perhaps rejecting some and systematizing the rest, it may seem to him that he has discovered his reasons. And it seems capable to do all this *a priori*. Ideally the account should explain why it seems possible to learn about reasons in this way.

Second, the rule should explain why some reason beliefs seem to be true by necessity. For instance, it seems true to Larry that the fact that he will have an enjoyable time at the Dodgers game is a reason for attending. Moreover, when he attempts to imagine whether this reason belief might not be true, he has some difficulty. Perhaps the earth will soon be destroyed by an asteroid. Or perhaps it never will. Regardless of what the background scenario happens to be, as long as it's true that Larry will have an enjoyable time at the game, it seems to him that this fact is a reason for attending. And as a consequence, it also seems true to Larry that, necessarily, the fact that the Dodgers game will be enjoyable is a reason for attending. It would be ideal if the account could explain why this is so.

Third, to many who are competent with the concept of a reason, it seems that beliefs about reasons are about a distinct, *sui generis* phenomena—something that is not capable of being explained in fundamentally different terms. For instance, when Larry believes that the fact that the Dodgers game will be enjoyable is a reason for attending, he thinks this belief is about some potentially justificatory relation between that fact and the response of attending. And this relation seems different in kind from other relations, such as causal relations or temporal relations. The account should explain why this is so.

Fourth, the psychology that underlies competence with concepts is complex. So we should not expect the account to be without exceptions. Still, for this account to be correct, these exceptions must not be fundamental departures from it. On closer inspection, these exceptions should be explained by extensions, refinements, or special cases of the rule and property it identifies. In other words, even when it seems that the account does not specify what regulates some particular instance of an agent's use of the concept,

some version of the account should still apply.

The guidance account satisfies each of these constraints. To see this it will be helpful to have a case in mind. Thus, consider Larry's belief that the consideration that the Dodgers game will be enjoyable is a reason for attending. The guidance account explains this belief in two ways. First, the consideration that the Dodgers came will be enjoyable is a discriminative stimulus for Larry. Second, the guidance account's rule holds that Larry is disposed to have this reason belief since there is an instrumental association within him between representations of the consideration that *the Dodgers game will be enjoyable* and of his response *attending*. This association is formed because his learning and deciding dispositions have detected a contingency between this consideration and response and some proposition which is the object of some preference of his. Without loss of generality, suppose this preference is that *he has an enjoyable time*.

Using this example, let's consider how the account satisfies the constraints. It satisfies the first constraint since it identifies some condition that must be met for representations of some consideration and response to be associated and then mapped to a proposition about a reason that the agent is disposed to believe. In Larry, representations of the consideration that the Dodgers game will be enjoyable and the response of attending become associated because Larry has learning and deciding dispositions. They detect the relevant contingency and form an instrumental association between the representations of the consideration that the Dodgers game will be enjoyable and the response of attending. Since those representations are associated, Larry is thereby disposed to believe that the consideration is a reason for that response.

The account straightforwardly satisfies the second constraint. According to the ac-

count, the rule that regulates the concept is a function from instrumental associations to reason beliefs. Instrumental associations explain how considerations guide agents to respond in particular ways. And they also dispose agents with the concept of a reason to believe that considerations are reasons for responding those same particular ways when representations of consideration and response feature in instrumental associations.

The guidance account also satisfies the third main constraint since it offers a viable candidate for what the concept of a reason might be about. The candidate is the property of being a discriminative stimulus. This is a property had by considerations when they relate to responses of agents in a certain way. What way? When there is some proposition that is the object of one of the agent's preferences and when there is a contingency between this proposition on the one hand and the consideration and response on the other.

To be a viable candidate this property of being a discriminative stimulus must have features that correspond to the orientation of a reason, as being either for or else against some response. Do discriminative stimuli possibly have such features?

They do, as can be seen in standard texts on instrumental conditioning.[26] Together, two features of discriminative stimuli determine this. The first is the valence of the preference related to the discriminative stimulus. For instance, when Larry prefers that he has an enjoyable time, the valence is positive in the sense that Larry's affective attitude toward the proposition that he has an enjoyable time is positive—he would like to have an enjoyable time, rather than dislike it. But a preference may also be against its object. Larry has a preference against it being the case that he is stuck in traffic. The valence of

[26] For example see Gray and Bjorklund (2014, ch 4). To avoid any confusion between conditioning processes and the field of reinforcement learning, I refrain from following Gray and Bjorklund's terminology of *positive* and *negative reinforcers* and *punishers* here.

his affective attitude toward this proposition is negative.

The second feature that determines the orientation of a reason is how the object of preference is contingent upon the consideration—that is, the discriminative stimulus—and the response. The contingency is *positive* when, all else being equal, it is more likely that the preference is satisfied when the consideration obtains and the agent responds to it in some particular way, compared to when the consideration does not obtain and the agent responds to it in that particular way. The contingency is *negative* when, all else being equal, it is less likely that the preference is satisfied when the consideration obtains and the agent responds to it in some particular way, compared to when the consideration does not obtain but the agent responds to it in that particular way.

Thus, for example, the contingency is positive with respect to Larry's preference that he has an enjoyable time, the consideration that the Dodgers game will be enjoyable, and the response of attending. For all else equal, it is more likely that Larry has an enjoyable time given that he attends the game when it will be enjoyable compared to when it will not be enjoyable but he attends.

Given the positive or negative dimensions of both the preference and the contingency, there are four kinds of discriminative stimuli. Two of them function in such a way to promote the agent's response to the consideration.

Response promoting discriminative stimuli

a. Positive preference, positive contingency:
By responding to the consideration, the agent increases the likelihood of something he likes. The consideration guides the agent to respond in this way.

b. Negative preference, negative contingency:
By responding to the preference, the agent decreases the likelihood of something he dislikes. The consideration guides the agent to respond in this way.

The remaining two function in such a way to inhibit the agent's response to the consideration.

Response inhibiting discriminative stimuli

a. Negative preference, positive contingency:
By responding to the consideration, the agent increases the likelihood of something he dislikes. The consideration guides the agent to avoid responding in this way.

b. Positive preference, negative contingency:
By responding to the preference, the agent decreases the likelihood of something he likes. The consideration guides the agent to avoid responding in this way.

Response promoting discriminative stimuli correspond to reasons with the orientation of *being for*, while response inhibiting discriminative stimuli correspond to reasons with the orientation of *being against*. So for instance, consider the contingency among Larry's preference that he has an enjoyable time, the consideration that the Dodgers game will be enjoyable, and his response of attending. Here the consideration that he will have an enjoyable time is a response promoting discriminative stimulus since by attending the game given that he will enjoy it, Larry makes it more likely that he will have an enjoyable time, which is something he likes. And, assuming Larry has the relevant instrumental association, Larry is disposed to believe that the consideration that the Dodgers game will be enjoyable is a reason for attending.

So far, so good. The guidance account identifies discriminative stimuli as the candidate that the concept of being a reason is about. And this candidate accounts for the orientation of reasons since a reason's orientation as being for or against may correspond to whether a consideration is a response promoting or response inhibiting discriminative stimulus.

Now we may turn to the weight of a reason. Do discriminative stimuli have some feature that might correspond to the weight of a reason? They do. Very roughly, and at first pass, this is the strength of the affective component of the preference, as opposed to, say, its motivational dimension.

Suppose that in addition to having a preference for having and enjoyable time, Larry also has a much stronger preference for having an amusing story to tell later. For Larry, the discriminative stimulus involving the second preference is much more strongly preference promoting than the first. And that feature may correspond to the weight Larry attributes to each reason. Specifically, both the consideration that the Dodgers game will be enjoyable and that he will have an amusing story to tell friends later guide Larry to consider attending. But the second does so more strongly. And that is why Larry is disposed to believe that both facts are reasons for attending but the second is a much stronger, better, or more weighty reason.

Since the property of being a discriminative stimulus has features that correspond to both the orientation and weight of reasons, it is a viable candidate for what the concept of being a reason might be about. But is it a plausible candidate?

I cannot consider that rigorously here. But I can quickly provide some loose remarks in what I take to be the right direction: I think some important, subtle refinements to the notion of a discriminative stimulus remain to be made. Not all discriminative stimuli are created equal. That's because the preferences on which they depend sometimes fade away, get rejected, or otherwise fail to play an important role in constituting an agent's identity.

Agents are constituted in part by their preferences. And as agents develop and

learn about themselves and the world, their preferences may change. With these changes also come changes in the identities of the agents themselves. These identities may be more or less well integrated or coherent. The preferences that endure and cohere in the agent may lay claim to being part of the agent's true or ideal identity. And if there are multiple sets of preferences that may cling together in this way—depending on how the agent's life goes or might have gone—then they may be alternate true or ideal identities of this agent. My hunch is that the process of reflecting and responding to considerations can change the possible identities of the agent. On this view there is no single ideal or true you, no best set of preferences above all others. The picture is much more fluid and dynamic. An analogy might be a person with a lantern in the dark. The outer limit of the lantern's light makes a circle. Each point on this circle represents a relatively local "ideal self." As you reflect upon or respond to your reasons, this may bring you closer to a point which was once an ideal. But as you've moved, so has your lantern and the outer limit of its light has moved too. There is a new set of ideal preferences or rankings on propositions—a new set of ideal identities given where you now stand, or who you have now become. Over time, you may reach terrain where your movements no longer shift the ideal so radically—this is steep hill country. You climb a mountain in the dark and the radius of the circle of light coming from your lantern is much closer to the lantern's source than ever before: you are acting more rationally than ever before. You are acting more autonomously too—i.e. you are expressing an identity that is, in some sense, more true to you than any identity you've expressed through your responses in the past.[27]

[27] For a discussion of a broadly similar link between rationality and self-expression—but not one that necessarily involves this lantern analogy—see Plunkett (2016).

Corresponding to this analogy we can tentatively introduce the notion of a *privileged discriminative stimulus*. This is a discriminative stimulus that is, in a certain sense, "free from local error". It is like a locally accurate weighting of a reason—it relates to a preference which is immune to certain local adjustments in relative strength. These local adjustments are of a sort that would arise through minimal shifts in the relative rankings of other preferences. And, perhaps also, they are free from error that depend on false or inaccurate information about the natural world. The idea of locality here is important since, as the lantern analogy suggests, there may be no single ideal set of preferences and preference rankings as reflection and responding to considerations very plausibly shapes this preference profile. All the objectivity we can hope for when it comes to reasons may be of this local sort that is "within reach."

The above arguments establish that the guidance account explains or satisfies the first three and main conditions on any plausible account of what regulates the concept of a reason. Now we may turn to the remaining constraints. First, the guidance account helps us understand why, for someone who is competent with the concept of a reason, it seems possible to learn about reasons *a priori*. Suppose Larry is unsure whether attending the Dodgers game with Monena a prostitute he has encountered (as from the TV show, *Curb Your Enthusiasm*), will end up costing more than $500.[28] Despite being unsure he can consider both possibilities. Supposing that attending with Monena will cost much more than $500, this strikes Larry as a strong reason against going. Supposing that it will cost much less than $500, this strikes Larry as a much weaker reason against going. Although Larry doesn't know how much it would actually cost to attend with Monena, it

[28] *Curb Your Enthusiasm* (2004).

seems to him that he's already able to know, *a priori*, some facts about his reasons against going.

This makes sense if the guidance account is correct. Instrumental associations can be formed *a priori* as Larry reflects upon different possibilities. And he might come to have new reason beliefs as a result of these reflections. According to the guidance account, *a priori* reflection about reasons is not fundamentally different than trial and error.

Next, recall that some beliefs about reasons are not just apparently true, but apparently necessary. The guidance account can explain this as well. Larry believes that the consideration that he will have an enjoyable time at the Dodgers game is a reason for attending. But when he tries to imagine whether this reason belief might not be true, he has some difficulty. For again, regardless of whether the earth will soon be destroyed, his wife will be more upset than he imagined, or there will be terrible traffic on the way home, as long as it remains true that Larry will have an enjoyable time at the game, it still seems true to Larry that this fact is a reason for attending. And that's why it seems necessarily true that, the fact that the Dodgers game will be enjoyable is a reason for attending.

That makes perfect sense if the guidance account is correct. For regardless of how the background scenario changes, as long as Larry holds fixed the consideration that the Dodgers game will be enjoyable, within Larry, there continues to be an instrumental association between a representation of that consideration and a representation of the response of going to the game.

Next, reason beliefs seem to be about something that is *sui generis*, and not capable of being explained in terms of, say, natural properties or relations. Can the account explain why this is so? It can. But a full discussion of how will have to wait until chapter

5, so I set the issue aside for now.

What about the fourth and final further constraint? Recall that this constraint accepts that since psychology is complex, competence with some concept will not be something that can be captured perfectly by some simple account. There will be exceptions. Still, to be the right account about what regulates a concept, these exceptions should not be fundamental departures from the account.. Instead they should be principled variations of it. So while it may appear that the account does not apply, some refinement of it should. Can the guidance account handle the apparent exceptions in this kind of way? Yes. Let's now turn to consider how it does in two kinds of cases.

2.2 Objection 1: "Preferences without Reason Beliefs"

One class of apparent exceptions to the guidance account stem from cases in which it seems the "input" condition of the account's rule is satisfied—and so there is a relevant instrumental association in the agent—but the agent is not disposed to believe that the consideration is a reason for responding in the relevant way.

Addiction cases seem to be a familiar version of this kind of exception. Imagine a methamphetamine addict who is trying hard to quit. This is a challenge for the addict in part because he continues to live in the same environment where he has cultivated his addiction. As a result, he is so often exposed to various considerations which prompt him toward addictive behaviors. While going to the strip mall grocery store he sees his drug dealer's car parked outside. Somewhat automatically that prompts the addict to consider approaching to buy just a little bit of methamphetamine. While at work he enters the

bathroom stall where, so many times before, it's been safe for him to snort meth. This calls to mind the urge to ask his coworker, who's a current user, if he has any meth to share. Representations of considerations like these have a strong association with representations of responses that have so often led him to relapse. One leading explanation of addiction is that deeply ingrained associations of just these kinds, coupled with constant reactivation when in the addictive environment, make it difficult for addicts to quit.

Although the addict too often responds to considerations like these in ways that prolong his addiction, he deeply wishes he could stop. And we may suppose that, despite having these associations among representations of consideration and response, he does not believe the considerations are reasons. He readily acknowledges, for instance, that the bathroom stall at work is a place to snort meth with relative discretion. But he strongly dismisses the idea that this consideration is a reason for trying to snort meth there.

Cases of drug addiction are an extreme example. But other cases share their structure. Consider conflicts which concern some major aspect of one's identity. A young adult growing up in a highly religious community finds herself attracted to members of the same sex, but would not say, and does not yet believe the consideration that she finds a female attractive is a reason for trying to begin a romantic relationship with her. A young boy in a well-adjusted family finds himself excited to dress in his sister's clothes but also feels ashamed and confused about it later, especially after being caught once. While eager to explore this aspect of his identity he may not think that is a reason to do so. Having been mortified and ashamed before, he thinks there is every reason to hate this aspect of himself and wish he didn't have these urges.

Finally, simple cases of temptation are a more benign variety of this same broad

sort of case. A dieter is tempted by the selection of pastries at the company meeting but may not believe their availability is a reason for breaking the rules of her regimen. And the scrupulous but cash-strapped custodian is tempted by the money in the lost wallet he's discovered. Yet, having been deeply affected by poverty and reflective enough to appreciate that the owner of the wallet may be in a similar position, he does not think there is any reason for keeping it.

Although cases like these may seem to be clear exceptions to the guidance account, there are two reasons for thinking that they are not problematic. The first is somewhat pedantic. In offering the rule of what regulates the concept of a reason, I am seeking to account for competence with that concept. But in at least some of the cases, there is reason to think that their failure is due to their use of another concept, or else a departure from competence with the concept of a reason. In either case, there would be no threat to the account being offered.

Consider thoughts to the effect that there is no reason for responding in some way. Typically what people have in mind by applications of that concept is the concept of *no good reason*, or *no sufficient reason*. The cases above may be similar. When pressed, the dieter and addict may concede that, strictly speaking, there are some reasons for indulging. At the very least, there are direct hedonic benefits of doing so. If the agents would concede that certain considerations are reasons after all, then it is clear that the original cases were not exceptions to the guidance account Alternatively suppose that when pressed, they would neither concede nor believe that there are any reasons for indulging, despite the acknowledged hedonic consequences of doing so. Then we would need to probe their use of the concept further to ensure that they are competent.

Suppose that their failure to apply the concept of a reason was widespread. For instance, although the dirty laundry piling up prompts the addict to consider doing the laundry, he denies the fact that it's piling up is any reason to do it. And while the dieter feels the urge to get to bed early given the hectic schedule at work tomorrow, she denies that's any reason for turning in early. If they fail to apply the concept in these and other cases, it becomes less and less clear that they are competent with the concept. Assuming that they apply the concept in at least some of these cases, however, then the guidance account seems to capture that use. And what we need is an explanation for why the agents depart from the account in these particular cases.

This brings us to the second reason to think these cases are not problematic—one that reveals an intimate relation between the concept of a reason and an agent's self-conception. Notice that in each of these cases there seems to be some obvious conflict between the agent's preferences. For instance, although the meth addict may strongly prefer that he lives a drug free lifestyle, at least when he is reminded of his addictive ways, he has some kind of preference for getting high once again. In the most extreme case, the preference may be a raw urge or motivation which brings no relief from any variety of pain, unease, or discomfort. He would be like Warren Quinn's radio man who, somewhat automatically adjusts the dials on any nearby radios without receiving any apparent relief from pain or experiencing any pleasure. Perhaps cases of this kind are entirely hypothetical—wouldn't there be *some* relief from scratching that addictive itch?

But it is clear we can imagine a disassociation among the affective or hedonic aspects of fulfilling some preference from the motivational ones. The conceptions of rewards and preferences in contemporary research in learning supports this possibility as

well.[29] In less extreme cases this preference for getting high one last time carries with it the prospect of some form of pleasure or freedom from pain. Yet in both cases, the addict disavows the apparent normative significance of these experiences. What matters to him most is knocking off this irresponsible behavior and getting clean once and for all.

The cases of identity and mere temptation also involve some clash of preference. Perhaps the young woman in the religious community would very much enjoy being in a romantic relationship with another female. But this would make her family and community upset. She imagines ending up embarrassed, ashamed, and ostracized, all outcomes she has overwhelming preferences against. The young man with confusion about his gender identity enjoys the experience of pretending to be a woman but was caught once by his sister. This mortifying experience makes him feel uneasy about his desire, and his fear that others may find out makes him feel there is no reason for him to be so curious about being a woman—he's a boy after all, and boys are supposed to become men.

On a much more superficial level there is a conflict of preference in the cases of the dieter and cash strapped custodian as well. The dieter would like to taste the delicious sampling of treats, and the custodian would be relieved to have another fifty dollars for expenses that might crop up before his next paycheck. But they have conflicting preferences for losing weight and helping others, respectively.

In cases like these I think conflicts among sets of preferences within the agent help explain why they may not be disposed to have some reason belief despite having the relevant instrumental association. Some evidence that a conflict of preference rather than non-applicability of the guidance account is at work stems from what reason beliefs they

[29] See the discussion of the components of preference in chapter 1.

would have absent any conflict. Before the meth addict acknowledged any problem—or indeed before there was any problem to acknowledge—it is plausible that he felt at least somewhat justified in his more stimulated lifestyle. It feels good to be high, and it's fun to party with others. It's also one hell of a way to find the motivation to organize all of one's belongings. When there is no acknowledged conflict of preference, it seems plausible that, to the user, considerations like these are reasons for getting high. Similarly, if raised in a social climate that tolerates exploring gender identity, perhaps the young boy would feel less confused and more confident and ultimately justified in his desires to appear as and possibly become a woman.

If this suggestion is right, then when an agent's preferences conflict, this may suppress some reason beliefs that the agent would be disposed to hold otherwise. But if so, and if the rule I've specified still regulates the concept of a reason, then when preferences conflict, what determines which preference will still seem normatively significant to the agent? This is an extremely interesting question and I can only offer a sketch here. As I've mentioned, I suspect there is an intimate relation between the concept of a reason and the agent's self-conception. Agents are constituted in part by a set of preferences. These preferences are not typically coherent and as integrated as one might hope. Still, some of these preferences are recognized by the agent as constituting part of an identity. This acknowledged identity need not be ideal in any way—the developing addict may love the lifestyle that's cultivating his addiction, and not see how it will one day threaten to undermine other things he cares about deeply. Still, when there is a conflict of preference, my sense is that the preferences which the agent himself believes—correctly or not—to constitute his identity will continue to exert normative influence, while those that

conflict with this identity are most liable to have any normative influence suppressed.

Thus, while in the throes of addiction, the addict may understand the concern of his loved ones, and their attempts to help him, as not simply challenges to his choice of lifestyle—how he spends his time, where, and with whom—but as threats to his identity. His budding preference against alienating them, and for seeking their help are part of this threat. Accepting and appreciating their role would constitute a self-abdication, a transfer of control from one identity to another.

A possible consequence of appreciating this existential threat from within, I think, is that the preferences which challenge the incumbent identity lack the normative influence that they might otherwise have. Part of the addict's failure to appreciate that he has any problem may consist in the failure of these preferences to help give rise to reason beliefs—in accordance with the guidance account. He feels there are no reasons to seek his family's help, and he feels there is every reason to resent how they have tried to interfere with his life. That should be expected—the concept of a problem is also normative.

Perhaps as he comes to appreciate how deeply his addictive and non-addictive preferences conflict, and so comes to understand and accept who he really is, what matters to him most, and what he must do to change, it is the lingering addictive preferences that will no longer exert normative influence on his thought and choice. And with that comes a new normative perspective of himself and his circumstances. It was a mistake to think that his loved ones were interfering with his life or trying to control him. They were only trying to help him, and they knew better than he did, that he was better than the addict he'd become, and that in some sense, that addict was never who he really was. What he once saw as genuine reasons for using drugs, he now understands as rationalizations.

What he once saw as justifications for not making any serious effort to change, he now views as excuses at best. With this gradual shift in self-conception comes a change in normative point of view. Combined, these non-normative and normative shifts in perspective yield a new, possibly more accurate narrative.

Reflections along these lines undermine any serious challenge these cases may pose to the account I'm offering. It is highly plausible that in these cases, some version of the guidance account still applies and regulates the concept of a reason. Thus, it's still plausible that when representations of some consideration and response have become associated in an agent, this disposes the agent to believe that consideration is a reason for responding in that way. But the nature of the instrumental association makes a difference. More precisely, if some instrumental association relates to a preference that conflicts with other preferences, and if the agent believes—correctly or not—that those other preferences or their objects are more central to his identity, then the disposition to make that reason belief may be suppressed. Rather than challenge the account, however, these cases help us appreciate its depth. By drawing attention to the relation between identity and our reasons, between who we take ourselves to be and what reasons we take ourselves to have, they reveal how, if the guidance account is true, the concept of a reason is intimately tied to our self-conception. That is exceedingly plausible.

2.3 Objection 2: "Reason Beliefs Without Preferences"

A second variety of apparent exceptions stem from cases in which there is evidence that there is no relevant instrumental association within the agent but the agent is disposed to

believe that the consideration is a reason for responding in some way. The most compelling cases of this kind involve patients with late-onset acquired sociopathy.[30] Congenital or developmental sociopaths display various anti-social attitudinal and behavioral traits, including a reckless disregard for the safety of others and a lack of remorse. These congenital or developmental sociopaths consistently display such traits throughout the course of their development as a result of their nonstandard genetics.

In contrast, patients with late-onset acquired sociopathy do not display such traits during the course of their development. Instead, as they develop into adults, their pro-social behavior develops in characteristic ways. Their burgeoning concern for themselves gradually extends to a concern for family members, and those in affiliated social groups. Sometime after this normal development they sustain frontal lobe damage and then begin to display traits typical of congenital sociopaths.

Remarkably, while their behavior is now distinguished by its antisocial character, these patients retain their prior, normal capacity for moral judgment and reasoning. When tested in moral reasoning and judgment tasks, their performance may be comparable to psychologically normal individuals.[31]

These patients were not directly tested to see which reason beliefs they hold. But given their normal performance in the moral reasoning and judgment tasks, we may suppose that although they are disposed to have reason beliefs typical of normal adults, they lack the affective and motivational dispositions of these adults. This may suggest that they are disposed to have beliefs about reasons without having the preferences which, ac-

---

[30] Damasio (1994). See Blair (2002) for a helpful overview of the important differences between acquired-onset and developmental sociopathy.

[31] Damasio (1994). But Greene (2013) contests this.

cording to the guidance account, are central to these beliefs. If so, these cases may be exceptions to the guidance account.

For illustration, suppose that Donald is a late onset sociopath who holds the typical reason belief that the fact that you would suffer if he punched you in the face is a reason against punching you in the face. Although Donald is disposed to have this normal reason belief, his affect and motivation do not seem to cohere in the usual way with that disposition. In particular, he is not relevantly disposed to refrain from punching you in the face. Nor would he be disturbed with himself were he to do so or express remorse after actually punching you. Such affective and motivational responses are evidence that Donald has no relevant preference—like a preference against causing anyone to suffer—which the guidance account seems to require him to have in order to explain that belief. Without some preference like this, however, regardless of his learning and deciding dispositions, there will be no instrumental association between token representations of the consideration that *you would suffer if he punched you in the face* and the response of *punching you in the face*. Because Donald would not have any association of this kind, if the guidance account is true, it seems he would not be disposed to believe that this consideration is a reason against responding this way. Since he is so disposed, that seems to be evidence that the guidance account is false.

Cases like Donald's are not fundamental departures from the guidance account since preferences are a complex mental disposition with distinct dimensions that can become disassociated from each other. As a result it is possible for an agent to have a degenerate preference in which only some of these dimensions function in the usual way. And as a result, an agent like Donald might seem to be an exception to the account when,

on closer inspection, the account still explains his disposition, though in a qualified form.

To see this, recall the different dimensions of preferences: their content, their hedonic quality, the affect toward the content, the motivation to respond, and the role in learning to form instrumental associations. My suggestion is that Donald's preference is degenerate. It retains its learning dimension, but lacks its affective and motivational dimensions. Presumably, for instance, Donald has some pro-social preference, like a general preference against harming others. Since this preference retains its learning functionality, when Donald learns that punching you in the face will cause you to suffer, he is disposed to believe that the fact that punching you in the face will cause you to suffer is a reason against doing so. But this degenerate preference lacks its affective and motivational functionality. As a result Donald does not seem to care whether you happen to suffer, much less when his punch happens to be the cause. Nor does he have the typical inhibition and restraint that would tend to prevent someone with that preference from punching you.

Support for this suggestion comes from three key sources. First, the brain regions which help regulate affect and motivation are damaged in late-onset patients. Indeed, cases of acquired sociopathy helped researchers discover that these regions of the brain play important roles in this regard. This suggests that if Donald had some preference against causing others harm prior to his injury, then this injury may have caused it to become a degenerate preference by impairing its affective and motivational functionality.

But what evidence is there that, prior to his injury, Donald had some preference against causing direct harm to others? This is the second source of support. On the one hand, it is common for most people to have some preference of this kind. This is clear

from our interaction with others, of course. But it was an important aspect in some of the early models of moral development. Specifically the model held that moral development occurs in stages with psychologically normal individuals having some concern for others. And more recent research suggests how this gradual development occurs. For instance, in a famous study, infants were found to display a preference for toys which behave in pro-social ways. The suggestion is not that this preference emerges as self-interested at first. But as the agent comes to develop capacities for recognizing minds of others and then empathy, they then come to have a preference for helping others. On the other hand, one of the distinctive facts about congenital sociopaths is that they have a marked lack of concern for others. They display lower levels of empathy toward others. Part of this seems related innate differences such as lower levels of fear.

Thus, roughly, on the assumption that Donald was a psychologically normal individual, through the course of his development, he would come to have some general pro-social preference against harming others directly. Thus his normal development would explain why he displays moral belief and reasoning typical of normal individuals.

A third source of evidence is especially telling and comes from cases of early- rather than late-onset acquired sociopathy. Early-onset patients sustain their injuries during infancy, childhood, or adolescence rather than adulthood,[32] and they exhibit behavior problems throughout the course of their development. In contrast late-onset patients exhibit serious behavioral problems only in adulthood, and their behavioral problems are less severe by comparison.[33]

---

[32] See e.g. Anderson et al.(1999), Damasio (1994), and Price et al. (1990).
[33] Ibid.

When tested as adults, early-onset patients performed much differently than the late-onset patients in moral reasoning and judgment tasks. As one researcher observes, these patients display moral judgment and reasoning capacities "characteristic of 10-year-olds, and... [are] surpassed by most adolescents."[34] This stage of moral development is dominated by short-term, self-centered concerns for pain avoidance. Patients falling under this stage of development display "limited consideration for the social and emotional implications of decisions" and "[fail] to identify the primary issues involved in social dilemmas."[35]

Why don't early-onset patients display the moral judgment and reasoning capacities of psychologically normal adults, but late-onset patients do? And why do late-onset patients make normal moral judgments and engage in normal moral reasoning but do not behave normally?

At least in outline, the answer seems relatively clear. These different groups of patients have different sets of preferences.[36] The late-onset patients had an ordinary course of development through adulthood. Part of this involved developing a theory of mind at an early age. This allowed them to gain an early and rich understanding of the mental lives of others. In turn, this facilitated the natural and early development of empathetic concerns and attitudes regarding the welfare of others. In addition to being disposed to develop these empathetic concerns and attitudes, they engaged in ordinary interactions with others. When they hit their peers and siblings, making them cry, they felt guilt and remorse. When they helped their friends or parents, they felt satisfied and pleased. Thou-

---

34 Anderson (1999, 1033).
35 Ibid.
36 Damasio (1994) suggests a similar explanation.

sands of experiences like these helped shape the content of what mattered to these patients, and made these concerns more robust. Therefore, late-onset patients came to develop pro-social preferences which would dispose them to have the relevant reason beliefs. And prior to their injuries, these patients also had the affective and motivational responses typically affiliated with such associations. As a result, they did not exhibit any serious behavioral problems until after sustaining their injuries. These injuries caused their pro-social preferences, and likely others, to become degenerate. While still disposed to make the relevant reason beliefs, they no longer displayed affective and motivational responses typical of those preferences.

In contrast, early-onset patients never developed the standard pro-social preferences. This may be largely because their frontal lobe damage limited their ability to develop a theory of mind, which in turn diminished their otherwise ordinary human disposition to develop empathetic concerns.[37] This explains their comparatively severe behavioral problems during their development. It also explains why unlike the late-onset patients these patients did not engage in normal moral reasoning or make normal moral judgments.[38]Thus, it is possible that late-onset patients may be disposed to have certain reason beliefs yet seem to lack the preferences the guidance account predicts they would have. They have degenerate preferences which still allow for them to form the relevant instrumental associations among representations of consideration and response, and these associations dispose them to have typical reason beliefs.

---

[37] Unlike acquired sociopaths, developmental sociopaths do not seem to have an impoverished theory of mind. For one explanation of the source of their behavioral problems see Blair (2002).

[38] What about autistic persons who engage in moral behavior while presumably lacking an ordinary theory of mind? Baron-Cohen (2011, ch 4) offers one possible answer.

2.4 The Detection Account

There is a second argument for the guidance account. Of all the candidates that might regulate the concept of a reason, it is the most plausible in the following sense. First, there is good evidence that the learning and deciding rules exist and are implemented in subjects and agents. Second, those instrumental associations which these learning rules produce seem capable of explaining token reason beliefs themselves, as well as accounting for the response-guiding nature of reasons. Third, the guidance account is capable of explaining why apparent exceptions to it are not genuine exceptions. And fourth, there is no alternative rule that satisfies these three conditions as well as the guidance account.

I assume that my remarks above provide good support for those first three claims. Here I will argue for the fourth claim. I'll consider some alternative account and argue it fails to account for these conditions as well. I assume that this alternative is arbitrary. If so, the argument generalizes. And that would provide still more support that the guidance account is correct.

What alternative rule might regulate the concept of a reason? More concretely, how might some rule account for Larry's belief that the consideration that the Dodgers game will be enjoyable is a reason for attending?

One possibility is that the concept of a reason is regulated in part by mechanisms or rules which *detect* normative or evaluative properties—for variety I will focus on evaluative properties rather than normative ones, but nothing hinges on the difference. In turn this evaluative detection mechanism would regulate evaluative concepts. And finally, these evaluative concepts might regulate normative concepts. Call this account the *detec-*

tion account.

As a first pass, for example, it is somewhat plausible to think that Larry has that reason belief in part because he believes—perhaps implicitly—that by attending when the Dodgers game will be enjoyable, this promotes something that Larry believes to bear value. We may suppose that this something is the proposition that he has an enjoyable time and that Larry's evaluative belief here is simply that it is *good* that he has an enjoyable time.

On closer inspection, then, the rule specified by the detection account as being the one that regulates the concept of a reason has two distinct parts. The first part is some further rule that regulates some evaluative concepts. This is the part of the rule that disposes Larry to believe, for instance, that it is *good* that he has an enjoyable time, or that it is *bad* that he has a miserable time.

The second part of the evaluative rule is functionally equivalent to the practical learning and deciding dispositions of the guidance account, which explain how instrumental associations are formed between representations of some consideration and response. However, instead of associating those representations when they relate to something the agent prefers, this part of the evaluative rule associates them when they relate to something the agent believes to bear value. For instance, suppose Larry believes that it is good that he has an enjoyable time. Suppose also that Larry learns, believes, supposes, or imagines that the proposition that he has an enjoyable time is contingent upon the consideration that the Dodgers game will be enjoyable and the response of attending. Under these assumptions, the second part of the detection account's rule holds that an association forms between token representations of this consideration and response.

According to the detection account's rule, when and because such an association between token representations of consideration and response is made, the agent is thereby disposed to believe that *this consideration is a reason for responding in this way*.

Now that we have the detection account as our arbitrary alternative, let's see if it might have trouble.

The problem with the detection account and its kin is simple. There is good independent evidence that the learning and deciding dispositions that underlie the guidance account exist—those mechanisms are extensively studied in the sciences. There is no comparable evidence that the mechanisms which underlie the detection account exist, assuming those mechanisms are not simply those of the learning and deciding dispositions. So there is no good evidence that the detection account is correct.

What are the mechanisms that underlie the detection account? Recall the first is some rule that regulates evaluative concepts like the concepts of good and bad. The constraints on the rule for value concepts is similar to those that hold for the concept of a reason. Recall that a belief that some consideration is a reason for some response is a belief comprised of token representations of some consideration and some response.

To regulate the concept of a reason, the guidance account's rule needed to specify some condition, K, that held within the agent between token representations of consideration and response so that, when it held, the agent would be disposed to believe that the consideration is a reason for responding in that way. And since that belief seems to be about a property that holds between some consideration and response which can vary in orientation and weight, K had to identify some plausible candidate for what that property might be. More exactly, K had to explain why the concept of a reason seems to be about a

property with those features.

The concept of value seems to be about a property that certain things can possess and, as in the reason case, this property can vary in orientation and weight. Positively oriented value concepts include *good*, *great*, and *important*. Negatively oriented value concepts include *bad*, *terrible*, and *abominable*. Value concepts seem capable of varying in weight since one thing might be very good, somewhat good, extremely good, and so on, at least with respect to some parameter.

Just as we used the terms 'consideration' and 'response' to refer to those things to which the concept of a reason applies, and between which the relation holds, let's say that something instantiating some value is a *value bearer*. Value bearers might be objects like this table or that coffee cup, properties like deceit or honesty, and propositions, like that today was sunny or that I ate breakfast.

The rule that regulates value concepts should therefore take the following form. It should identify some condition, V, such that when a token representation of a value bearer might satisfy this condition within some agent, the agent is thereby disposed to believe that this value bearer instantiates the value. Further, V must identify some viable candidate for what the value concept might be about, or else at least help explain why the value concept seems to be about a property capable of varying in orientation and weight. Thus, in Larry, when the token representation of the value bearer that *he has an enjoyable time* satisfies condition, V, Larry is thereby disposed to believe that *it is good that he has an enjoyable time*.

So, what regulates evaluative concepts? One broad possibility is that preferences do so. On this view, when the value bearer is the object of a preference—when, for in-

stance, it is strongly preferred—then this disposes the agent to believe that the object is, say, *very good*, or *very important*, at least in some respect.

This view has advantages. We've seen that this approach seems to work in the reason case, and so it is tempting to think it might work here as well. Further, if preferences regulate evaluative concepts, the overall account would be somewhat parsimonious and independently supported by evidence since we know that agents have preferences.

Of course, if this detection account holds that preferences play this role, then it simply collapses into some version of the guidance account. For suppose preferences play a privileged functional role in determining the rule for evaluative concepts. Then, indirectly, the detection account would hold that preferences help to form instrumental associations and that is what regulates the concept of a reason. But that is already a key tenet of the guidance account. So, if preferences regulate evaluative concepts, the detection account would not be relevantly different from the guidance account. Thus, to be a genuine alternative, the detection account cannot hold that evaluative concepts are regulated by preferences—or any functionally equivalent mental state, such as any state that relates to the learning and deciding dispositions in the way preferences do.

There is another broad possibility. Rather than featuring as the object of preference, the token representation of some value-bearer might feature in some other mental state. And when it does, that may be what's needed to account for the agent's disposition to believe that the value bearer is, say, good.

What should this state be like? Given the option we've just rejected, it's safe to assume this state must have a functional role that is sufficiently distinct from preferences. The only alternative that comes to mind is some truth apt state, perhaps akin to belief or

perception.

Because evaluative talk is sometimes expressed in terms of perception, I'll assume that the state is a perceptual state which might somehow detect whether something bears value. I will say little about what this perceptual faculty is like. It will be enough to consider how it works in its simplest terms.

I assume that when this state features a token representation of some value bearer, it also features a token representation of some value property that value bearer instantiates. Thus, when Larry considers chocolate ice cream, he "evaluatively perceives" that it is very good. And that evaluative perception is what disposes Larry to believe that chocolate ice cream is very good.

So given this rough understanding of evaluative perceptions, perhaps there is nothing particularly mysterious about what the rule that regulates evaluative concepts is like. Arbitrary token representations of value bearers can be objects of this evaluative perceptual faculty. And when they are, this perceptual faculty also features token representations of evaluative properties those value bearers instantiate.

The viable candidate for the property of goodness might then be some *sui generis* evaluative property. The property has distinct valence and weight, we perceive those features of goodness by way of our evaluative perceptions, and that is why the concept *good* seems to be about a property that has those features.

We now have the full version of the detection account. Value concepts are regulated by evaluative perceptions. And the concept of a reason is in turn regulated by evaluative perceptions and some disposition to form associations between representations of consideration and response, when the agent learns, believes, supposes, imagines, etc. that

the value bearer is contingent upon some consideration and response.

We are now in a position to see why the guidance account is better than any alternative like the detection account. First, while there is direct evidence that something plays the functional role of preferences, there is no comparable evidence that anything plays the functional role of evaluative perceptions. In fact, in a major recent review of the learning and decision mechanisms, the possibility that there might be some such perceptual faculty is explicitly considered and rejected.[39] The considerations are twofold. First, the neuronal structure of other perceptual systems is well-known. But while there are receptors for various stimuli like sound, light, taste, color, and so on, the brain structures implicated in learning and decision lack a functionally comparable neuronal structure. In short there is no retina for detecting the value (or normativity) of rewards. Second, that there isn't makes sense given how considerations guide behavior and given the likely evolutionary origins of the learning mechanisms. Given the apparent variability and distribution of evaluative properties, it may have been more flexible to extract or create evaluative information from whatever representational capacities subjects and agents already possess.[40] Indeed, while it is clear there are evolutionary pressures for agents to develop some preferences and learning and deciding dispositions associated with them, there is no direct evidence that there is any pressure for agents to develop any evaluative perceptual system that is accurate.

For notice that it doesn't matter whether such perceptions work at all. First, assume the evaluative perceptions do not affect what one prefers. For instance, suppose that

---

[39] See the discussion in Schultz (2015, 861) of the "reward retina".
[40] Schultz (2015, 861-862).

some rodent, Mary, has malfunctioning evaluative perceptions. She cannot detect whether some value bearer instantiates positive or negative value—she is effectively blind to the orientation of value. Or consider her ancestor, Gary. Rather than being blind to the orientation of value, he perceives orientations besides just the familiar positive and negative ones. Yet, as we may assume, there are no such alternate orientations. Regardless, these agents have learning and deciding dispositions and preferences. And there are selective pressures for them to care about certain kinds of things—like their future well-being, and that of those who cooperate with them. Since evaluative perceptions made no difference to how these subjects responded to their environment, there were no selective pressures to come to perceive values more rather than less accurately.

Now assume the capacities for value perception exist and directly influence preferences. On these assumptions, of course there would be pressures for the perceptual faculties to perceive certain things as good and others as bad. For if the perceptions malfunctioned, then subjects would have maladaptive preferences. But notice two things here. First, strictly speaking, the selective pressures would be for *the faculties to work in a certain way*, not necessarily for *them to correctly represent value*.[41] And second, we can only establish that there were selective pressures for the faculties to work in certain ways, rather than others, on the assumption that they exist and affect preferences. So, we don't even get that weak conclusion unless we have independent evidence that these faculties exist.

Finally, given the discussion of the guidance account in this and the previous chapter, I take it to be *prima facie* plausible that it may be adapted to explain what regu-

[41] Street (2006) makes a similar point.

lates evaluative concepts as well. And, of course, if that is so, then the detection account would be superfluous by comparison.

Since there is no independent evidence for the existence of some perceptual system that detects normative or evaluative phenomena, and there is ample evidence that agents have functional states like preferences and learning and deciding dispositions, it is much more plausible that the guidance account rather than any alternative like the detection account is correct.

But wait. An objector might insist that although they don't have any viable alternative account something must be wrong since we seem to grasp abstract truths about possibility, necessity, and mathematical objects in a similar way. Perhaps there really is some perceptual system for abstract properties, relations, and objects. The fact that we appear to have mathematical, logical, modal, and other abstract knowledge seems to support the idea that we do. If so, maybe this is evidence—perhaps indirect—that we have such a faculty.

The main problem with this objection is that it rests on the assumption that there is no similar account in the cards for what regulates those concepts. Yet this conflicts with empirically informed views about how concepts work.[42] Roughly, on this kind of view, abstract concepts are regulated by rules governing relations among comparatively less abstract concepts. Clearly this is just a proposal about abstract concepts and not a robust defense. But, combined with my remarks here, I think the opponent of the guidance account should not be especially hopeful that the analogy with abstract concepts will help him.

---

[42] Lazareva and Wasserman (2008).

Conclusion

The guidance account seems to characterize the rule and property that regulate the concept of a reason as it occurs in thought and for actualized first-personal cases. I suspect this is a central if not the central variety of cases on which other versions of the concept depend. Thoughts about the reasons of others, or hypothetical versions of yourself are likely to be constrained by the workings of the practical learning and deciding dispositions and the preferences which one cannot imagine away. The brute necessity of selective pressures stamped upon our predecessors a tool for solving fundamentally practical problems. That tool guides not only how we solve those problems but how we conceptualize them normatively.

Chapter 3

Realism about Reasons

Introduction

The guidance account is not a philosophical theory about reasons. It is an empirical theory about what regulates the concept of a reason. But it can be used to defend a particular theory of reasons. I begin that defense here and continue it through the rest of this dissertation.

In section 1 I distinguish between two views about reasons, realism and anti-realism. And for simplicity I focus on two varieties of realism. The first is discriminative stimulus internalism about reasons, which is a variety of internalism inspired by the guidance account. The second is a generic form of non-naturalism about reasons.

In section 2 I consider the question of whether there is better evidence for realism or anti-realism about reasons. The presumption is for realism. For our normative beliefs and practices seem to constitute evidence that there are reasons. That evidential support can be undermined by what I call the parsimony challenge.

In section 3 I present that challenge and argue, in sketch form, for the first substantive philosophical implication of the guidance account: the parsimony challenge undermines the presumption for non-naturalism, but not for discriminative stimulus internalism. In other words, although our normative beliefs and practices seem to provide evidence of reasons, there is reason to think they do not provide evidence of reasons of the

kind non-naturalism proposes—we have a more parsimonious explanation of those beliefs and practices. But they seem to and very well may provide evidence of reasons of the kind discriminative stimulus internalism proposes.

3.1 Realism and Anti-realism

As I will understand the term, a *realist* theory of normative practical reasons is one according to which the concept of a reason is about some property, and that at least some consideration is a reason because it has or instantiates that property. It is because a realist theory holds that the concept of a reason *is* about some property and that at least some consideration has this property that thought and talk involving this concept might correctly or incorrectly represent reality, and so be true or false in a substantive rather than a deflationary sense.

In contrast, *anti-realist* views come in two broad types. On the one hand there are those which, thinking realism implausible but its surface appearance attractive, deny that there is some property which the concept of a reason is about but insist that thought and talk involving these concepts might be true; these include versions of expressivism. On the other hand there are more purely anti-realist views which deny that there is any sense to be made of truth or falsity when it comes to thought and talk of reasons.

Before considering specific kinds of realism about reasons, I should highlight what I mean by 'property'. Properties are typically distinguished from relations, with properties being had by particular elements and relations had by sets of elements. For instance, the property of being a human is a feature that you and I and many other particu-

lar things seem to possess. But no particular thing might simply have the relation of being taller than. While Smith may be taller than Jones, it makes no sense to claim, simply, that Smith is taller than. Thus while relations are had or instantiated by at least two elements, properties are had by only one.

The concept of a reason is more accurately characterized as being about a relation that holds, minimally, among four elements: a set of circumstances, a consideration, an agent, and some response the agent might perform. As an example, suppose that in some circumstance, the consideration that it is raining is a reason for Larry's carrying an umbrella. Here, the relation of being a reason holds among the set of circumstances, the consideration that it is raining, the agent Larry, and his response of carrying an umbrella.

Nevertheless, for brevity, I will continue to speak of the property of being a reason, which some consideration may have, rather than the relation of being a reason, which holds among a set of elements including that consideration. No part of the argument below hinges on this terminological shortcut.

The focus here will be on two types of realism. The first kind of realism holds that the relation of being a reason is non-natural. What does this mean? It is difficult to characterize what non-natural properties are in general, in part, because it is difficult to characterize the complimentary class of natural properties in a way that is not trivial. So, it may be more helpful to offer putative examples of each kind of property. Properties like being water, being fifteen feet wide, being a person, being the cause of a building's collapse, and so on, are generally thought to be natural. In contrast, non-natural properties are thought to include the properties of being a square, or being a prime number. The rough idea is that although natural properties are the kinds of things that might be

amenable to empirical observation, non-natural properties are not.

But aside from these negative remarks is there more that can be said about what the non-naturalist's candidate property is like? This is unclear. Some non-naturalists do an admirable job emphasizing the relational nature of the property. Scanlon, for instance, makes it explicit that reasons are a four place relation.[43] But the naturalist will agree that reasons at least appear to be of this sort. So the relational structure itself does not necessarily tell us what is distinctive about the non-naturalist's candidate.

Despite this, perhaps we can gain some insight about the non-naturalist's candidate by observing another class of non-natural properties—supernatural properties, like those of mythology, folklore, and fiction.[44] We can appreciate the idea of some witch casting a supernatural spell on someone, thereby cursing her. And we can appreciate the gods supernaturally approving of some course of human activity, and demonstrating this through crop-sustaining rainfall. And, as per the film franchise *Star Wars*, we can appreciate some act of revenge as instantiating the dark side of the Force, while some act of benevolence as instantiating its light side.

Thus perhaps the non-naturalist's candidate property is like one of these supernatural properties. Indeed, it is useful to go one step further and observe what some supernaturalist theory of reasons might offer as a candidate property. The property of being a reason might be that of being part of the light side of the Force, approved by the gods, or blessed by some magical spell. Or, perhaps more parsimoniously, it might be that of being a non-natural reason which, like these supernatural candidates, posits something non-

---

43 Scanlon (2014).

44 See McPherson (2012) who distinguishes among varieties of non-naturalism.

natural—namely, non-natural normative practical reasons—but unlike them, does not posit supernatural phenomena, like the Force or magic.

And of course, since it will be relevant below, it is important to remember that even when we set the supernaturalist candidates aside, there may be other varieties of non-naturalism about reasons. One such candidate is the paperclip non-naturalist which we will meet more fully in chapter 4. Like the familiar non-naturalist, he thinks the property of being a reason is non-natural and not supernatural. But he disagrees with the familiar non-naturalist about which considerations have this property since his normative worldview centers on making and sustaining paperclips.

One last note on non-naturalist varieties of realism. In the literature, non-naturalists take themselves to fall into two varieties. On the one hand there are those like David Enoch, William FitzPatrick, and Terence Cuneo, who claim that non-natural properties exist in some robust or ontological sense.[45] In contrast, others like Thomas Scanlon and Derek Parfit insist that their candidate property exists, but not in a robust or ontological sense.[46]

It is not clear to me what sense something might exist if not in some ontological sense. I will assume these non-naturalists might be able to explain what they have in mind more precisely and so there might be some non-ontological sense in which some considerations could possess non-natural properties. But nothing in what follows hinges on this distinction or that assumption. For either there is some sense of existence besides ontological existence, in which case Parfit and Scanlon qualify as realists and will face

[45] Enoch (2011), Cuneo and Shafer-Landau (2014), FitzPatrick (2008, 2011), and Heathwood (2015) (on at least some reading).

[46] Examples include Nagel (1986), Parfit (2011), Scanlon (1998, 2014), Dworkin (2011), Skorupski (2010), and Kramer (2009).

the challenges I describe below, or there is not, in which case their views would either be incoherent or non-realist and so face other challenges related to those below. Either way, non-naturalists will have some problem.

The kind of naturalism I'll consider is a version of internalism. According to such views, if some consideration is a reason for an agent to respond in some way, this is always because of how this consideration relates to some actual or hypothetical preference of the agent. So much for the gloss of internalism. But can we say something more specific about the candidate property being offered?

At first approximation, the candidate property is the property of being a discriminative stimulus. Thus, suppose that in some circumstance, the consideration that it is raining is a reason for Larry's carrying an umbrella. Then if the internalist's naturalist candidate is the right one, this must be because the consideration that it is raining relates to some preference of Larry's. Perhaps, for instance, Larry prefers that he remain dry and by carrying an umbrella, he's more likely to.

There is some reason to think that the property of being a discriminative stimulus isn't the best candidate that the naturalist might offer. To see why notice that discriminative stimuli relate to actual preferences of the agent. But considerations that relate to actual preferences in this way do not always seem to be reasons or reasons of the right kind.

For instance, suppose as a youth, you have come to find rock and roll music incredibly rewarding. And suppose you now have some overwhelming preference that you become an important musician and that this preference is much stronger than your preference for maintaining your relationships with loved ones. Suppose it's true that you can have a rewarding career as a rock and roll musician only if you commit to your craft over

the next year. Thus the consideration that you can have a rewarding career only if you commit to your craft over the next year is a discriminative stimulus for you. It relevantly relates to your actual overwhelming preference that you become an important musician.

Yet suppose that if you do so commit, you will learn something about yourself. In particular, you will learn how much those personal relationships matter and mattered to you. Perhaps, after you reflected on your efforts the past year, you would think that, while the original consideration for committing to your craft was still a reason, it was not as strong of a reason as you originally thought. You might even conclude that the original consideration was considerably weaker in comparison to your reasons against committing to your craft.

It would be an advantage for a naturalist theory to be able to vindicate this familiar occurrence. But if the naturalist's candidate property is that of being discriminative stimulus then it is not immediately obvious how this might be done.

The form of the solution may, however, be clear. The naturalist internalist should shift focus from the relation a consideration may have to occurrent preferences to ones involving some special class of preferences which we may call privileged preferences. The task then would be to fill in what this privileged class of preferences is. I suggested in chapter 2 that it may be some local or accessible preference that is free from radical change with respect to its relative ranking among other preferences you have. The improved naturalist candidate property is that of being a *privileged discriminative stimulus* which is a discriminative stimulus that relates to one of these preferences. Call the realist version of naturalism and internalism which holds this property to be its candidate *discriminative stimulus internalism*.

3.2 The Presumption for Realism

Is there any evidence that supports the conclusion that there are normative practical reasons? And if so, what is it?

Our normative beliefs about reasons seem to be evidence that there are reasons. For instance, it seems that if acting in some way would be agonizing, this would be a reason to try to avoid acting in that way.

Next consider normative practices which seem to involve reasons. Consider Sally who has recently accepted a job in a new city. Initially she might think that the fact that the bus stop is near her apartment is a good reason for taking the bus. But over time she might come to discover that the bus is terribly crowded and she is never on time for work. After trial and error she might come to believe that there is greatest reason to walk to work. And she might understand that she has learned about what her reasons really are through this experience.

In principle of course, she might have reached a conditional version of that conclusion from the armchair. For instance, she might conclude that if the walking option will be invigorating, faster, and less expensive compared to the bus, then there is greatest reason to walk to work. All that would remain would be to discover which set of assumptions obtains.

But the courses of experience that seem capable of improving normative beliefs are not limited to this sort. For sometimes scrutinizing our ends allows us to improve our normative beliefs. Suppose Sally begins to consider her course of career. The company

she works for supports animal testing. She knows this harms animals. By working in this capacity she is promoting its ability to do that. Originally she understood this as perfectly sensible. But she has gradually come to believe that her justifications are really rationalizations—they are not good enough. The reasons she originally thought to support her career choice now seem inadequate. True, there are still reasons for working here. The financial rewards are good and the work is challenging and helps her grow. But she knows that by participating in these advertising campaigns she is responsible for promoting the suffering of animals.

What she does as deliberative practice, normative philosophers do for a living. Part of the business of normative philosophy seems to be engaging in what is sometimes called *reflective equilibrium*.[47] Suppose our agent is concerned not just with whether she has good enough reason to work at this job, but whether she has good enough reason for engaging any number of actions. On one approach, she begins by considering her firmest beliefs, intuitions, and judgments about her reasons. Then she might seek to explain some principle that accounts for them. She might then test these principles against others. In this way she might come to reach a wide range of normative conclusions about various actions and form a more systematic, coherent, and mutually supporting set of normative beliefs.

In contrast, there are courses of experience that do not seem capable of improving her normative beliefs, at least in some cases. For instance, when deliberating about whether there are better reasons for taking the bus or walking, Sally doesn't think it helpful to point to random sentences in a book about tectonic plates, or to make observations

[47] Rawls (1971).

about the number of grains of rice in any of several bags in the local bodega. Similarly, some courses of action seem to Sally to impair her judgments about reasons and also her ability to reason. While socializing she knows she is more enthusiastic about making future plans. But she also knows that she later regrets these plans as misguided. When in a depressed mood she knows she tends to devalue everything and prioritize sleeping. But she also knows that while sleep is important, other things are still worth doing, just after she rests. And when she's distracted or tired or after too many drinks she also knows she can make normative judgments that are not in line with what she really believes.

There are other aspects of normative practice. We offer reasons to each other as advice, and we can later assess this advice as having been good or bad, and so the reasons as being genuine or not. And we offer normative practical reasons as personal and interpersonal explanations or narratives, and some of these explanations seem correct.[48] For instance, I might explain my decision to vote for some political candidate in normative terms, saying that the opponent gave me every reason to vote against him. Similar explanations may aptly characterize broader social events or movements. The best explanation for the dissolution of some political party may be that many party members saw good reason to vote for an unorthodox, polarizing candidate, thereby causing a schism.

Thus, normative beliefs themselves—which seem to be about reasons—as well as normative practices—which identify ways we might or might not improve normative beliefs, as well as explanations that appeal to normative properties—seem to constitute evidence for the existence of reasons.

---

[48] See e.g., Sturgeon (1988), Boyd (1988), Railton (1986), and Loeb (2005).

3.3 The Parsimony Challenge

There is a simple way this putative evidence for realism might be undermined. Suppose there is some adequate explanation of our normative beliefs and practices involving actualized first-personal reasons that does not presuppose the existence of any consideration's having some property offered by a particular realist theory of reasons. Then that would provide some evidence against this particular realist theory about reasons, on the basis of parsimony. Call this the *parsimony challenge*.[49] If no realist theory could plausibly avoid it, that would constitute evidence against realism in general and in favor of anti-realism about reasons.

How do our two realist theories—non-naturalism and discriminative stimulus internalism—fare with respect to this challenge? This objection seems to undermine evidence for non-naturalist theories which stem from basic observations of normative beliefs and practices. For recall the guidance account. According to it, the concept of a reason is regulated by discriminative stimuli. This explanation does not presuppose the existence of any non-natural property. And assuming the guidance account is broadly correct, discriminative stimuli adequately explain the use of the concept of a reason. I assume that allows them to adequately explain the bulk of our normative beliefs and practices involving the first-personal concept of a reason. So to account for these basic aspects of normative beliefs and practices, there is no need to suppose that any consideration has the non-naturalist's candidate property. And as a consequence, this undermines the presumption for realism of a non-naturalist variety. That is, it is not clear that normative beliefs and

[49] For a the a classic contemporary version of this challenge see Harman (1977). For discussion see, for example, Sturgeon (1988) and Loeb (2005).

practices themselves constitute evidence for realism about reasons in the non-naturalist's sense given the adequacy of the guidance account.

Of course, in a similar way, it is not clear that normative beliefs and practices themselves constitute evidence for a supernaturalist's conception of realism about reasons. In other words, although you might believe that the fact that it is raining is a reason for carrying an umbrella, this does not constitute evidence that the fact that it is raining has some supernatural property of being a reason for carrying an umbrella either.

Clearly these parsimony based arguments against the presumption for non-natural and supernatural varieties of realism are sketches. The details would need to be filled in more robustly. But the generality of a concern for parsimony—across a wide range of domains of inquiry—combined with the level of specificity of the guidance account make the challenge especially promising in this case.

Can the presumption for discriminative stimulus internalism, based on our normative beliefs and practices, also be undermined by the parsimony challenge?

Because I have not characterized the naturalist's candidate precisely, it is not possible to provide a definitive answer. But since the candidate property offered by discriminative stimulus internalism is a special case of the property identified by the guidance account, I think there is good reason to think it avoids the parsimony challenge.

This may seem unclear at first. For the guidance account appeals to the property of a discriminative stimulus, but the naturalist's candidate is a special instance of that property—that of a privileged discriminative stimulus. Further, it seems possible to explain many normative beliefs and aspects of normative practice without supposing that any consideration is a privileged discriminative stimulus rather than simply a discrimina-

tive stimulus. Indeed, I may regard the consideration that it is raining as a very good reason for carrying an umbrella in some circumstance. Assuming the guidance account is correct, then to explain that belief we must presuppose that the consideration is a discriminative stimulus. But do we need to presuppose it is a *privileged* discriminative stimulus?

Probably not. But that does not mean that discriminative stimulus internalism falls to the parsimony challenge. For there are other cases where the concern is to explain more than just occurrent normative beliefs. In these cases the concern is broader: how might an agent's beliefs and practices change and stabilize? For instance, suppose after much deliberation and reflection, I come to believe—perhaps originally to my surprise—that whenever eating animal products contributes to the suffering of animals that consideration is a strong reason against eating animal products. This belief strikes me as an important discovery. Indeed, originally it may lead to a partial overhaul or reassessment of my identity, as I reflect on how much of my past eating habits have been grossly unjustified. That belief goes on to influence others. And it influences the choices I make and what I recommend to others across a wide range of circumstances I encounter throughout my life. And, it would have a similar influence on my thought and action across merely possible circumstances that I didn't but *might* have encountered. In contrast, other considerations do not play such a prominent, lasting, and modally robust role.

What explains these normative beliefs and practices of mine centering around this consideration? Arguably, the explanation must presuppose that that consideration is not merely a discriminative stimulus but a privileged discriminative stimulus. That is, to account for the change in my belief, the level of surprise and my regarding this belief as a discovery, and the modally robust role which this consideration had with respect to guid-

ing my responses, we must presuppose this consideration is a privileged discriminative stimulus.

Conclusion

So what do these preliminary reflections on the parsimony challenge suggest? They suggest that our normative beliefs and practices seem to constitute evidence for some variety of realism about first-personal reasons. But, of the two realist theories we've considered in this chapter, they only provide evidence for discriminative stimulus internalism. To account for normative beliefs and practices, we do not need to assume that any consideration possesses any non-natural property of being a reason. But to account for important aspects of our normative beliefs and practices, such as how those beliefs might change and have lasting and robust effects on an agent's identity and responses, it seems we need to assume that some considerations are privileged discriminative stimuli.

Thus we have a limited presumption in favor of realism. Our normative beliefs and practices *do* seem to provide evidence for the existence of reasons. But not reasons according to just *any* conception of reasons. Rather, the presumption is for reasons which are privileged discriminative stimuli. Insofar as those beliefs and practices provide evidence for the existence of reasons, that is *because* they provide evidence for the existence of privileged discriminative stimuli.

Chapter 4

The Extension Objection

Introduction

In this chapter, I discuss the extension objection, which is an objection to all varieties of internalism, including discriminative stimulus internalism. According to internalism, if some consideration is a reason for you to respond in some way, this is always because of the following: by responding in this way given this consideration, you promote some actual or hypothetical preference of yours.[50] For example, suppose you can prevent me from drowning and that this is a reason for you to toss me a life preserver. If internalism is true, this must be because, by tossing me a life preserver given that you can prevent me from drowning, you promote, say, your actual or hypothetical preference that I survive, or that you cash in on your fifteen minutes of fame.

The extension objection appears to be an obvious and devastating objection to internalism. At first glance, it seems that some considerations are your reasons, and others are not, *no matter what you prefer*. It is deeply plausible, for instance, that the consideration that you can prevent me from drowning is a reason for you to toss me a life preserver. Your preferences do not affect the normative status of that consideration. So, since internalism clearly fails to account for the correct extension of your reasons, it must be

[50] For brevity I use the term 'reasons' rather than 'normative practical reasons'. I set aside concerns over normative reasons that are not practical (such as epistemic reasons), and over reasons that are not normative (such as explanatory or motivating reasons).

false. Or so the objection goes.[51]

In metanormative theory, the extension objection is widely regarded to be strong evidence against internalism and in favor of externalism. It is not uncommon to find philosophers with externalist sympathies endorsing it as "obvious" and "decisive".[52] Nor is it uncommon to find externalists—eager to mount new, powerful objections of their own—first giving the extension objection a perfunctory hat tip.[53] (And, however clever these new objections may be, it is natural to suspect that their plausibility derives in no small part from the plausibility of the extension objection itself.)

Internalists seem to appreciate its force too. This explains why, overwhelmingly, they seek to ground reasons in terms of ideal rather than actual preferences.[54,] Such a move does not dispatch the objection entirely, of course. But it does protect internalism against egregious instances of the objection.[55] And what about varieties of internalism that don't appeal to ideal preferences? They aim to avoid the objection in other ways.[56]

Not to be left out, proponents of other foundational views—such as constructivism and expressivism, as well as those that are more difficult to classify—all boast of

---

51 I largely focus on versions of the extension objection that concern reasons. But the objection may concern any normative or evaluative phenomena.

52 See for example Parfit (2011, 81), McGinn (1997, 10-12). See also Foot (2001, 60-61), who endorses the objection—against her earlier work in Foot (1972)—while developing an Aristotelian metanormative view, which has elements of both internalism and externalism.

53 Scanlon (2014, 4-5), Shafer-Landau (2003, 41, 185-188).

54 Mill's (1861, ch 2) competent judges test can be seen as a proto-idealization maneuver, which Sidgwick (1907, bk 1, ch 9) builds upon. More fully developed positions are found in Firth (1952), Brandt (1979), Williams (1981), Railton (1986), and Smith (1994, 1995). For important critiques of idealization accounts see Loeb (1995), Rosati (1995), Sobel (1994), and Velleman (1988). See also Johnson (1999, 2003) for an important focused discussion of competing tensions for internalist views.

55 Enoch (2005) raises the worry that internalists may have no good justification for invoking these idealization maneuvers. For similar concerns see Shafer-Landau (2003, ch 2). See Sobel (2009a) for an internalist's reply to this concern. See also Johnson (1999, 2003) and Plunkett (2016) for discussion.

56 See for instance Schroeder (2007, ch 5-7). For objections to Schroeder's solutions see Enoch (2011b), Gregory (2009), and Sobel (2009b). For a particularly interesting recent approach see Plunkett (2016).

their theory's apparent ability to avoid the objection.[57]

A particularly telling indication of the objection's significance is found in a recent, admirably frank remark by David Enoch, a prominent externalist.[58] In what he describes as a confession, Enoch reports accepting externalism not because of abstract arguments concerning ontology, semantics, or the nature of action, but for a "much less abstract, and perhaps even much less philosophical" reason. The reason? Externalism's rivals—which fail to be "objectivist in some important, intuitive sense"—have objectionable first-order normative implications.

Of course, highlighting the objectionable first-order normative implications of internalism is precisely what the extension objection purports to do. *It is the quintessential objection of this kind*. For Enoch, this fact alone is reason enough to reject internalism and endorse externalism—the only alternative, it seems, that is untouched by this problem. I join Enoch in the suspicion that he is not the only one to assign such a hefty dialectical role to the objection.

It makes sense, I think, for Enoch to characterize his attitude here as a confession—shouldn't there be some better, deeper, more intellectual, and less obvious reason for rejecting internalism and accepting externalism? How could the matter be this straightforward?

---

[57] Examples of constructivists include Korsgaard (1996, ch 4), (2009, ch 4), and Markovits (2011, 2014). Examples of expressivists include Blackburn (1984, ch 6), (1993), (1998, appendix), and Gibbard (1990) and (2003). Chang (2013) suggests a "hybrid" theory that blends elements of different metanormative approaches in an attempt to avoid the extension objection, among others. See also Foot (2001)'s sketch of an Aristotelian theory, where arguing that such an approach might avoid instances of the objection is a looming concern.

[58] Enoch (2010, 111-112). In this work and in his (2011a, ch 2-3), Enoch offers new objections which aim to highlight troubling first-order normative implications of non-externalist theories. Despite their novelty, I think these objections may be little more than circuitous versions of the extension objection. See footnote 6 above.

My aim here is to explain why the extension objection is seriously misunderstood. In section 1, I observe the objection's structure. It is clear that the objection always rests on a substantive normative premise. The question is whether this premise is true. Clearly it is no good, dialectically, to simply assume that the premise is true. In section 2 I consider the first problem with this objection, which is that it cannot be defended in an unproblematic way. I illustrate this by considering a version of the objection that might be offered on behalf of unfamiliar variety of externalism. Focusing on this unusual instance of the objection gives us the critical distance necessary to appreciate the limitations of the objection more generally. Despite prevailing assumptions, the extension objection cannot possibly be strong evidence against internalism and in favor of externalism.

But matters may be worse for externalism. For, as I explore in section 3, given the guidance account, the real concern behind the extension objection may have little to do with extensions at all. Rather, the concern may be more practical: we want considerations to be reasons only if we can relevantly respond to them by exercising our procedurally or instrumentally rational capacities. Of course, this constraint on reasons is something internalists have embraced all along. And, although they don't realize it, I argue it's what externalists really want from a theory of reasons too.

The extension objection is not the decisive evidence against internalism and in favor of externalism that it is widely taken to be. Either its status as evidence is unclear, or else—to great surprise—it is actually evidence against externalism and in favor of internalism since it supports a constraint that is best explained by internalism.

4.1 How the Extension Objection Works

Instances of this objection are easy to come by and feature the usual cast of peculiar characters—sadists, bigots, depressives, and those with other bizarre concerns.[59] Without loss of generality, we can consider Derek Parfit's recent and especially forceful version which centers on the normative significance of agony.[60] His version begins with the following scenario.

> *Agony Scenario*
> Although you may soon be in agony, you can easily avoid it. However, your preferences are bizarre. Being in agony would not undermine any actual or ideal preference of yours, and avoiding the experience would not promote any such preference.

Parfit then argues as follows. If internalism is true, then the fact that agony feels the particular way it does—that is, the fact that it is agonizing—is not a reason for this imagined version of you to try to avoid being in agony. Nor is it the case that the imagined you ought to try to avoid being in agony, all things considered. Intuitively, however, things are quite the opposite. Indeed, you *ought* to try to avoid being in agony all things considered, and the fact that agony is agonizing *is* a reason for you to try to avoid being in agony. So, internalism is false—or so it seems to follow.

Each instance of the extension objection shares this two part structure. It begins with some scenario, like the one above, and is followed by a two premise argument which references that scenario and takes the following form.

The Extension Objection

---

[59] As Sharon Street (2009) nicely puts it, "[s]ome strange characters inhabit the world of metaethics." For examples of such characters and conditions see Gibbard (1990, 165-176), (1999, 145); McGinn (1997, 21); Parfit (1984, 124); Shafer-Landau (2003, 41, 185-189); and Thomson (2008, 253-256). See Street (2009), Railton (1998), Plunkett (2016) for critiques of the extension objection which differ from the one I pursue here.

[60] I have made unproblematic changes to the case for ease of discussion. For the original case and argument, see Parfit (2011a, ch 3, sec 11).

Extension premise
In the given scenario, some normative property is instantiated. For instance, some consideration *is a reason for* the agent to respond in some way. Or responding in this way is what the agent *ought* to do, all things considered.[61]

Conditional premise
If the target theory is true, then, in this scenario, this normative property is *not* instantiated. For example, the consideration is *not* a reason for the agent to respond in that way. Or it is *not* the case that responding in that way is what the agent ought to do, all things considered.

Conclusion
So, the target theory is false.

Since the argument is valid, the question is whether its two premises are true. We may assume that the conditional premise is true. For this is easy to ensure provided the commitments of the target theory are made explicit and the scenario is well-constructed. On this assumption, the soundness of the extension objection rests entirely on the truth of the extension premise.

4.2 The First Problem

So, are the extension premises in relevant instances of the objection true? Well, to many people, these premises certainly *seem* true. The extension premise in Parfit's version above, for instance, seems deeply plausible. And Parfit's version of the objection is perfectly typical in this regard.

Indeed, the deep plausibility of its extension premise is what makes the extension objection itself so convincing. This is precisely why agents with eccentric or utterly bizarre preferences frequently appear in the objection's scenarios—*clearly*, it seems, in-

[61] Again this premise might concern other thin or thick normative properties, such as permissibility, wrongness, reasonableness, goodness, honor, virtue, bravery, treachery, evil, etc.

ternalism has implausible implications about what reasons *these* peculiar agents have.

This is also why externalists don't ever raise instances of the objection which feature implausible extension premises. Presumably you have no relevant preference for torturing puppies for its own sake. If internalism is true, then—given the appropriate details—it is not the case that you ought to torture puppies for its own sake. It would be entirely unconvincing for an externalist to suggest that, since you ought to torture puppies for its own sake, internalism must be false. After all, it's deeply implausible that you ought to do that.

Ultimately, then, the deep plausibility of the extension premise itself explains the great dialectical weight the extension objection as a whole carries in the debate between internalism and externalism. In the extension objection we have a compelling reason to favor externalism over internalism—a reason which rests upon a deeply plausible extension premise. Dialectically speaking, isn't this precisely as it should be? Should an objection, which is one of the strongest reasons for rejecting internalism and accepting externalism, rely on some deeply implausible premise instead?

No, of course not. But that is not the issue. The dialectical problem here is this. The extension objection is not a strong reason for favoring externalism and rejecting internalism. The presumption that it is stems from an uncritical acceptance of its extension premise—an acceptance which is unwarranted given that dialectic.

Let's take a more critical eye toward the extension premise. To do so, consider an instance of the objection that might arise from an unfamiliar externalist. Call this unfamiliar externalist the *paperclip externalist* since his normative theories place great non-instrumental significance on paperclip production, and comparatively little non-instru-

mental significance on much else.

The details of how this externalist came to hold this view need not occupy us too much. Perhaps, in their planet's distant past, a now defunct artificial intelligence presence selected the paperclip externalist's ancestors to be disposed to have an abundant non-instrumental concern for paperclip production.[62] Or perhaps instead of having roots in artificial selection, this concern for paperclip production arose as a result of some natural selective process, like sexual selection—I leave the curious details here to the adventurous reader.

These paperclip people may believe other things are normatively significant. For instance they may care about the welfare of kin, truth, trust, and so on. And they may even believe there are often good, non-instrumental reasons for avoiding agony. Or maybe not. However, in at least some cases, and in all the cases that concern us here, they find it overwhelmingly obvious and deeply plausible that what matters most, normatively, is making paperclips.

Some of these paperclip people are externalists about normativity and one in particular is eager to object to both internalism and our familiar variety of externalism—henceforth *familiar externalism*—by having us consider the following instance of the extension objection, beginning with the following scenario.

> *Paperclip Scenario*
> You can either prevent yourself and your loved ones from experiencing agony or else not do so but thereby make one paperclip. Importantly, in this scenario, you are yourself—or, perhaps some more fully informed and procedurally rational counterpart. Thus, unlike in the agony scenario above, you are not some imaginary version of yourself who happens to have some bizarre preference set. Thus—

[62] My paperclip externalist might be the lucky survivor of the artificial intelligence gone awry found in Bostrom (2009).

> I hope—you have an overwhelming preference to try to prevent yourself and your loved ones from being in agony, and you have little to no interest in making paperclips for its own sake.

The paperclip externalist then argues:

> *Paperclip Extension Premise*
> In this scenario, the fact that you can make a paperclip by allowing yourself and your loved ones to experience agony is a decisive reason for you to do so. Further, this is what you ought to do, all things considered.
>
> *Paperclip Conditional Premise*
> If either internalism or familiar externalism is true, then, in this scenario, it is not the case that you ought to allow yourself and these loved ones to experience agony. Also, it is not the case that the fact that you can make a paperclip by allowing yourself and your loved ones to experience agony is a decisive reason for you to do so.
>
> *Conclusion*
> So, both internalism and familiar externalism are false.

This argument is valid and we may assume its conditional premise is true. The fate of internalism and familiar externalism, it seems, hinges on the truth of the paperclip externalist's extension premise.

To argue that his extension premise is true the paperclip externalist can appeal to at least four facts. First, to him and many other paperclip people, this extension premise *seems* true. Second, he can amplify this claim with an appeal to doxastic and epistemic modalities. The premise does not merely seem true, rather, it seems that it *must* be true; it seems true so *clearly* and *obviously;* it is *deeply compelling* or *overwhelmingly plausible;* he is *highly confident*, *sure*, or perhaps even *certain* that it is true. Third, the paperclip externalist adds, it seems that the extension premise is *necessarily* true and *couldn't possibly* be false. This apparent necessity is of a broadly metaphysical sort, though exactly

which kind need not concern us here.[63]

Fourth, the paperclip externalist might appeal to the distinct way this extension premise seems to relate to other normative propositions which strike him much like this one does. For instance, the extension premise might be entailed by a normative principle which the paperclip externalist finds plausible, perhaps due to its explanatory power with respect to other normative propositions. On the paperclip conception of utility, for instance, making paperclips is what matters most. And so the extension premise might be entailed by the principle that, in all circumstances, and all things considered, every agent ought to try to maximize expected utility. On the paperclip conception of virtue, the virtuous agent creates paperclips rather than prevents agony, and so some principle of virtue entails the extension premise. And according to an influential version of paperclip contractualism, agents ought to act only in ways that are permitted by principles that no agent could reasonably reject. Since paperclip people regard it highly unreasonable not to make a paperclip when it is possible to do so, some relevant principle entails the truth of the extension premise.

Alternatively, the paperclip externalist might claim that the extension premise is not so much a consequence of some normative principle, but is more akin to an axiom within the set of normative propositions he accepts. From this broadly foundationalist perspective, the paperclip externalist reports, the extension premise seems to explain various other apparent normative truths. These include the apparent truths that, often, there is greater reason to destroy one paperclip to save five, that it sometimes permissible for one to abort some paperclip production process, and that, one ought to donate a substantial

[63] Perhaps, for instance, the necessity is a distinctly normative kind as Fine (2002) argues.

portion of one's income to aid agencies like Réparations Trombone Sans Frontières or, because it's more cost effective, OxClip.

It will be helpful to have a brief way of referring to these four facts. Call some normative proposition, like an extension premise, *especially compelling* to someone if and because, to him, it seems true, obviously so, as a matter of metaphysical necessity, and because it seems to relate in relevant ways to other normative propositions which also seem true to him.

The paperclip externalist might then argue that because his extension premise is especially compelling to him, it *is* true, or at least, this is *strong evidence* that it is true, or *good reason for everyone to believe* that it is true.[64] Others who find his extension premise especially compelling will likely agree. Case closed! Internalism and familiar externalism are done for, right?

Not so fast. The paperclip extension premise is not especially compelling *to us*. If anything its negation is. And so it is no surprise that *we* are not persuaded by this defense.

But this is no exercise in rhetoric—the paperclip externalist is not in the business of trying to persuade for its own sake. As he sees it, the fact that his extension premise is not especially compelling to us is explained by some epistemic or rational shortcoming on our part. If only we were not so normatively blind, we would recognize that, in the Paperclip Scenario, there *is* decisive reason to make the paperclip and this is what we *ought* to do! We don't recognize this obvious fact since our normative beliefs are askew—so

---

[64] There is a weaker interpretation of his argument on which the fact that some normative proposition is especially compelling to some agent A is a reason for A to believe it is true, but not necessarily any good reason for arbitrary agents (who may not find the proposition especially compelling) to believe it is true. I set aside discussion of this weaker argument here. This weaker argument would not carry the dialectical weight the extension objection is taken to have.

askew, it seems, that even upon reflection, we still don't grasp the obvious normative truths. In contrast, his normative beliefs are properly attuned to the normative facts. And when some normative proposition is especially compelling to him, this is good evidence that the proposition is true. Or so the paperclip externalist insists.

And, I take it, he's *right* to so insist—at least, given his commitment to paperclip externalism. After all, how could it be that his finding the paperclip extension premise to be especially compelling constitutes good evidence for its truth unless *our* normative sensibilities are askew?

Further, setting his commitment to paperclip externalism aside, is his insistence on this point adequate? Given the dialectic of this foundational debate about reasons, is the fact that the paperclip extension premise is especially compelling *to him* enough to constitute strong evidence against internalism and familiar externalism, and in favor of paperclip externalism?

Certainly not. The familiar externalist might offer an extension objection which, if sound, would undermine paperclip externalism. And to defend its extension premise, the familiar externalist could clearly provide an analogous argument regarding his epistemic superiority over the paperclip externalist. If only the paperclip externalist's normative beliefs were properly attuned to the normative properties and truths, he would find it especially compelling that agents ought to try to prevent themselves and their loved ones from being in agony instead of not doing so to make some paperclip. Although you and I may regard familiar externalism as the more plausible view, dialectically, the two externalisms are certainly at an impasse with none enjoying any advantage over the other.

Rather than seeking a way to justify some extension premise by appealing simply

to other normative propositions that he accepts, the paperclip externalist might appeal to some partially external mode of justification. Here, explanations of normative beliefs may seem particularly relevant. According to the evolutionary explanation of normative beliefs, for instance, it is possible to provide an explanation of why some agent is disposed to have some normative beliefs without appealing to any normative truths. Given various evolutionary forces, like natural selection and genetic drift, we are now disposed to believe that avoiding agony is far more important than making paperclips, while the paperclip people are now disposed to believe the opposite.

So what now? Both people and paperclip people seem to have equal claim to having some evolutionary heritage which bestows upon them more accurate normative belief forming capacities. Each may claim they descend from a species that was selected in part for being rational.[65] Being rational, we may suppose, involves, among other things, being relevantly disposed to have true normative beliefs. Although the theorists would agree to that innocuous unpacking of 'rational', they would fill it out differently by specifying which normative propositions, and so which normative beliefs are true. It is far from clear how either the familiar externalist or the paperclip externalist can argue that only they possess the appropriate evolutionary pedigree in any uncontroversial way.

Similar comments apply to the more localized explanation of reason beliefs supplied by the guidance account from chapters 1 and 2. According to that account, an agent's reason beliefs are explained in terms of his instrumental associations, or non-normative beliefs and preferences. Assuming the guidance account is correct, both kinds of

[65] Parfit (2011, ch 32, sec 114) calls this the Darwinian Answer, and offers it as part of his response to the causal challenges to normative and mathematical belief found in, e.g., Street (2006) and Field (1998, 396). While I doubt his proposal could succeed, I set that issue aside for the sake of illustrating my broader point.

externalist might insist that this does not undermine the truth of those beliefs. For again, they are rational agents. A rational agent is one who forms preferences for those propositions that there are normative reasons to prefer, and in proportion to the weight of the reasons for preferring them. But, as before, the externalists differ with respect to which normative propositions are the ones there are reasons to prefer. The paperclip externalist will insist it is rational to prefer making paperclips over preventing people from being in agony, and the familiar externalists will disagree. As before, this appeal to external factors that might influence normative beliefs seems to give neither side any clear advantage.

What should we make of this apparent impasse between the familiar and paperclip externalists? Suppose the familiar externalist suggests that although he and other externalists may be at an impasse, his extension objection still affords some kind of dialectical advantage over the internalist.

Why might this be so? One possibility is that since the paperclip people differ too radically from the kinds of agents we happen to be, we need not be concerned with a theory of normativity that applies to them. This suggestion would be too great a concession to the relativist features of internalism. The paperclip people are instrumentally rational agents by stipulation. They must have reasons and obligations of some kind. If they differ too radically from us to have the same non-instrumental reasons and obligations as we do, this is plausibly due to their different normative priorities. Assuming the guidance account is correct, this in turn is due to what propositions they are disposed to prefer.

If radical differences in preference might make some agents have different reasons and obligations than we do, then less radical differences in preference might do the

same. But this is precisely what internalists suggest. Once the familiar externalist has conceded that the paperclip people may have different reasons and obligations, it is not clear why Parfit's hypothetical version of you, or Gibbard's anorexic and Caligula could not also have different reasons because of their different preferences.

Instead of ignoring paperclip externalism, the familiar externalist might concede that the two theories face a stalemate with respect to each other. But familiar externalism still enjoys some kind of advantage over internalism since, unlike paperclip externalism, internalism is a departure from familiar externalism in the following sense. Both internalists and familiar externalists begin with broadly similar normative beliefs. Internalists come to accept a theory which requires that they reject or at least withhold any commitment to some of those pre-theoretic normative beliefs. For instance, internalists come to find some plausibility in the suggestion that Caligula ought to torture his subjects for fun, that the anorexic ought to starve herself to death, and that if you have an unusual preference set, the fact that agony is agonizing might not be a reason for you to try to avoid being in agony. So, of course internalists no longer find the familiar externalist's extension premises plausible. But the fact that both he and the familiar externalist may have found such premises plausible and that his current view requires that he reject those pre-theoretic propositions is a liability.

This suggestion is not promising in at least three respects. First, in general it is not clear why we should cling to pre-theoretic commitments come what may. Pre-theoretically, perhaps, it seemed that the sun orbits the earth and that there are no imaginary numbers. But in each case, inquiry provided reasons to believe otherwise.[66] We should pre-

[66] "As late as the sixteenth century we find mathematicians referring to the negative roots of an equation

sumably cling to some belief if there is not enough reason to abandon it. The claim here is not that the internalist's rejection of certain pre-theoretic commitments is certainly an improvement. Rather, the point is simply that the externalist must defend the suggestion that rejecting those pre-theoretic beliefs is a liability. How can he do this without assuming, at some point, that those pretheoretic beliefs are true? Reasoning has to start and stop somewhere, to be sure—there is nothing wrong with making assumptions. But there is something wrong with making assumptions in the context of a debate where the truth of what's assumed is clearly at issue, where there are grounds for skepticism, and where we aspire for something beyond epistemic complacency.

Second, we may suppose there are internalists who share no relevant "pre-theoretic" normative beliefs with familiar externalists. Paperclip internalists, for instance, are not at all tempted to find the familiar externalist's extension premises plausible. So, even if some internalists may be faulted for rejecting some of their pre-theoretic beliefs, it is not clear that internalism itself is similarly at fault.

Third, the stalemate with the paperclip externalist remains relevant even if we set it aside. The stalemate exists between the two theories since neither is able to justify in any uncontroversial way that his favored extension premise is true. The familiar externalist's inability to do so does not suddenly become an ability to do so when we set the stalemate with paperclip externalism to one side. Even if some particular *internalists* can be faulted for rejecting some of their pre-theoretic normative beliefs, the fault need not be any indication that *internalism* is false.

It is not at all clear that internalists, much less *internalism,* is compromised by the

---

as *fictitious* or *absurd* or *false*" Nahin (2007, 6). Italics in the original.

familiar externalist's extension objections. The objection may seem plausible because its extension premise seems true. But we must be careful to distinguish between seeming true or plausible and being true.

The problem here is general and applies to the internalist as well. Like the familiar externalist, the internalist might offer an extension objection against paperclip externalism. The crucial premise might be that, in the scenario described by the paperclip externalist, all things considered, you ought to try to prevent yourself and your loved ones from experiencing agony, rather than to make a single paperclip. Like the familiar externalist, the familiar internalist might try to defend that premise by appealing to the fact that it is especially compelling to him. But this move has all the problems seen above.

Perhaps from a distance it may seem that the internalist has a more favorable option on the assumption that one's preferences help to explain one's normative beliefs. For instance, presumably you have a decisive preference for preventing yourself and several of your loved ones from experiencing agony rather than for making a single paperclip. Further, we may suppose that you find the following normative propositions especially compelling: the fact that you can prevent yourself and your loved ones from experiencing agony is a decisive reason for you to do so; all things considered, you ought to try to prevent yourself and your loved ones from experiencing this agony rather than not do so but make a single paperclip. If your normative beliefs are explained by your preferences, and your preference in this case is so decisive, this explains why you find those normative propositions especially compelling. If internalism is true, so are these normative propositions.

So far, so good. But all this merely yields the conditional that *if* internalism is

true, then these two normative propositions are true as well. This is a hollow victory. Trivially, if internalism is true, then internalism's extension objection against the paperclip externalist is sound. But its soundness would not rely on the fact that its extension premise was especially compelling to you. The extension premise follows simply from the truth of internalism itself. Similar remarks apply all around. Trivially, *if* paperclip externalism is true, then both internalism and familiar externalism are false. The soundness of the paperclip externalist's extension objection would not rely at all on the fact that its extension premise was especially compelling to the paperclip externalist. No normative theory is so easily challenged by the extension objection.

4.3 The Second Problem

It is natural to understand the extension objection as highlighting an extensional desideratum which the correct theory of reasons must satisfy. It is especially compelling to you that certain considerations are reasons and others are not. Assuming your beliefs are accurate, any theory of reasons which fails to comport with this extension must be false. Internalism, of course, fails to comport. And from any externalist's perspective, at least, externalism doesn't. If this desideratum is correct, then this is evidence against internalism and in favor of externalism.

Why this talk of whether this desideratum is correct? Because, despite appearances, the extension objection may not be best understood as giving rise to an extensional desideratum. There is another way of understanding the objection, one that gives rise to a very different desideratum. Although this alternative desideratum is not obvious at first, it

may be much more fundamental. And because externalism fails to satisfy it, and internalism has no trouble satisfying it, this alternative understanding of the extension objection provides strong evidence against externalism and in favor of internalism. In this way, I will now argue, the extension objection is turned on its head.

Suppose the guidance account offered in chapters 1 and 2 is basically right. Then instrumental associations or sets of preferences and non-normative beliefs explain which considerations an agent believes to be reasons. But that explanation is technically restricted to actualized first-personal cases. I will need to consider an extension of the guidance account which applies to cases in which one thinks about the reasons of others. There is good evidence that thinking about the reasons of others—like the hypothetical versions of oneself in these wild Paperclip and Agony Scenarios—involves a relatively automatic attempt to simulate their experience or perspective, but one that is naturally constrained by the beliefs and preferences of the one doing the simulating.[67] Thus I will assume that when one thinks about what reasons other agents might have, one somewhat automatically simulates the perspective of the intended other. But, by default, this simulated perspective of the other is constrained by the instrumental associations that the agent doing the simulating already possess and is capable of possessing. So I will say that one's beliefs about the reasons of others are explained not by one's own instrumental associations, but by a *modest departure* from one's own instrumental associations.

If the guidance account and this way of extending it are basically right, a very different desideratum emerges from the extension objection. While on the surface the extension objection seems to require that a theory of reasons imply that certain considerations

[67] See for example Cushman (2015) and Lamm, Decety and Singer (2011).

are reasons, this may be a red herring. Oversimplifying this point a bit: it is a red herring since we only care about some set of considerations—about some extension—because those are the considerations that play a reason-like role for us. Similarly, the paperclip people only care about a different set of considerations—a different extension—because those are the considerations that play a reason-like role for them.

And of course, those considerations play that role simply because we have the preferences and non-normative beliefs that we do. What manifests itself as a superficial concern about an extension is, at bottom, a concern that reasons be things that can play a reason-like role for us. That's what you and I want from a theory of reasons. That's what the paperclip people want too. It's what we all have in common, even if we disagree on the surface about which considerations are reasons.

Thus, the idea here is that, although we may not know it, what we *really* want from a theory of reasons is that the things that are reasons according to this theory are the kinds of things that can play some relevant reason-like role for us, by guiding our responses.

Considerations that relate to our preferences and non-normative beliefs in the reason-like way not only guide our responses, but also give rise to our beliefs about reasons. Indeed, given that we believe these considerations *are* reasons, these are the considerations that—we think—would guide our responses, insofar as we are fully rational. Since 'rational' here is relative to none other than those very reason beliefs, it is best understood as referring to procedural or instrumental rationality.

Because this desideratum calls for a theory of reasons to make an agent's reasons be things that this agent can respond to, insofar as this agent is fully procedurally rational,

we can call this the *procedural desideratum*.[68]

An immediate point of emphasis is in order. In the overly simplified sketch of the procedural desideratum, I mentioned that *it* and not the extensional desideratum may be what we *really* want from a theory of reasons, although we may not know it. *That* was no oversimplification. Unlike the extensional desideratum whose influence on theory choice is readily apparent—favoring one extension, ruling out theories that conflict with this extension—, the procedural desideratum influences theory choice in a subtle and relatively opaque way. It is subtle because it still wields influence over which kinds of theories we find plausible, according to our starting stock of normative beliefs. The theories we find plausible differ greatly from the theories paperclip people find plausible because different considerations can guide our responses, given the kinds of preferences we have and might come to have through experience and reflection. Its influence is opaque because it is not obvious—and it may at times seem implausible—that our reason beliefs are influenced by our preferences.

Is the suggestion that there might be such a surreptitious desideratum at play, influencing our normative theorizing, a cheat on my part, or a strained attempt to inject metaethics with quasi-Freudian vigor? No, and no. If the guidance account from chapters 1 and 2 is roughly correct, then we've been unaware of something shared by all considerations we believe to be reasons. They all relate to certain of our preferences and non-normative beliefs in the same kind of way. It is no cheat, nor a misguided exercise in psychoanalysis to suggest that, surprising as it may be, what we really want from a theory of reasons is that the considerations that are reasons relate to our capacities for procedurally ra-

[68] Williams (1981).

tional agency in the right way. Rather, it's an alternative way of understanding what reason beliefs might be about—one that suggests a candidate constraint on the correct theory of reasons.

Of course, *if* the procedural desideratum is the correct desideratum to emerge from the extension objection, this spells big trouble for externalism. Consider paperclip externalism, for instance. On this view, in the Paperclip Scenario, the fact that you can make some paperclip is a decisive reason for you to do so. But suppose that your failure to have a decisive preference for making paperclips in this circumstance is no failure of procedural rationality on your part. Then given the procedural desideratum, this consideration cannot be a decisive reason for you in this circumstance. So paperclip externalism must be false. Similar remarks apply to familiar externalism. As long as the paperclip externalist is procedurally rational in decisively preferring to make paperclips, as irrational as it may strike us, making paperclips—and not, say, preventing others from being in agony—is what he has decisive reason to do. So familiar externalism must be false too.

While externalism fails to satisfy the procedural desideratum, at least some varieties of internalism may do more than just satisfy it. In particular, recall the ideal preferences variety of internalism. On this approach, if some consideration is a reason for some agent to respond in some way, this is because it relates to one of the agent's ideal preferences. Exactly which preferences count as being ideal varies from one version of this approach to the next. But all these approaches gesture toward a similar idea—ideal preferences are those that withstand the lessons of experience and the scrutiny of procedurally rational reflection, at least to some degree.

According to varieties of internalism—including discriminative stimulus internal-

ism—which incorporate preferences of this broad kind, such preferences help to ground, or fundamentally explain what makes considerations reasons for agents. This is a metaphysically strong claim—it is stronger than the procedural desideratum itself. Importantly, then, if an internalism of this kind is correct, it not only satisfies but also explains the truth of the procedural desideratum. Discriminative stimulus internalism may do so in a way that is even more appropriate since it may avoid the pitfalls of ideal preference versions of internalism—though it would have others since it makes reasons much more relative by focusing on more local idealizations. But I cannot explore that issue here.

If this understanding of the objection is correct, then, to great surprise, the extension objection is turned on its head. It is evidence for internalism—at least varieties of internalism capable of explaining the procedural desideratum—and evidence against externalism. Could the procedural desideratum, not the extensional desideratum, really be the correct or most legitimate desideratum to emerge from the extension objection?

At first glance it may seem easy enough to reject this idea. After all, aren't normative truths true by necessity? Suppose we happened to be like the paperclip agents, having bizarre preferences for making paperclips, and so also having correspondingly bizarre reason beliefs. Yet, the normative facts would remain the same. It would still be true that, in the Paperclip Scenario, the fact that we can make a paperclip is not a decisive reason for making a paperclip. It couldn't be since, in that scenario, the fact that agony is agonizing is a decisive reason for trying to prevent ourselves and our loved ones from being in agony.

Don't simple observations like this reveal that the extensional desideratum is all we care about? Don't they reveal that what we really want from a theory of reasons is that

it comports with some particular set of considerations, and not that it comports with, say, our procedurally rational preferences?

Not clearly. For remember, we are assuming that our reason beliefs are explained in part by our preferences and non-normative beliefs, or instrumental associations, as the guidance account suggests. If so, then reflections on cases like the one above do not clearly endorse the first desideratum over the second. For when we consider cases in which imagined versions of ourselves have radically different preferences, we—the ones *doing* the considering—are stuck with the actual preferences and non-normative beliefs we started with. So it is not surprising at all that we think our imaginary counterparts are the ones with the false normative beliefs. Our confidence in our *actual* reason beliefs may persist and manifest itself as a simple concern for an extension *tout court*—not an extension that relates to our procedurally rational capacities in some relevant way. But that doesn't rule out the possibility that we are unwittingly attracted to the procedural desideratum.

Perhaps a *better* way of testing which of the competing desiderata is the more legitimate one or correct one to emerge from the objection is to consider a case in which the normative facts are radically different from what we take them to be. Hence, suppose an oracle, concerned about our normative ignorance, reveals to us that paperclip externalism is true. We are given a great tome with all the true normative propositions. These apparent normative truths strike us as incredibly bizarre. Since they *are* the normative truths, however, our entire normative practice as a species, spanning centuries of apparent intellectual and practical achievements of seeming normative significance has, in fact, yielded no substantive normative truths. Abolishing slavery, broadening the liberties of a

greater variety of persons, extending our concerns to the welfare of non-human animals —none of this has brought us any closer to the way we ought or have reason to live. From the perspective of what truly matters, the efforts of our species have been a waste.

I take it that the oracle's revelatory tome would be no more than a curious anecdote. Of course some efforts of even the greatest normative minds have yielded normative falsehoods. Kant did not *rightly* think it is always impermissible to lie. And while defending the rights of man, Thomas Jefferson probably thought he had good reasons for holding slaves. Nevertheless, it simply could not be that what normative theorists have been up to for centuries could be completely undermined by the oracle's revelation. There is surely *some* sense to humanity's normative enterprise.

The same is true on a more personal scale. The oracle's tome could not wholly undermine every substantive normative belief and apparent lesson acquired over the course of a life. It really was foolish to play with those fireworks as a child, as these scars now attest. It surely was right to be assertive with the sleazy car salesman. And perhaps only time will tell whether moving abroad was a mistake—but the matter won't hinge on any foregone opportunity to make paperclips. Some of these normative propositions, which we meet with a sense of discovery, and which we find especially compelling, *must* be true.[69]

The scenario has a complement. Instead of bringing terrible news, the oracle may be delighted to reveal to us that many of the normative beliefs we hold deeply are true after all—paperclip externalism really is as absurd as it seems. As the tome of normative

[69] Gibbard (1990, 202) makes a related point in a similar case, asking, "Should we leave our judgments hostage to the normative sensibility of beings in a far galaxy?"

truths reveals, the human enterprise of normative inquiry *has* genuinely progressed. Now that our lives have definite significance, now that our normative inquiry has not been in vain, we can be relieved!

Of course, the oracle's pronouncement would warrant no such relief. We were not at all vexed by the possibility that paperclip externalism might be true. *That* possibility really *is* absurd. Our normative beliefs seem fully capable of defending themselves. They are not made more credible by the oracle or its tome, nor *could* they be. No tome could vindicate the significance of eradicating war, famine, disease, and poverty. We already know those to be worthy ends.

Thus, I take it, the proper—or at least natural—response to these oracle scenarios is a healthy mix of incredulity and pragmatism. On the one hand it is incredibly hard to believe that the normative facts could be so different from what now seems especially compelling to us. It is incredibly hard to believe that our entire normative enterprise could be so radically misguided.

And regardless of the merits of the oracle's pronouncements, what could we really do in response, anyway? We already have deeply held beliefs about our reasons. These beliefs withstand the scrutiny of our best attempts at procedurally rational reflection. It would be foolish, it seems, to make any serious effort to study much less follow the apparent truths of the oracle's tome. It makes considerably more sense to carry on with our lives and our normative practices in broadly the way we do now.

This natural reaction provides some evidence that, of the two desideratum, the procedural desideratum is most fundamental. Why is it so hard to seriously consider that radically different considerations could be reasons? Why would it be foolish and fruitless

to try to be guided by them? Because those considerations couldn't possibly be reasons *for us*. And they couldn't be reasons for us, they couldn't play that role, because they fail to relevantly relate to our preferences in such a way that, through the exercise of *our* procedurally rational capacities, we could respond to them. They are not the kinds of things we could ever believe to be reasons or treat as reasons insofar as we exercise our capacities for procedurally rational agency.

Of course, *some* evidence is not *conclusive* evidence. It is possible to interpret the natural response to the oracle scenarios in a way that does not favor the procedural desideratum. True, it *is* difficult to imagine the normative facts to be otherwise—at least radically otherwise. But this need not be any evidence of an unwitting attraction toward a conception of reasons which satisfies the procedural desideratum. Maybe it is just a byproduct of what normative beliefs are like and not any deep indication of what normative reasons are like.

Perhaps. But this is a tenuous proposal for the externalist to rely on. For, in all likelihood, the externalist only accepts this proposal with a major qualification. If the difficulty in imagining normative facts to be so radically different is *simply* a byproduct of what normative beliefs are like, and if not every possible agent with normative beliefs has true normative beliefs, then some possible agents are normatively hapless. Their normative beliefs are false, and the normative truths are radically different from those these agents accept. Indeed, *we* might be normatively hapless.

But wait. Might we really be normatively hapless? If the externalist is intellectually honest, here, he should bite this bullet with zest. He can't rule out the possibility—not in any clear way. But few externalists will be so bold. Instead, they will bite it only

tentatively insisting that, while *some* agents may be normatively hapless, *we* are not those agents (and so, not to worry!). But this desperate qualification would seem to reveal a persistent attraction for the procedural desideratum. How else can the externalist reasonably explain that we are not the hapless ones? On what grounds? He could cite the very considerations that he believes to be reasons as such grounds. But so too can the paperclip externalist, and many other bizarre externalists besides. But since these considerations are those that play the reason-like role for these agents, aren't they once again revealing their attraction for the procedural desideratum? By trying to resist its allure, the externalist can't help but affirm it.

Conclusion

I have argued that the extension objection has been greatly misunderstood. Many normative theorists regard it as constituting clear, compelling evidence in favor of externalism and against internalism. This is far from the case.

On the assumption that the objection might favor externalism at all, it is not clear how it does so without making some substantive, dialectically contentious assumption either about whether some consideration is a reason, or whether something is evidence that it is.

And most importantly, the extension objection may not even favor externalism at all. Its apparent extensional desideratum is not the only possible desideratum to emerge from the objection. For suppose our beliefs about whether some consideration is a reason are explained by the preferences and non-normative beliefs that explain why considera-

tions guide our responses in reason-like ways. Then the objection's ultimate concern may be for an agent's reasons to be the kinds of things he can respond to by exercising his capacity for procedural rationality. Maybe what we really want from reasons are considerations that we can treat as reasons, through our thought and action, insofar as we engage in reasoning. Maybe what we want to avoid from a theory of reasons are reasons that are wholly incapable of playing this role, given how we might reason, reflect, and deliberate while still being the kinds of agents we are with the kind of procedural rationality we are capable of.

The evidence that this procedural desideratum is the correct desideratum to emerge from the extension objection stems from how difficult it is to accept the possibility that our entire normative enterprise may be ill-conceived and that the normative truths may be radically different than we imagine. The reason it is so difficult to imagine that our reasons could be so different, I suggest, is because they could not be. An agent's reasons must be the kinds of things that can engage his capacity for procedural rationality. What we want most from a theory of reasons—though we may not realize it—is for reasons to be the kinds of things that can guide our thought and action, in reason-like ways, insofar as we, as practical reasoners exercise our capacities for practical reasoning. This is something internalists have long realized. And it is something externalists—and many others besides—don't yet realize they want most.

Chapter 5

The Normativity Objection

Introduction

According to what we may call the *normativity objection*, the property of being a reason cannot be natural since normative properties are fundamentally different from natural properties. Some of the most compelling and important arguments against naturalism about reasons—including ones made by Sextus Empiricus, David Hume, Henry Sidgwick, G. E. Moore, and, more recently, Jonathan Dancy, Guy Kahane, Thomas Nagel, and Derek Parfit—may be versions of the normativity objection.[70]

To establish this conclusion these arguments appeal to a wide range of considerations ranging from intuitions about reasons, possibility, meaning, concepts, knowledge, cognitive significance, and content. Yet, I suspect that, at their core, they rest upon what David Enoch calls the just too different intuition: the fact that the property of being a normative reason seems fundamentally different from any natural property is good evidence that the property of being a reason is not identical with or reducible to any natural property.[71] Although the normativity objection remains influential, the broad contours of a compelling defense against it seem within reach.[72] Consider the Fregean proposal that, in

---

[70] Sextus Empiricus (1976, 249). Hume (1738). Sidgwick (1907, 396). Moore (1903, ch 1). Dancy (2006). Kahane (2010). Nagel (2002). Parfit (2011). McGinn (1997, ch 2-3).

[71] Enoch (2011, chs 3 and 5).

[72] Wedgwood (2007, 2013). Not all non-naturalists find this argument compelling. See for instance Wedgwood who seems to agree with naturalist's about its limitations.

some sense, concepts are "modes of presentation" for properties they may represent.[73] If that's right, then it is possible that the concept of a reason and the concept of a given natural property present that same property in fundamentally different ways. As a result the property of being a reason will seem to be different from any given natural property. Yet normative properties may be natural after all.

But naturalists have not reached far enough. As Parfit's versions of the normativity objection reveal, it is not enough for naturalists to appeal to this familiar Fregean strategy. For while that strategy may account for the apparent but prosaic difference between properties like being water and being H2O, it does not seem up to the task of accounting for the apparently fundamental difference between properties like being a reason and being a discriminative stimulus.

In this chapter, I reach all the way and offer this defense in detail. In section 1 I canvass some arguments which may be instances of the objection and highlight how they may rely on the just too different intuition. In section 2 I present the naturalist's Fregean defense and Parfit's criticism of it. This paves the way for a more detailed defense in Section 3. The core of this defense relies on identifying a unique feature of normative concepts—their inconspicuous practicality. This feature accounts for the just too different intuition in the right way, by meeting Parfit's criticism.

5.1 Possible Instances of The Normativity Objection

The normativity objection aims to establish that no normative property—such as the property of being a normative practical reason—is identical with, grounded by, reducible

[73] Frege (1948).

to, or otherwise explained by any natural property.[74] I suspect many important arguments against naturalist varieties of realism about reasons rely on the just too different intuition for support.[75] Applied to the case of normative reasons, that intuition is as follows.

> *Just too different intuition*
> The fact that the property of being a normative practical reason seems fundamentally different from any natural property is good evidence that the property of being a reason is not identical with or reducible to any natural property.

I'll now quickly canvass some possible instances of this objection and highlight how the just too different intuition may be driving them.

Begin with David Hume's famous passage about the is-ought gap. Hume expresses surprise when, from propositions about natural properties, writers conclude that propositions about normative properties are true, and cautions against such reasoning. It is natural to interpret Hume's remarks as a version of the normativity objection. This is surprising since the normative concepts express "some new relation"—like the relation of being a reason. Yet it "seems altogether inconceivable" that some consideration might instantiate "this new relation" of being a normative reason in virtue of its instantiating any natural properties or relations "which are entirely different from it".[76] Clearly the just too different intuition may underlie Hume's concern here: the fact that it seems inconceivable that the property of being a reason might be some natural property is good evidence that it isn't.

---

[74] For simplicity I will only discuss the normative property of being a reason rather than normative properties in general. I will assume that others are discussing reasons as well, even when they were in fact discussing obligations or values. But I think my discussion could be extended to cover other normative properties. Versions of this kind of objection might also aim to establish that no evaluative property is a natural property. With further work, I believe the same strategy I pursue could succeed in the evaluative case as well.

[75] I suspect the just too different intuition may have a hand in arguments besides the normativity objection. Enoch seems to agree (2011, ch5)—he thinks it plays a role in Moral Twin Earth cases. It seems to pay a role in Rosati (1995a) and Harman (1986) as well.

[76] Hume (1738, bk 3, pt 1, sec 1).

G.E. Moore's famous remarks could be plausibly driven by the just too different intuition as well.[77] After all, he seems to think that the naturalistic fallacy doesn't arise when dealing with at least some natural properties. And this supports a reading on which whatever puzzle he thought to exist was a distinctly normative one. It is plausibly at the heart of his worries about the inability to define normative concepts in terms of natural ones, and the ability to genuinely question whether a consideration is a reason on the basis of its having certain natural properties. If this is something we can question then perhaps this is because the two kinds of properties seem to be just too different from each other.

Jonathan Dancy, David Enoch, and Derek Parfit have all recently appealed to the deliberative perspective of the agent in support of the normativity objection's conclusion.[78] Suppose you are curious about what your reasons are and what you should do. Jack tells you that the consideration that children are suffering from preventible disease is a discriminative stimulus for you. Jill tells you that the consideration that children are suffering from preventable disease is a reason for you to donate money to an effective aid agency that will relieve their suffering. It is plausible that Jill's answer provides you with importantly different information than Jack's answer. And perhaps that is evidence that the property of being a reason is importantly different in kind from any natural property. Of course these intuitions about importantly different information may be versions of the just too different intuition.

More recently, Guy Kahane presents us with a normative analogue of Frank Jack-

---

77 Moore (1903, sec 13).

78 Parfit (2011 ch24, ch 25, s 91). Dancy (2006), Enoch (2011).

son's Knowledge Argument about color.[79] Kahane's argument centers on a peculiar agent, Zeno, who has never experienced what it is like to be in the mental state of agony. But he knows various natural facts, including that when he experiences what it is like to be in agony, he will have a strong preference against being in this state—it will be a state he dislikes very much and that he is very motivated to escape. He also knows that, whatever the experience of agony is like, the consideration that agony feels that way has various natural properties, which the naturalist might offer as the candidate for the property of being a reason. Thus, for instance, Zeno knows that the consideration is a discriminative stimulus for him.

Kahane then appeals to an intuition about normative knowledge. Despite knowing that the consideration has any such natural property, Zeno does not know that the consideration is a reason to try to avoid being in agony. He gains that knowledge only after experiencing agony for the first time. Since Zeno knows various natural properties the consideration has but does not know of its normative properties, normative properties are not natural properties.

Although Kahane's argument has other interesting features, the just too different intuition seems to play a crucial role. After all, set aside knowledge and turn to normative thought more generally. To think that a consideration is a reason seems to be much different from thinking that it is a discriminative stimulus. And that incarnation of the just too different intuition is central to Kahane's claim that Zeno learns about a new property that the consideration has—one that cannot be a natural property.

More than any other contemporary philosopher, Derek Parfit has offered the

[79] Kahane (2010).

greatest number of normativity objections. His reductio ad absurdum fact stating argument is particularly interesting to consider since, in it, the just too different intuition may play a somewhat insidious psychological role.[80]

For the reductio premise, Parfit supposes that naturalism is correct. Then, for instance, for some consideration to have the property of being a reason just is for it to have some particular naturalistic property, like being a discriminative stimulus. Parfit argues that from this and other admittedly "drab and dreary" claims, it is *not* the case that for some consideration to have the property of being a reason just is for it to have that or any other naturalistic property.[81] Contradiction! So normative properties could not be natural after all.

Insofar as reductio arguments identify a problem with their target premise they often do so indirectly—the contradiction is merely a symptom of the underlying problem. So we should expect one of Parfit's drab and dreary premises to identify what naturalism's problem is more directly. So, what's the source of the problem?

Disappointingly, the argument's key premise is simply this: the fact that some consideration has some particular natural property "could not state [or be] a normative fact."[82] In other words, a consideration's having some natural property, like being a discriminative stimulus, could not be the same thing as its being a reason. Of course, that's

---

80 The fact stating argument is in Parfit (2011, vol 2, ch 26, sec 94).

81 Ibid, 339.

82 Ibid. 339. This is his premise (6), which Parfit agrees is the key premise. It reads "This non-normative claim could not sate a normative fact." In the context this non-normative claim is one which "[states] the same fact" as some normative claim by giving us "the same information". His phrase "the same information" is ambiguous but not in a way that affects my point. Regardless of whether by "the same information" he means something like *have the same content* or *be about the same property*, his key premise (6) clearly relies on the just too different intuition for support. Charitably, I assume that by "the same information" Parfit means *have the same content* or *be about the same property*. His premise (5) is obviously problematic if that is not what he means—if, for instance, information might be pragmatic.

just a denial of the reductio premise of naturalism. So, the reductio ad absurdum really is superfluous. What matters is Parfit's *defense* of that premise.[83]

Before stating that defense, it is worth highlighting that anyone who has the just too different intuition may already find Parfit's key premise plausible. Indeed, I suspect this is what drives Parfit himself to accept and seek to defend that premise. This bit of psychoanalysis becomes even more plausible when we find his defense of that premise to rest on yet another claim which relies on the just too different intuition. Here's that defense: a consideration's having some natural property could not be the same thing as its being a reason because, in that case, the property of being a reason would not be "a distinct normative [property]".[84] So, normative properties must be *distinct* from natural properties and that's why they cannot be natural? If Parfit has the just too different intuition, it's obvious why he might find this defense compelling—normative properties just seem importantly different from natural ones. But it is problematic since it is not an argument against the naturalist's position. It's simply a rejection of it.

What could possibly explain why Parfit guides us on this drab and dreary detour-by-reductio only to offer a question-begging argument against naturalism, which he fails to realize as such, and which he takes himself to defend by offering yet another question begging argument against naturalism? No teetotaler, I suspect that Parfit is under the influence of the just too different intuition.[85]

---

[83] Am I being uncharitable by suggesting that the form of the argument is a reductio? Perhaps. But if it is not a reductio, the argument's structure becomes straightforwardly question-begging. For, at least on one obvious recharacterization, the key premise—that normative properties could not be natural—would occur under no assumed reductio premise. That's not how Parfit seems to present it. And it would make the argument immediately question-begging since that's the intended conclusion. I think characterizing the argument as not-striaghtforwardly-question-begging is more charitable than the alternative and more faithful to his presentation.

[84] Ibid., 339-341 especially. The quoted phrase comes from a portion of his claim (X).

[85] Parfit's Triviality Objection relies on the intuition as well. I set that aside for now.

I take these quick observations to be suggestive rather than conclusive. Despite appealing to different considerations, it is plausible that each of these arguments is some version of the normativity objection. If so, and regardless of whether their proponents realize it, the most charitable interpretations of these arguments may rely on the just too different intuition for evidential support. Rather than begging the question, for instance, proponents of normativity objections would do better by straightforwardly embracing the intuition and explicitly acknowledging it as a premise in their arguments.

In what follows I will assume the just too different intuition is the key premise in the normativity objection. If this is right, this is a major liability for non-naturalists about normativity. For naturalists would be able to undermine all these arguments by providing an explanation of the just too different intuition that does not appeal to any non-natural properties.

5.2 The Naturalist's Fregean Defense and Parfit's Criticism

Naturalists appear to have an important defense against the normativity objection.[86] Following Gottlob Frege, suppose a concept is, in some sense, a "mode of presentation" for any property it may represent or be about.[87] Then distinct concepts may be about the same property. Yet, to one who is competent with each concept, it may seem that they are about different properties.

For instance, consider the concept of oxidane. Oxidane is a kind of hydride. A hydride is a type of chemical compound in which hydrogen is combined with another ele-

[86] Here are some people who at least suggest some defense along these lines. Gibbard (2003). Wedgwood (2013).

[87] Frege (1948).

ment, generally by way of an ionic, metallic, or covalent bond. Oxidane is the hydride such that hydrogen shares a covalent bond with two hydrogen atoms. The concept of oxidane is distinct from the concept of water. Water is the clear, potable liquid that flows in rivers and streams, and falls from clouds as rain, and sustains life on Earth. Plausibly oxidane and water are the same property. Yet it is possible to be competent with the concepts of oxidane and water without realizing that they are *about* the same property.

Putting the point in the picturesque terms, which Frege's phrase suggests, thinking that something is water may present it to your mind in a particular way, namely, one which brings to mind rivers, rain, and clear potable liquids. Thinking that something is oxidane may instead bring to mind the class of hydrides, the element oxygen, and the covalent bond between two hydrogen atoms and an oxygen atom. Because different things come to mind when thinking that something is water compared to when thinking that something is oxidane, it may seem that something's being water is neither the same as nor is reducible to its being oxidane. Yet as we may assume, the properties are the same. So, assuming the notion of modes of presentation can be filled in more rigorously—both linguistically but also psychologically—the fact that water may seem to be a different property than oxidane is not good evidence that it isn't.

The naturalist's Fregean strategy against the normativity objection is to appeal to an analogous maneuver to account for the just too different intuition. Just as the concepts of water and oxidane might seem to be about different properties when they are not, the same might be true for normative and natural concepts. So the fact that normative and natural concepts seem to be about different properties is not particularly good evidence that they are.

As Parfit has recently argued, this defense has a problem.[88] The Fregean strategy is generic. It promises to account for why some property seems different from another when it may not be—that's a general metaphysical concern that is not limited to the debate between naturalists and non-naturalists about normativity. Crucially, normative properties do not merely seem different from natural properties, they seem fundamentally different. So the Fregean strategy does not provide an adequate explanation of why normative properties seem different from natural ones in the fundamental way they do. And so the normativity objection still stands.

For instance, suppose you are competent with the concepts of water and oxidane but do not realize they are about the same property. Parfit grants that as a result of your competence with these distinct concepts, it may seem to you that they are about different properties. Yet, your competence with the concept water seems to leave open the possibility that water is or is reducible to oxidane. Although it is by no means obvious, it seems possible that something's being the clear, potable liquid that flows in rivers and streams just is or is reducible to its being the hydride involving a covalent bond with oxygen.
The concept of water tells us how the property of water fits among other properties. The chemical candidate of oxidane presents us with a property that may not simply fill that role, but also *explain* why it does so.

Parfit suggests that the case of normative concepts is different since the concept of a reason seems to rule out the possibility that it might be any natural property. But how? As one of several comparison cases, he has us consider whether the property of being a

---

[88] Parfit thinks this strategy and the views they support have various problems. See much of Parfit (2011, vol 2, pt 6, ch 25-26). I focus on what I take to be his most successful criticism.

river might be the same as the property of being a sonnet.[89] Perhaps, in some sense, we can conceive of the possibility that the property of being a river is the same as the property of being a sonnet. Still, we may suppose, there is some important sense in which these properties are not and could not be the same. Given what rivers are, they could not be sonnets. This possibility seems to be ruled out in some way, perhaps physically, metaphysically, conceptually, semantically. Though "*much* less obviously", Parfit insists "this... is the way in which" normative properties could not be identical with or reducible to natural properties.[90]

I think Parfit's criticism of the naturalist's strategy is important. The apparent difference between normative and natural properties is a distinct kind of difference. It is not like the comparatively generic difference between concepts of water and oxidane. So naturalists are too quick to dismiss the normativity objection on the basis of a general Fregean strategy.

To better understand the nature of normativity, naturalists should refine the Fregean strategy to address Parfit's criticism. To my knowledge naturalists have not done this.[91] I aim to remedy this oversight in what follows. The task is to identify a unique—but natural—feature of normative concepts. This feature must explain why normative properties seem to be importantly different from natural properties. If it does, the just too different intuition will have been accounted for without appealing to any non-natural properties. The normativity objection—in all its varieties—would be undermined, depriv-

---

89 Ibid vol 2, ch 25, s 91, 324.

90 Ibid 325. Parfit's italics.

91 Many philosophers develop views which allow them to develop the kinds of position I offer below. In a future work I hope to discuss how, even if they did so, my discussion offers a unique and complimentary way of doing so. Some of these philosophers include for instance Wedgwood (2013), Gibbard (2003), Brandom (1994), Egan (2012), Dreier (1990).

ing non-naturalists one of most compelling and frequently invoked arguments against naturalism.

5.3 Improving the Naturalist's Fregean Defense

The unique feature of normative concepts is that they are *inconspicuously practical*. Focusing on the concept of a normative practical reason, let's first turn to its *practicality*. Recall that, according to the guidance account, the concept of a reason is practical because the rule that regulates competent use of the concept of a reason is a function of the more basic rule that makes some consideration response-guiding. The output of the rule that allows a consideration to guide your response in a particular way is an instrumental association. That instrumental association makes it possible for you to not only believe that some consideration is a reason, but also for that consideration to guide your response.

For example, suppose you are disposed to believe that, in some circumstance, the consideration that your dog Fido has been hit by a car is a reason for you to run to his aid. You are so disposed because of some initial input mental state, which for simplicity, consists of a belief, preference, and a resulting instrumental association between token representations of some consideration and response. In this case, we may suppose the preference is that Fido survives and the belief is that whether Fido survives is contingent upon the consideration that he has just been hit by a car and your response of running to his aid. The association that forms is between the token representation of this consideration and response.

This input mental state does not merely dispose you cognitively, to believe that

the consideration is a reason for running to Fido's aid. Rather, and perhaps primarily, it disposes you practically. But how, exactly?

It is difficult to say precisely what this more practical way is. Here I tentatively offer the following proposal: the initial mental state disposes you *to consider responding in this way on the basis of this consideration*. This state (or process) of considering what to do and why is a practically oriented mental state which is or may be the precursor to intention. Thus the concept of a reason is practical in the sense that the rule that regulates this concept is part of a broader rule. This broader rule also regulates how the considerations and responses falling under your competent use of that concept factor into your practical reasoning, by offering candidate responses and bases for these responses when you deliberate about what to do.

I take it that this deliberative state exists. We often wonder what to do. Sometimes the answer seems unclear and multiple responses come to mind as candidates. When relatives have just arrived in town after a late flight, would it be better to let them rest or take them out for dinner and drinks? Typically, competing responses do not come to mind alone, but are backed by reasons. The consideration that the flight was a short one supports taking the relatives out, since the flight itself wasn't exhausting. Still, your relatives are jet-lagged—it is now 3AM at their point of departure. That's a reason to let them rest.

How strongly a consideration seems to support a response depends on the preference with which it is associated. You prefer that you do what they prefer most, and assume that they will want to rest. So the consideration that they are jet-lagged seems to support letting them rest more strongly than any reason supports taking them out. And as a result you find yourself with the candidate intention of letting them rest, if this is what

they want to do. When you discover they are not at all jet-lagged and are eager to go out, the consideration that they want to go out seems to strongly support taking them out.

The concept of a reason is practical since the rule that regulates its use is part of a broader rule which also regulates an important aspect of your practical reasoning: the responses which occur to you as candidates for choice are accompanied by considerations that you believe to be reasons for choosing and intending these responses. But the concept is *inconspicuously* practical since your competence with the concept does not depend on your awareness that the considerations to which you apply the concept serve this practical function. Indeed, in a kind of confabulation, insofar as you both believe that some consideration is a reason and recognize that it plays such a role, it may seem to you that it plays that role *because* you are rational and recognize *that* it is a reason.

In contrast, while some natural concepts may be practical, they are conspicuously so. For instance consider the concept of a first-personal practical intention—an intention about something you intend to do. It is possible that the rule that regulates your competence with this concept is also practical in the sense that it is part of a broader rule which also regulates some aspect of your practical reasoning. For instance, when your friend asks you to join him to watch a movie, your standing intention to meet your relatives at the airport comes to mind and prompts you to not join your friend, or to consider revising that intention. Further, and intuitively, you are competent with the concept of a first-personal intention only if you believe you intend to do something on the basis of some first-personal practical intention of yours. In this way, we may suppose the rule that regulates your competence with this concept is a function from an initial input mental state, consisting perhaps of some first-personal practical intention—to another output mental state,

a belief about something you intend to do. Yet because that initial state of an intention affects your practical reasoning, we may suppose that the concept of an intention is practical since the rule that regulates it is part of a broader rule that regulates an important aspect of your practical reasoning.

Yet the practicality of the concept of a first-personal intention is *conspicuous*. Your competence with the concept consists in part on your implicit beliefs about how you are practically disposed with respect to the things you intend to do. You do not competently use the concept of a first-personal practical intention if you believe you intend to meet your relatives at the airport tonight while acknowledging that there will be no point later today when you try to meet them there, and that you are completely indifferent whether you do.

As a consequence of being inconspicuously practical the phenomenal character plays a comparatively more significant role with respect to the rule governing their use. Thus, for instance, consider the phenomenal character of the state of intending to do something. It is distinct from other states, like hoping that something happens, or being excited. Yet if the only anchor we had for understanding what the concept is about were its phenomenal character, like normative properties, it too may seem just too different to be a natural property. Fortunately we have as a vital anchor its relational nature. Things you intend to do can be situated in the natural order as states of mind you have now which relate to responses you will attempt to do later.

As a result of being inconspicuously practical, the concept of a reason, like all normative concepts, has a distinct phenomenal character which explains its unique mode of presentation. When a consideration and response presents itself to your mind as a rea-

son for responding in some way, this has a phenomenal character that reflects the role that things falling under this concept play in your practical reasoning. In the case of reasons, the role is for the consideration to prompt you to consider responding in a particular way —more or less strongly—or for this response to come to mind as a candidate for what to do on the basis of this consideration.

Although things falling under this concept bear this important practical relation to you, the fact that they do so is inconspicuous to you. More generally, your competence with the concept does not rest upon your belief that this consideration and response have any other salient, natural property. As a result, as it seems to you, you base your belief that some consideration is a reason for responding in a particular way upon the distinctive way that a consideration and response present themselves to you. This distinct phenomenal character, which seems to have no basis in your recognition of any natural property had by this consideration, is the unique mode of presentation of the concept of a reason.

Because only normative concepts are inconspicuously practical, normative properties seem just too different from natural properties. Suppose you believe that the consideration that Fido has just been hit by a car is a reason for you to run to his aid. Could that consideration's being a reason to run to Fido's aid just be the same as its being a discriminative stimulus for you?

I assume that, intuitively, it doesn't seem so. And it's clear why that's the case. The concept of a discriminative stimulus is *not* inconspicuously practical and is more overtly theoretical. Thus when you suppose that the consideration that Fido has just been hit by a car is a discriminative stimulus for you, insofar as this has a distinct kind of phenomenal character, I suspect it is one you understand and can articulate in terms of con-

cepts such as *preference*, *contingency*, *reward*, and *cause*. This phenomenological character and resulting mode of presentation is in terms of these concepts. And it is markedly unlike what it is like when it seems that some consideration is a reason for responding in some way. *That* distinct phenomenological character is not primarily linked to concepts but seems, by comparison, much less conceptual. It seems different in kind. And a consideration's being a discriminative stimulus definitively seems to lack that phenomenological character. As a consequence, the normative property of being a reason seems just too different from the property of being a discriminative stimulus. Assuming the concepts of both a discriminative stimulus and of a normative practical reason are arbitrary for their kind, the explanation generalizes: normative properties seem just too different from natural ones because only normative concepts are minimally theoretical and inconspicuously practical.

Conclusion

Normative concepts are unique because they are inconspicuously practical. This makes them have a distinct phenomenal character which accounts for the psychological aspect of their unique mode of presentation. This mode of presentation is unique to normative concepts and makes normative properties seem fundamentally different from natural ones. The fact that normative properties *seem* different from natural ones is not particularly good evidence that they aren't natural. A simpler explanation, in the reason case, is that the guidance account of chapters 1 and 2 is true. That account explains why, when a consideration is a discriminative stimulus for some agent, it can seem to this agent to

have a distinct phenomenal character. A consideration's being a discriminative stimulus may be the same thing as its being a reason. But that need not be obvious to the agent.

Conclusion to the Dissertation

In the first two chapters I proposed and defended the guidance account, an empirical theory about what regulates the concept of a reason. In the remaining chapters I demonstrated how that account can be used to defend discriminative stimulus internalism, a philosophical view about reasons. Strictly speaking this defense is at times in sketch form and applies only to realism about first-personal normative practical reasons. Further, I have only compared it to a generic non-naturalist rival and not seriously considered anti-realism in detail. And of course, many positive details about the theory itself must be filled in. Still, it has key virtues which constitute evidence of its approximate truth.

First, the theory is explicitly structured around the property of a discriminative stimulus which figures centrally in the natural sciences and has correlates in artificial intelligence. If some version of naturalism about reasons is true, then since reasons are central to the lives of agents this is precisely what we should expect. Specifically, the property of being a reason should figure centrally in sciences concerning the behavior and apparent rationality of humans, non-human animals, and varieties of artificial intelligence.

Second, the theory seems capable of avoiding the parsimony challenge. The argument for that claim was the loosest but I think it is clear enough that the details are forthcoming. In any event, since the theory is structured around discriminative stimuli, and since they regulate the concept of a reason, it is clear enough how, compared to any realist rival, this theory is better suited to avoid the parsimony challenge. This is a key advantage. For, all else equal, it would emerge as the only realist theory that enjoys the support

of the presumption for realism. Insofar as normative beliefs and practices provide evidence of reasons, that is because they provide evidence of discriminative stimuli.

Third, in two key cases when discriminative stimulus internalism was challenged by some realist rival—that is, when there was a question about whether all else really was equal, the guidance account provided simple explanations which undermined the bases of these challenges. The defensive maneuver was similar in each case and seems perfectly capable of undermining other potential challenges to the discriminative theory which I have not considered, including, for instance, those based on apparent facts about disagreement, non-relativity, and the like. In other words, the guidance account is a versatile tool that explains in a relatively simple way why putative evidence against the discriminative stimulus internalism is not actually evidence against it.

Fourth, discriminative stimulus internalism will be able to avoid worries about epistemic access and content determination. This gives it a considerable advantage over views about reasons, like non-naturalism and varieties of naturalism that require convergence of judgment across agents, which are more prone to facing these difficulties.

To my knowledge, these virtues are unmatched by theories which depart from discriminative stimulus internalism in the case of first-personal reasons. To adapt a line from the younger Thomas Nagel, I conceive of rationality as a branch of psychology.[92]

---

[92] Nagel (1978, 3).

Bibliography

Anderson, Steven, Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio Damasio. 1999. “Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex.” *Nature Neuroscience* 2 (11): 1032-1037.

Baldassarre, Gianluca. 2011. “What Are Intrinsic Motivations? A Biological Perspective.” *2011 IEEE International Conference on Development and Learning* (2): 1-8.

Balleine, Bernard W., and Anthony Dickinson. 1998. “Goal-Directed Instrumental Action: Contingency and Incentive Learning and their Cortical Substrates.” *Neuropharmacology* 37, (4): 407-419.

Bandura, Albert. 1977. *Social Learning Theory*. Englewood Cliffs: Prentice Hall.

Baron-Cohen, Simon. 2011. *The Science of Evil: On Empathy and the Origins of Cruelty*. New York: Basic Books.

Beecher, Michael. 1988. “Some Comments on the Adaptationist Approach to Learning.” In *Evolution and Learning*, edited by Robert Bolles and Michael Beecher, 239-248. Hillsdale, New Jersey: Lawrence Earlbaum Associates.

Berridge, Kent, and Terry Robinson. 2003. “Parsing Reward.” *Trends in Neurosciences* 26, (9): 507-513.

Berridge, Kent, Terry Robinson, and J. Wayne Aldridge. 2009. “Dissecting Components of Reward: ‘Liking’, ‘Wanting’, and Learning.” *Current Opinion In Pharmacology* 9, 65-73.

Blackburn, Simon. 1984. *Spreading the Word*. Oxford: Clarendon Press.

———. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press.

———. 1998. *Ruling Passions*. Oxford: Oxford University Press.

Blair, James, R. J. R. 2002. “Neuro-Cognitive Models of Acquired Sociopathy and Developmental Psychopathy." In *The Neurobiology of Criminal Behavior*, edited by Joseph Glicksohn, 157-186. New York: Springer.

Block, Ned. 1986. “Advertisement for a Semantics for Psychology.” *Midwest Studies in Philosophy* 10, 615-678.

Boghossian, Paul. 1989. “The Rule-Following Considerations.” *Mind* 98, (392): 507-549.

———. July 24, 2011. “The Maze of Moral Relativism,” *New York Times*.

Bostrom, Nick. 2009. “Ethical Issues in Advanced Artificial Intelligence.” In *Science Fiction and Philosophy: From Time Travel to Superintelligence*, edited by Susan Schneider, 277-284. Malden: Wiley-Blackwell, 2009.

Boutilier, Craig, Richard Dearden, and Moises Goldszmidt. 1995. “Exploiting Structure in Policy Construction.” In *IJCAI* 14, pp. 1104-1113.

Boutilier, Craig, Thomas Dean, and Steve Hanks. 1999. “Decision-Theoretic Planning: Structural Assumptions and Computational Leverage.” *Journal of Artificial Intelligence Research* 11, (1): 1-94.

Boyd, Richard. 1988. “How to Be a Moral Realist.” In *Essays on Moral Realism*, edited by Geoffrey Sayre-McCord, 181-228. Ithaca: Cornell University Press.

Brandom, Robert. 1994. *Making It Explicit*. Cambridge: Harvard University Press.

Brandt, Richard. 1979. *A Theory of the Good and the Right*. Oxford: Clarendon Press.

Brink, David. 2001. “Realism, Naturalism, and Moral Semantics.” *Social Philosophy and Policy* 18, 154–176.

Broome, John. 2000. “Normative Requirements.” In *Normativity*, edited by Jonathan Dancy, 78-99. Oxford: Blackwell Publishers.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4, (1): 73-122.

———. 1986. “Intellectual Norms and Foundations of Mind.” *The Journal of Philosophy* 83, (12): 697-720.

———. 1990. “Frege on Sense and Linguistic Meaning.” In *The Analytic Tradition*, edited by David Bell and Neil Cooper. Oxford: Blackwell.

Cabanac, Michel. 1996. “On the Origin of Consciousness, a Postulate and its Corollary.” *Neuroscience & Biobehavioral Reviews* 20, (1): 33-40.

Cabanac, Michel, Arnaud J. Cabanac, and André Parent. 2009. “The Emergence of Consciousness in Phylogeny.” *Behavioural Brain Research* 198, (2): 267-272.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

———. 2012. *Constructing the World*. Oxford: Oxford University Press.

Chalmers, David, and Frank Jackson. 2001. "Conceptual Analysis and Reductive Explanation." *The Philosophical Review* 110, (3): 315-360.

Chang, Ruth. 2013. "Grounding Practical Normativity: Going Hybrid." *Philosophical Studies* 164, (1): 163-187.

Chomsky, Noam. 1959. "A Review of B.F. Skinner's Verbal Behavior." *Language* 35, (1): 26-58.

Cuneo, Terence, and Russ Shafer-Landau. 2014. "The Moral Fixed Points: New Directions For Moral Nonnaturalism." *Philosophical Studies* 171, (3): 399-443.

Curb Your Enthusiasm. February 8, 2004. "The Car Pool Lane". Season 4, episode 6. Directed by Robert B. Weide. Written by Larry David. HBO.

Cushman, Fiery. 2015. "From Moral Concern to Moral Constraint." *Current Opinion in Behavioral Sciences* 3, 58-62.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason, And The Human Brain*. New York: Penguin.

Dancy, Jonathan, editor. 2000. *Normativity*. Oxford: Blackwell Publishers.

———. 2006. "Nonnaturalism." In *The Oxford Handbook of Ethical Theory*, edited by David Copp, 122-145. Oxford: Oxford University Press.

Darwall, Stephen. 1983. *Impartial Reason*. Ithaca: Cornell University Press.

Degris, Thomas, and Olivier Sigaud. 2010. "Factored Markov Decision Processes." In *Markov Decision Processes in Artificial Intelligence*, edited by Olivier Sigaud and Olivier Buffet, 99-126. Hoboken, New Jersey: Jon Wiley & Sons.

Dickinson, Anthony, and Bernard Balleine. 1994. "Motivational Control of Goal-Directed Action." *Animal Learning and Behavior* 22, (1): 1-18.

Dickinson, Anthony, and Deborah J. Charnock. 1985. "Contingency Effects with Maintained Instrumental Reinforcement." *The Quarterly Journal of Experimental Psychology* 37, (4): 397-416.

Dickinson, Anthony, and C. W. Mulatero. 1989. "Reinforcer Specificity of the Suppression of Instrumental Performance on a Non-Contingent Schedule." *Behavioural Processes* 19, (1): 167-180.

Dreier, James. 1990. "Internalism and Speaker Relativism." *Ethics* 101, (1): 6-26.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.

Domjan, Michael. 2010. *The Principles of Learning and Behavior*. Belmont, CA: Wadsworth, Cengage Learning.

Donahoe, John W., Jose E. Burgos, and David C. Palmer. 1993. "A Selectionist Approach to Reinforcement." Journal of the Experimental Analysis of Behavior 60, (1): 17-40.

Dworkin, Ronald. 2011. *Justice for Hedgehogs*. Cambridge: Harvard University Press.

Egan, Andy. 2006. "Secondary Qualities and Self-Location." *Philosophy and Phenomenological Research* 72, (1): 97-119.

———. 2012. "Relativist Dispositional Theories of Value." *The Southern Journal of Philosophy* 50, (4): 557-582.

Eliasmith, Chris. 2000. *How Neurons Mean: A Neurocomputational Theory of Representational Content*. PhD thesis, Washington University in St. Louis, Department of Philosophy, St. Louis, MO.

———. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Enoch, David. 2005. "Why Idealize?" *Ethics* 115, (4): 759-787.

———. 2010. "How Objectivity Matters." *Oxford Studies in Metaethics* 5, 111-152.

———. 2011a. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press, 2011.

———. 2011b. "On Mark Schroeder's Hypotheticalism: A Critical Notice of *Slaves of the Passions*." *Philosophical Review* 120, (3): 423-446.

Fadiman, Clifton, general editor. 1985. *The Little, Brown Book of Anecdotes*. Boston: Little, Brown and Company.

Falk, W. D. 1963. "Action-Guiding Reasons." *The Journal of Philosophy* 60, (23): 702-718.

Feinberg, Todd and Jon Mallatt. 2016. *The Ancient Origins of Consciousness*. Cambridge: MIT Press.

Field, Hartry. 1977. "Logic, Meaning, and Conceptual Role." *Journal of Philosophy* 74, (7): 379-409.

———. 1998. "Mathematical Objectivity and Mathematical Objects." In *Contemporary Readings in the Foundations of Metaphysics*, edited by Stephen Laurence and Cynthia Macdonald, 387-403. Oxford: Wiley-Blackwell.

Fine, Kit. 2002. "Varieties of Necessity". In *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne, 253-281. Oxford: Oxford University Press.

Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12, (3): 317-345.

FitzPatrick, William. 2008. "Robust Ethical Realism, Non-naturalism, and Normativity." *Oxford Studies in Metaethics* 3, 159-206.

———. 2011. "Ethical Non-naturalism and Normative Properties." In *New Waves in Metaethics*, edited by Michael Brady, 7–35. Basingstoke: Palgrave.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press.

Foot, Philippa. 1972. "Morality as a System of Hypothetical Imperatives." *The Philosophical Review,* 305-316.

———. 2001. *Natural Goodness*. Oxford: Clarendon Press.

Frankfurt, Harry. 1988. "Freedom of the Will and the Concept of a Person." In *The Importance of What We Care About*, 11-25. Cambridge: Cambridge University Press.

Gallese, Vittorio. 2013. "Bodily Self, Affect, Consciousness, and the Cortex." *Neuropsychoanalysis* 15, (1): 42-45.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge: Harvard University Press.

———. 1999. "Morality as Consistency in Living: Korsgaard's Kantian Lectures." *Ethics* 110, 140-164.

———. 2003. *Thinking How to Live*. Cambridge: Harvard University Press.

Glickman, Stephen E., and Bernard B. Schiff. 1967. "A Biological Theory of Reinforcement." *Psychological Review* 74, (2): 81-109.

Gold, Joshua I., and Michael N. Shadlen. 2007. "The Neural Basis of Decision Making." *Annual Review of Neuroscience* 30, 535-574.

Gottlieb, Daniel and Elizabeth Begej. 2014. "Principles of Pavlovian Conditioning: Description, Content, Function." In *The Wiley Blackwell Handbook of Operant and Classical Conditioning*, edited by Frances K. McSweeney and Eric S. Murphy, 417-451. Oxford: John Wiley & Sons.

Gray, Peter, and David F. Bjorklund. 2014. *Psychology.* 7th edition. New York: Worth Publishers.

Greenberg, Mark. 2009. "Moral Concepts and Motivation." *Philosophical Perspectives* 23, (1): 137-164.

Greenberg, Mark and Gilbert Harman. 2006. "Conceptual Role Semantics." In *The Oxford Handbook of Philosophy of Language*, edited by Ernest LePore and Barry Smith, 295-322. Oxford: Oxford University Press.

Gregory, Alex. 2009. "Slaves of the Passions? On Schroeder's New Humeanism." *Ratio* 22, (2): 250-257.

Guestrin, Carlos, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2003. "Efficient Solution Algorithms for Factored MDPs." *Journal of Artificial Intelligence Research* 19, 399-468.

Haidt, Jonathan. 2001. "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, (4): 814.

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

Hammond, Lynn J., and Michael Weinberg. 1984. "Signaling Unearned Reinforcers Removes the Suppression Produced by a Zero Correlation in an Operant Paradigm." *Animal Learning & Behavior* 12, (4): 371-377.

Harman, Gilbert. 1975. "Meaning and Semantics." In *Semantics and Philosophy*, edited by Milton Munitz and Peter Unger, 1-16. New York: New York University Press.

———. 1977. *The Nature of Morality*. New York: Oxford University Press.

———. April 18, 1986. “Moral Agent and Impartial Spectator.” The Lindley Lecture 25, The University of Kansas, Department of Philosophy. http://hdl.handle.net/1808/12400.

———. 2015. “Moral Relativism Is Moral Realism.” *Philosophical Studies* 172, (4): 855-863.

Harnad, Stevan. 1990. “The Symbol Grounding Problem.” *Physica D: Nonlinear Phenomena* 42, (1): 335-346.

Heathwood, Chris. 2015. “Irreducibly Normative Properties.” *Oxford Studies in Metaethics* 10, 216-244.

Hieronymi, Pamela. 2011. “Reasons for Action.” *Proceedings of the Aristotelian Society* 111, 407-427.

Hume, David. *A Treatise of Human Nature.* 1738. Reprinted from the original edition in three volumes and edited by L. A. Selby-Bigge. Oxford: Clarendon Press, 1960.

Jackson, Frank. 2000. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.

Johnston, Mark. 1989. “Dispositional Theories of Value.” *Proceedings of the Aristotelian Society* 63, 138-174.

Johnson, Robert. 1999. “Internal Reasons and the Conditional Fallacy.” *Philosophical Quarterly* 49, (194): 53-71.

Johnson, Robert. 2003. “Internal Reasons: Reply to Brady, Van Roojen, and Gert.” *Philosophical Quarterly* 53, (213): 573-580.

Jozefowiez, Jérémie. “Reinforcement Learning and Conditioning: An Overview.” Unpublished, 2002. Accessed June 29, 2016. https://www.researchgate.net/publication/228762452_Reinforcement_learning_and_conditioning_an_overview

Kahane, Guy. 2010. “Feeling Pain for the Very First Time: The Normative Knowledge Argument.” *Philosophy and Phenomenological Research* 80, (1): 20-49.

Kamm, Frances M. 1985. “Supererogation and Obligation”. *The Journal of Philosophy* 82, (3): 118-138.

Korsgaard, Christine. 1986. “Skepticism about Practical Reason.” *The Journal of Philosophy* 83, (1): 5-25.

———. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

———. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Kramer, Matthew. 2009. *Moral Reason as a Moral Doctrine*. Oxford: Blackwell.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language*. Cambridge: Harvard University Press.

Lamm, Claus, Jean Decety, and Tania Singer. 2011. "Meta-analytic Evidence for Common and Distinct Neural Networks Associated with Directly Experienced Pain and Empathy for Pain." *Neuroimage* 54, (3): 2492-2502.

Lazareva, O. F. and E. A. Wasserman. 2008. "Categories and Concepts in Animals." In *Learning and Memory: A Comprehensive Reference Volume I*, edited by Randolf Menzel, 198-226. Oxford: Academic Press.

Lieberman, David. 2012. *Human Learning and Memory*. Cambridge: Cambridge University Press.

Loeb, Don. 1995. "Full-Information Theories of Individual Good." *Social Theory and Practice* 21, (1): 1-30.

———. 2005. "Moral Explanations of Moral Beliefs." *Philosophy and Phenomenological Research* 70, (1): 193-208.

Markovits, Julia. 2011. "Why be an Internalist about Reasons?" *Oxford Studies in Metaethics* 6, 255-279.

———. 2014. *Moral Reason*. New York: Oxford University Press.

McGinn, Colin. 1997. *Ethics, Evil, and Fiction*. New York: Oxford University Press.

McPherson, Tristram. 2012. "Ethical Non-naturalism and the Metaphysics of Supervenience." *Oxford Studies in Metaethics* 7, 205-234.

McSweeney, Frances, and Eric Murphy, editors. 2014. *The Wiley Blackwell Handbook of Operant and Classical Conditioning*. Oxford: John Wiley & Sons.

Menzel, Randolf, editor. 2008. *Learning and Memory: A Comprehensive Reference Volume I*. Oxford: Academic Press.

Michael, Jack. 1975. “Positive and Negative Reinforcement, a Distinction that is No Longer Necessary; or a Better Way to Talk about Bad Things.” *Behaviorism* 3, (1): 33-44.

Mill, John Stuart. 1998. *Utilitarianism*. 1863. Edited by Roger Crisp. Oxford: Oxford University Press.

Miller, Ryan and Fiery Cushman. 2013. “Aversive for Me, Wrong for You: First-person Behavioral Aversions Underlie the Moral Condemnation of Harm.” *Social and Personality Psychology Compass* 7, (10): 707-718.

Moore, G. E. Principia Ethica. 1903. Reprint of the first edition. Cambridge: Cambridge University Press, 1922.

Morf, Carolyn and Walter Mischel. 2012. “The Self as a Psycho-Social Dynamic Processing System Toward a Converging Science of Selfhood.” In *Handbook of Self and Identity*, edited by Mark Leary and June Tangney, 21-49. New York: Guilford Press.

Murphy, Eric S., and Gwen J. Lupfer. 2014. “Basic Principles of Operant Conditioning.” In *The Wiley Blackwell Handbook of Operant and Classical Conditioning*, edited by Frances K. McSweeney and Eric S. Murphy, 165-194.

Nagel, Thomas. 1978. *The Possibility of Altruism*. Princeton: Princeton University Press.

———. 2002. “Ethics without Biology.” In *Mortal Questions*, 142-146. Cambridge: Cambridge University Press.

———. 1986. *The View From Nowhere*. New York: Oxford University Press.

Nahin, Paul. 1998. *An Imaginary Tale: The Story of* $\sqrt{-1}$. Princeton: Princeton University Press.

Newman, George E., Paul Bloom, and Joshua Knobe. 2013. “Value Judgments and the True Self.” *Personality and Social Psychology Bulletin,* 1-14.

Öhman, Arne and Susan Mineka. 2001. “Fears, Phobias, and Preparedness: Toward an Evolved Module of Fear and Learning.” *Psychological Review* 108, (3): 483-522.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

———. 2011. *On What Matters*. 2 volumes. New York: Oxford University Press.

Parker, Geoffrey A., and J. Maynard Smith. 1990. “Optimality Theory in Evolutionary Biology.” *Nature* 348, (6296): 27-33.

Pavlov, Ivan Petrovich, and Gleb Vasīl′evīch Anrep, translator. 1927 (2003). *Conditioned Reflexes*. Courier Corporation.

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

———. 1998. "Implicit Conceptions, Understanding and Rationality." *Philosophical Issues* 9, 43-88.

Perry, Clint J., Andrew B. Barron, and Ken Cheng. 2013. "Invertebrate Learning and Cognition: Relating Phenomena to Neural Substrate." *Cognitive Science* 4, (5): 561-582.

Pettit, Philip. 1990. "The Reality of Rule-Following." *Mind* 99, (393):1-21.

Plato. 1945. *The Republic of Plato*. Translated with an introduction and notes by Francis MacDonald Cornford. New York: Oxford University Press.

Plunkett, Carolyn. 2016. *Actions, Reasons and Self-Expression: A Defense of Subjectivist-Internalism about Reasons*. PhD thesis, CUNY, Department of Philosophy, New York City, New York. CUNY Academic Works. http://academicworks.cuny.edu/gc_etds/1269

Price, Bruce H., Kirk R. Daffner, Robert M. Stowe, and M. Marsel Mesulam. 1990. "The Comportmental Learning Disabilities of Early Frontal Lobe Damage." *Brain* 113, (5): 1383–1393.

Quinn, Warren. 1993. "Putting Rationality in its Place." In *Morality and Action*, edited by Philippa Foot, 228-255. Cambridge: Cambridge University Press.

Rachels, Stuart. 1998. "Counterexamples to the Transitivity of Better Than." *Australasian Journal of Philosophy* 76, (1): 71-83.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95, (2): 163-207.

———. 1998. "Red, Bitter, Good." In *Facts, Values, and Norms: Essays Toward a Morality of Consequence,* 131-147. Cambridge: Cambridge University Press.

———. 2003. "Aesthetic Value, Moral Value, and the Ambitions of Naturalism." In *Facts, Values, and Norms: Essays toward a Morality of Consequence*, 85-130. Cambridge: Cambridge University Press.

———. 2014a. "The Affective Dog and its Rational Tale: Intuition and Attunement." *Ethics*, 124 (4): 813-859.

———. 2014b. “Reliance, Trust, and Belief.” *Inquiry*, 57 (1):122-150.

Rast, Erich. 2012. “De Se Puzzles, the Knowledge Argument, and the Formation of Internal Knowledge.” *Analysis and Metaphysics*, (11): 106-132.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Belknap Press.

Raz, Joseph. 2000. “Explaining Normativity: On Rationality and the Justification of Reason.” In *Normativity*, edited by Jonathan Dancy, 34-59. Oxford: Blackwell Publishers.

Rescorla, Robert. 1966. “Predictability and Number of Pairings in Pavlovian Fear Conditioning.” *Psychonomic Science* 4, 383–384.

———. 1967. “Pavlovian Conditioning and Its Proper Control Procedures.” *Psychological Review*, 74, 71–80.

Rescorla, Robert, and Allan Wagner. 1972. “A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.” In *Classical Conditioning II: Current Research and Theory*, edited by Abraham Black and William Prokasy, 64-99. New York: Appleton Century Crofts.

Rolls, Edmund T. 2000. “Precis of *The Brain and Emotion*.” *Behavioral and Brain Sciences* 23, (2): 177-191.

Rosati, Connie. 1995a. “Naturalism, Normativity, and the Open Question Argument.” *Nous* 29, (1): 46-70.

———. 1995b. “Persons, Perspectives, and Full Information Accounts of the Good.” *Ethics* 105, (2): 296-325.

Robbins, Trevor W., and Barry J. Everitt. 1996. “Neurobehavioural Mechanisms of Reward and Motivation.” *Current Opinion in Neurobiology* 6, (2): 228-236.

Romo, Ranulfo, and Wolfram Schultz. 1990. “Dopamine Neurons of the Monkey Midbrain: Contingencies of Responses to Active Touch During Self-Initiated Arm Movements.” *Journal of Neurophysiology* 63, (3): 592-606.

Ross, Taylor, Middleton, Nokes. 2008. “Concept and Category learning in Humans.” In *Learning and Memory: A Comprehensive Reference Volume I*, edited by Randolf Menzel, 535-536. Oxford: Academic Press.

Scanlon, Thomas. 1998. *What We Owe to Each Other.* Harvard University Press.

———. 2014. *Being Realistic about Reasons*. Oxford: Oxford University Press.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

Schroeter, Laura and François Schroeter. 2013. "Normative Realism: Co-reference without Convergence?" *Philosophers Imprint* 13, 1-24.

Schroeter, Laura, François Schroeter, and Karen Jones. 2015. "Do Emotions Represent Values?" *Dialectica* 69, (3): 357-380.

Schultz, Wolfram. 1998b. "Predictive Reward Signal of Dopamine Neurons." *Journal of Neurophysiology* 80, (1): 1-27.

———. 2015. "Neuronal Reward and Decision Signals: From Theories to Data." *Physiological Reviews* 95, (3): 853-951.

Schultz, Wolfram, Paul Apicella, and Tomas Ljungberg. 1993. "Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task." *The Journal of Neuroscience* 13, (3): 900-913.

Schultz, Wolfram, Léon Tremblay, and Jeffrey R. Hollerman. 1998a. "Reward Prediction in Primate Basal Ganglia and Frontal Cortex." *Neuropharmacology* 37, (4): 421-429.

Schultz, Wolfram, Peter Dayan, and P. Read Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275, (5306): 1593-1599.

Schulz, Armin. 2011. "The Adaptive Importance of Cognitive Efficiency: An Alternative Theory of Why We Have Beliefs and Desires." *Biology and Philosophy* 26, 31-50.

———. 2013. "The Benefits of Rule Following: A New Account of the Evolution of Desires." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44, 595-603.

Sextus Empiricus. 1976. *Outlines of Pyrrhonism*. Translated by R. G. Bury. 4 volumes. London: William Heinemann LTD.

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

Sidgwick, Henry. *The Methods of Ethics*. 1907. Facsimilie of the seventh edition, first published 1907. London: Macmillan and Company Limited, 1962.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–243.

Singh, Satinder, Richard L. Lewis, Andrew G. Barto, and Jonathan Sorg. 2010. "Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective." *IEEE Transactions on Autonomous Mental Development* 2, (2): 70-82.

Skinner, B. F. 1938. *The Behavior of Organisms*. New York: Appleton Century Crofts.

Skorupski, John. 2010. *The Domain of Reasons*. New York: Oxford University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

———. 1995. "Internal Reasons." *Philosophy and Phenomenological Research* 55, (1): 109-131.

Smith, Michael. 1997. "In Defense of T*he Moral Problem*: A Reply to Brink, Copp, and Sayre-McCord." *Ethics* 108, 84–119.

Sobel, David.1994. "Full Information Accounts of Well-Being." *Ethics* 104, (4): 784-810.

———. 2009. "Subjectivism and Idealization." *Ethics* 119, (2): 336-352.

———. 2009. "Review of Mark Schroeder, *Slaves of the Passions*." *Notre Dame Philosophical Reviews,* April 25, 2009. http://ndpr.nd.edu/review.cfm?id=15905.

Sripada, Chandra. 2016. "Self-expression: A Deep Self Theory of Moral Responsibility." *Philosophical Studies* 173, (5): 1203-1232.

Stevenson, Charles. 1944. *Ethics and Language*. New Haven: Yale University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, (1):109-166.

———. 2009. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters." Philosophical Issues 19, (1): 273-298.

Sturgeon, Nicholas. 1988. "Moral Explanations." In *Essays on Moral Realism,* edited by Geoffrey Sayre-McCord, 229-255. Ithaca: Cornell University Press.

Suri, Roland E., and Wolfram Schultz. 1999. "A Neural Network Model with Dopamine-like Reinforcement Signal that Learns a Spatial Delayed Response Task." *Neuroscience* 91, (3): 871-890.

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

Temkin, Larry S. 1987. “Intransitivity and the Mere Addition Paradox.” *Philosophy and Public Affairs* 16, (2): 138-187.

———. 2011. *Rethinking the Good*. Oxford: Oxford University Press.

Thomson, Judith. 2008. *Normativity*. Chicago: Open Court.

Thorndike, Edward. 1911. *Animal Intelligence: An Experimental Study of the Associative Processes in Animals*. New York: The Macmillan Company.

Velleman, J. David. 1988. “Brandt’s Definition of ‘Good’.” *The Philosophical Review* 97 (3): 353-371.

———. 1992. “The Guise of the Good.” *Nous* 26, (1): 3-26.

Waelti, Pascale, Anthony Dickinson, and Wolfram Schultz. 2001. “Dopamine Responses Comply with Basic Assumptions of Formal Learning Theory.” *Nature* 412, (6842): 43-48.

Way, Jonathan, and Daniel Whiting. 2016. “Reasons and Guidance (or, Surprise Parties and Ice-Cream).” *Analytic Philosophy* (forthcoming)

Wedgwood, Ralph. 2007. *The Nature of Normativity*. Oxford: Oxford University Press.

———. 2013. “Taking Morality Seriously: A Defense of Robust Realism. By David Enoch.” *The Philosophical Quarterly* 63, no. 251 (2013): 389-393.

Weiss, Stanley. 2014. “Instrumental and Classical Conditioning: Intersections, Interactions, and Stimulus Control.” In *The Wiley Blackwell Handbook of Operant and Classical Conditioning*, edited by Frances K. McSweeney and Eric S. Murphy, 417-451. Oxford: John Wiley & Sons.

Williams, Bernard. 1981. “Internal and External Reasons.” In *Moral Luck: Philosophical Papers, 1973-1980,* 101-113. Edited by Bernard Williams. Cambridge: Cambridge University Press. Originally published in Ross Harrison, editor, *Rational Action* (Cambridge: Cambridge University Press, 1980).