THE RATIONAL UNITY OF THE SELF

by

Graham Hubbs

BA, Washington University in St. Louis, 1999, *Summa Cum Laude*

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Graham Hubbs

It was defended on

October 9, 2007

and approved by

Kieran Setiya, Assistant Professor of Philosophy

Peter Machamer, Professor of the History and Philosophy of Science

Sebastian Rödl, Professor of Philosophy [Ordinarius], University of Basel

Dissertation Advisor: John McDowell, University Professor of Philosophy

THE RATIONAL UNITY OF THE SELF

Graham Hubbs, PhD

University of Pittsburgh, 2008

The topic of my dissertation is selfhood. I aim to explain what a self is such that it can sometimes succeed and other times fail at thinking and acting autonomously. I open by considering a failure of autonomy to which I return throughout the dissertation. The failure is that of self-deception. I show that in common cases of self-deception the self-deceived individual fails, due to a motive on his part, to be able to explain the cause of some belief or action of his. There are several philosophical projects that arise when one reflects on this failure. They are presented by the following questions: what are our minds like, such that this failure is possible? For what should we criticize the self-deceived individual, given that he has a *motivated* lack of self-knowledge but does not know he is so motivated? Is the self-deceived individual epistemically criticizable for lacking explanatory *self*-knowledge in a way that he is not criticizable for lacking knowledge that would help him explain another's thoughts and actions? By answering these questions I provide an account of the rational unity that goes missing in self-deception and in the related phenomenon of epistemic akrasia. This unity can—and I argue, should—be present in bodily action as well. When a person acts without this unity, he acts in a weak-willed, akratic way. I provide an account of this disunity, which, when added to my account of the disunity of self-deception, reveals the rational unity of an autonomous agent, the rational unity of the self.

TABLE OF CONTENTS

1.0 INTRODUCTION

The topic of this dissertation is selfhood. We speak of the self as something that can be known, deceived, unified, and divided. Often, claims about the self are at once descriptive and normative: for example, when we say that someone is self-deceived, we often mean that the person is in a particular sort of psychological condition that is bad for him to be in. When we at once describe and evaluate a person in this way, what we are responding to, I argue, is the presence or absence of the rational unity that is characteristic of autonomous thought and action. To defend this thesis, I provide accounts of what this rational unity is, of why we should think and act with this unity, and of how we are such that we sometimes succeed but other times fail to be so unified. The result is an understanding of the self that has consequences for ethics, the philosophy of action, epistemology, and the philosophy of mind.

The phenomenon of self-deception allows one to raise questions that help in developing all of these accounts, and so I focus on self-deception throughout the dissertation. I open by characterizing some common cases of self-deception in order to bring out those features of self-deception that are relevant to my discussion. I show that in common cases of self-deception the self-deceived individual fails, due to a motive on his part, to be able to explain the cause of some belief or action of his. I review the contemporary literature on self-deception and argue that most of the accounts one finds there lack the resources to describe this failure. The review of the literature is organized by assessing how the various accounts handle an explanatory challenge posed by self-deception. That challenge is to explain how the self-deceived individual manages to satisfy an end of his given that, without lying, he denies that he is pursuing any such end.

Answers to this challenge fall into two broad classes: intentionalist and non-intentionalist. An intentionalist answer makes sense of the fact that an end is pursued by saying it is intentionally pursued. This approach is rejected due to its inability to describe the epistemology of the self-deceived individual. The non-intentionalist succeeds where the intentionalist fails, but the non-intentionalist faces difficulty when describing what the intentionalist describes with ease, *viz.*, how the self-deceived individual satisfies an end of his. I argue that since the intentionalist cannot be right, some non-intentionalist account must be right, but I find all of the non-intentionalist accounts that I review to be wanting.

I close the first substantive chapter in section 2.6 by listing some of the philosophical projects that arise when one reflects on the self-deceived individual's self-explanatory failure. One of these is the project pursued by intentionalists and non-intentionalists alike, which is simply to describe how our minds are such that this failure of self-deception is possible. I take up this explanatory project in the second substantive chapter, chapter 3.0. I develop my non-intentionalist account through a discussion of mental activity. It is a general fact of activity that its purpose may be to bring about affective satisfaction. This is obvious in the case of bodily activity: we often perform bodily activities in order to bring about the affective satisfaction of bodily pleasure. While there are clear differences between mental activity and bodily activity, mental activity is also subject to the causal influence of affective satisfaction. Sometimes we form and maintain beliefs not because they are warranted but because they are affectively satisfying to hold. This may happen either with our awareness, in which case we are epistemically akratic, or without it, in which case we are self-deceived. I argue that the awareness that is present in epistemic akrasia but absent in self-deception results from the fact that in self-deception, but not in epistemic akrasia, the person is confused about the

cause of her condition. The self-deceived individual's epistemic failure is made possible, then, by the mind's susceptibility both to the causal influence of affective satisfaction and to being confused about this influence when it is so influenced.

It is important to explain self-deception and epistemic akrasia in terms of each other, for doing so allows one to see how a person in either condition fails to exhibit the virtue of epistemic courage. In either case, the person should believe something that she does not, and in either case, the tendency of the mind to form affectively satisfying thoughts causes the condition. By contrast, when a person is epistemically courageous, her mind is not susceptible to the causal influence of affective satisfaction. When a person is epistemically courageous, her mind's activity aims at nothing but the truth. In talking of a failure of virtue on the part of the self-deceived and epistemically akratic individuals, I address one normative project that the self-explanatory failure of self-deception raises. That project is to say for what we should hold the self-deceived individual accountable, given that she has a *motivated* lack of self-knowledge. As I have just indicated, I think we should hold the self-deceived individual accountable for a failure of virtue—I argue for this conclusion at the end of the second substantive chapter in section 3.6.

This is not the only normative project raised by the self-deceived individual's self-explanatory failure. The self-deceived individual seems to be criticizable for some uniquely first-personal epistemic wrongdoing because of his particular lack of *self*-knowledge. His failure thus invites us to give an account of the way in which self-knowledge is epistemically unique in order to explain the uniquely first-personal character of his failure. I take up this project in the third substantive chapter, chapter 4.0, by arguing against an imagined sceptic who does not believe that there is anything epistemically unique about self-knowledge. His scepticism is limited to the denial of the

need for a special epistemological story to be told about self-knowledge. I review a variety of contemporary accounts of self-knowledge and find that they give this sceptic no reason to give up this limited scepticism. I show that he cannot maintain this limited scepticism, however, because he cannot without becoming a more radical sceptic explain what it is for a person to understand her own reasons. This shows that our uniquely first-personal capacity to understand our own thoughts and actions is the source of the epistemic uniqueness of self-knowledge. It is thus because the self-deceived individual fails to understand his own reasons that his epistemic wrongdoing is uniquely first-personal.

Even when a person understands himself, however, he may act in a way that can be rightly considered self-divisive: this is what happens when a person performs a weak-willed action. The unity that fails to obtain in this case is the topic of the final chapter, chapter 5.0. The chapter begins with a critique of hierarchical theories of autonomy. These theories count an action as autonomous just in case the acting person "owns" the desire that causes her action, and they cash out this metaphor of ownership in terms of a hierarchical motivational structure within the person. I show that we should explain autonomy not in terms of the ownership of desires but instead in terms of the ownership of actions themselves. To own an action is to hold a causally efficacious attitude towards the action that evaluates it as no worse than any relevant alternative. When this attitude is present with an action, it provides a unity of the person with her action (and, in turn, with the desire that causes her action) that is missing when the person acts in a weak-willed manner. I make sense of the way in which an evaluative attitude may be causally efficacious through some remarks on G. E. M. Anscombe's account of practical rationality.

I close the chapter and the dissertation with some broad remarks on the nature of autonomy. Building on the work of the whole dissertation, I consider several different epistemic and practical conditions a person might be in and say whether or not we should count them as autonomous. I work my way down to what I consider the worst case, in which a self-deceived individual does something that is out of step with his own sense of what is reasonable because he is self-deceived. The point of the dissertation, however, is not reveal our capacity for rational turpitude. Rather, the goal is probe the ways in which we may be rationally weak in order to discover who we are when we are at our best.

2.0 SELF-DECEPTION AND SELF-KNOWLEDGE

Many recent discussions of self-knowledge in analytic philosophy have taken as their starting point an examination of the epistemic relationship between an individual and her most pedestrian beliefs. These discussions of self-knowledge have been shaped by the apparent challenges posed for the possibility of a person having any unique sort of knowledge of her own mind by otherwise laudable theories of mental content. Those who write positively on the matter aim to show that neither "The Private Language Argument," nor Davidsonian interpretivism, nor anti-individualistic accounts of mental content require that one understand the privilege and authority of self-knowledge in terms of some mythic and disreputable Cartesianism.[1]

Something striking about these authors is the uniform tendency in their essays to bracket failures of self-knowledge, such as self-deception, as beyond the scope of their discussions. Crispin Wright, for example, opens the first section of his "The Problem of Self-Knowledge (I)" as follows: "People can be variously deluded about themselves, self-deceived about their motives, for instance, or deluded about their strengths of character and frailties. But it is nonetheless a truism that for the most part we know ourselves best—better than we know others and better than they know us."[2] Here we see Wright conceding that cases of self-deception are exceptions to the truism that we know ourselves best; he makes this concession in order to set such cases aside before discussing what he takes to be the problem of self-knowledge. Given the worries that shape the

[1] In respective order, I have in mind here Crispin Wright's response to his own reading of Wittgenstein in Wright 2001b, Donald Davidson's response to his own theory of interpretation in Davidson 2001c, and Tyler Burge's response to his own anti-individualist theory of content individuation in Burge 1988. Such topics dominate the essays in recent anthologies on self-knowledge—see, for example Ludlow and Martin 1998 and Smith, Wright, and MacDonald 1998. There are still other examples.
[2] Wright 2001b, p. 320.

positive projects of writers such as Wright, this tendency to bracket issues like self-deception is understandable. Inasmuch as these projects are shaped by their particular sorts of worries, however, little that has been written by these authors is of much use for the person who wonders about the problems phenomena like self-deception raise about knowledge of one's own mind.[3]

Perhaps surprisingly, the vast majority of philosophers writing on self-deception in the same period fail to characterize self-deception as a failure of self-knowledge. The tendency, rather, is to understand self-deception as a mysterious failure of rationality in which a person has overwhelming evidence to hold some belief yet goes against his better reason and believes just the opposite. I do not wish to dispute that when a person holds a belief in this manner it is curiously irrational, nor that in many standard cases of self-deception this correctly describes a fault of the self-deceived individual. I shall argue, however, that the lack of attention contemporary writers have paid to the failure of self-knowledge that occurs in self-deception has led them to mischaracterize the phenomenon, which, in turn, leads them to give either implausible or inadequate explanations of the cause of self-deception. These are significant explanatory shortcomings, for one must take on serious commitments both about the epistemology of self-knowledge and about the nature of the mind in order to account for self-deception. I will argue that a specific epistemic error in common cases of self-deception is a motivated failure of self-explanation. A central goal of this chapter is to bring out the philosophical relevance of

[3] Again, this tendency to bracket self-deception runs rampant in the literature on self-knowledge. Burge 1998a discusses "mistakes of haste, bias, and self-deception" in a footnote (cf. fn. 8, p. 251) and again briefly in a paragraph towards the paper's end (cf. p. 258); he has little positive to say on the topic other than that "these issues are complex" (fn. 8, p. 251). Shoemaker 1988 ends by saying, "It is entirely compatible with [my position] that there are failures of rationality that manifest themselves in failures of self-knowledge. . . . Fortunately, that is a matter for another paper, which I haven't the slightest idea how to write." Donald Davidson, as we shall see, has written on self-deception, but not in the context of writing about self-knowledge; when he writes about self-knowledge, as in, *e.g.*, Davidson 2001c, he sets aside "self-deception and other anomalous or borderline phenomena" before providing his account of self-knowledge (p. 18). The extended treatment of self-deception in Bilgrami 2006 is a noteworthy exception to this general trend; I will discuss and criticize that account in section 2.5.

this fact, and I will conclude by noting some of the philosophical projects to which the fact is relevant. This will set the stage for the rest of this dissertation.

2.1 A CASE OF SELF-DECEPTION

Let us begin this examination of self-deception by considering what is perhaps the most popular example of the phenomenon in the contemporary philosophical literature: the self-deceived cuckold. This cuckold recognizes facts that, by his own rational standards, ought to lead him to conclude that his wife is unfaithful, yet despite the evidence before him, he goes on believing that his wife is faithful to him. In contemporary writing there are two broad classes of accounts for how the cuckold gets in this self-deceived state: we may label the one class *intentionalist* and the other *non-intentionalist*.[4] Intentionalists insist that for the cuckold's case to be a genuine instance of self-deception, he must have done something intentionally to get himself into his state. Non-intentionalists claim that the self-deceived individual does not arrive at his state by acting on any intention, and so they attempt to explain the appearance of intentionality in self-deception without resorting to intentional vocabulary in their characterization of the phenomenon. Perhaps the most well-discussed contemporary non-intentionalist account is that of Alfred Mele; we shall begin our discussion of this variety of account by considering Mele's view.

Mele offers the following as sufficient criteria for a person's arriving at a condition of self-deception:

1. The belief that *p* which the person *S* acquires is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of <*p*> in a motivationally biased way.
3. This biased treatment is a nondeviant cause of *S*'s acquiring the belief that *p*.

[4] I am following W. J. Talbott in drawing this distinction between the two different classes of accounts. See Talbott 1995, esp. pp. 30-33.

4. The body of data possessed by *S* at the time provides greater warrant for <~*p*> than for <*p*>.[5]

Consider how the self-deceived cuckold is characterized on this view. Satisfying the first condition, the cuckold falsely believes that his wife is faithful; satisfying the fourth condition, he has better evidence to believe that she is cheating. An intentionalist need not have any disagreement with Mele on including these conditions in a characterization of self-deception, but she will disagree on the inclusion of the second and third, formulated as such. According to Mele, the cuckold is motivationally biased in his treatment of the evidence that might confirm his belief that his wife is faithful, and this biasing is what leads him to believe that she is faithful. This explanation is non-intentionalist because it does not understand the cuckold as intentionally seeking the relevant evidence; he treats the evidence in a motivationally biased way without cognizing his motivation as part of an intention.[6]

Mele's style of explanation is not without its merits—indeed, the account to be offered here is relevantly non-intentionalist. As it stands, however, Mele's account is insufficient. Consider the following counterexample, which is from a critique of Mele by Richard Holton, one of only a handful contemporary writers on self-deception to emphasize the importance of self-knowledge in characterizing the phenomenon.[7] Imagine Catherine, who has recently been applying without any luck for a new job.

---

[5] See Mele 1997.

[6] Non-intentionalists come in many stripes. Mele describes the cause of self-deception in a decision-theoretic manner; see, *e.g.*, Mele 1987, 1997, and 2001. Others appeal to anxiety as the source of the self-deception; see, *e.g.*, Johnston 1988 and Barnes 1997. Others appeal more generally to emotions; see, *e.g.*, Lazar 1999. Still others appeal to the application of sub-personal stereotypes; see, *e.g.*, Patten 2003. Still others require that self-deception be understood within the framework of a "dynamic unconscious" as posited by psychoanalysis; see, *e.g.*, Lockie 2003. Undoubtedly there are other alternatives.

[7] See Holton 2000. Barnes 1997 also recognizes the necessary role of a failure of self-knowledge in self-deception; this position will be discussed at greater length in section 2.4. Patten 2003 is also worth mentioning in this regard, as it attempts explicitly to extend Holton's project. Noordhof 2003 is similarly worth mentioning; it makes a point that is very similar to Holton's in terms of "consciousness," claiming that a self-deceived individual cannot at once be self-deceived and "conscious" of his self-deceived condition. Bilgrami 2006 discusses self-deception at length over the course of his account of self-knowledge, but, as we shall soon see, he mischaracterizes self-deception.

Catherine has developed a pattern of fantasizing excitedly about how much better than her current job her potential jobs will be, only to be disappointed when she is not hired for the new jobs. She determines that such fantasizing is a problem for her, so when she sets out to apply for her next job, she decides to spend as little time as possible ruminating about it, lest she end up disappointed again. As it turns out, the latest job for which she has applied is much worse than her current job, and she would recognize this if she reflected just a bit on the job for which she is applying. Catherine satisfies Mele's four conditions. She falsely believes that the job for which she is applying would be better than her current job, and this is a belief that, by her own epistemic standards, she should not have—thus she satisfies the first and fourth conditions. Furthermore, she is nondeviantly caused to hold her belief by treating the evidence for it in a motivationally biased way—thus she satisfies the second and third conditions. "But," Holton asks, "is she self-deceived? Surely not. She is simply, and quite rationally, refusing to look at certain evidence that she has."[8]

One might object here that Catherine is not being "quite rational" in her refusal to consider the relevant evidence; on a conception of rationality according to which it is in the nature of rationality "to aim at the truth," Catherine is not being quite rational in failing to pursue the truth of the matter about the job. I shall have more to say about this sort of case in the next chapter when I discuss epistemic akrasia. Even if he is wrong to say that Catherine is "quite rational" in refusing to consider the relevant evidence, Holton is right to argue that, even though she satisfies Mele's four conditions, Catherine is not self-deceived. The reason she is not self-deceived, Holton tells us, is that Catherine *knows* that she is keeping herself ignorant. Because she knows both what she is doing and the result of her so doing (*viz.*, not reflecting about the job and thereby maintaining

[8] Holton 2000, pp. 60-61.

her unreflective judgment about the merits of the job), she is not self-deceived. If she were less aware of her strategy, Holton claims, or less aware of her status as being relatively uninformed about the job, then she would be self-deceived; the only difference between Catherine in the scenarios just mentioned and in the initial scenario is a difference of self-knowledge on her part. Holton concludes from this line of thought that if a set of conditions resembling Mele's is to be sufficient for characterizing self-deception, it will have to be formulated so that the content of the attitudes that make a person self-deceived includes "mistaken beliefs about the self. . . , mistake[s] in the kind of belief[s] that, were one to get [them] right in the right sort of way, would count as self-knowledge."[9]

I think Holton is exactly right to claim that self-deception involves mistaken beliefs about the self, and I think he is exactly right that such mistakes are about the sorts of beliefs that, if gotten right in the right sort of way, would count as self-knowledge. Unfortunately, Holton shies away from any extended discussion of the nature of the self and of self-knowledge that might illuminate his discussion of self-deception, saying that providing the needed account of self-knowledge is "no easy matter."[10] We get a hint of what the needed account looks like, however, if we further isolate the problem of self-knowledge present in self-deception. That is the task of the next section, where I show this failure of self-knowledge, at least in a wide variety of common cases of self-deception, is a motivated failure of self-explanation.

[9] Holton 2000, p. 63.
[10] Holton 2000, p. 63.

The cuckold and Catherine are similar in the following way: each believes something that, given his or her epistemic position and his or her own epistemic standards, he or she should not believe.[11] As we have just learned from Holton, though, they do not hold comparably similar beliefs about their respective epistemic wrongdoings. Whereas Catherine truly believes that her belief about her job prospects is not warranted, the cuckold falsely believes that his belief about his wife is warranted. What is true here of the cuckold is true generally of self-deceived individuals. At least tacitly, a self-deceived individual must believe that he is warranted to believe his unwarranted self-deceptive belief, for if he did not take himself to be warranted he would not be self-deceived. If a person believes something without sufficient warrant and furthermore knows that she believes what she does without warrant, she is not self-deceived; she may be irrational, but to the extent that she recognizes her irrationality, she is self-aware, not self-deceived. Such a person knows that it is wrong to explain the maintenance of her unwarranted belief by appeal to what she has most warrant to believe. The self-deceived individual lacks this self-insight; the self-deceived individual at least tacitly believes that the maintenance of the relevant belief accords with his best epistemic principles.[12]

---

[11] In the cuckold's case, this epistemic position is determined by the evidence he has concerning his wife's fidelity. I do not believe that this is always how the self-deceived individual's epistemic position is determined, however, for I do not believe that in all cases this epistemic position is an evidential matter. In cases in which an individual has a self-deceptively false belief about some motive of hers, her false belief concerns something she should know not, I believe, because of any evidence she has regarding these motives but rather because the motives are *her* motives. That a person should know her own motives, I believe, is a fact of rational agency; I discuss this in Chapter 4.0.

[12] Kent Bach would disagree with the characterization presented here of the self-deceived cuckold, for Bach would claim that while the self-deceived cuckold might entertain thoughts about his wife's fidelity, this cuckold actually believes that his wife is unfaithful. Bach 1981 characterizes self-deception as a sort of motivated avoidance of thinking certain thoughts, not as an avoidance of holding certain beliefs. As will become evident, there is much to this line of thought with which I am sympathetic, but I think it patently clear that in cases of self-deception like the cuckold's, the self-deceived individual does in fact believe what he claims to believe; the cuckold's faith in his wife is no mere thought about her fidelity but is a *bona fide* belief that she is faithful. I grant that there may be cases of self-deception of the sort Bach considers,

This being so, the self-deceived individual, so long as he is self-deceived, cannot correctly explain why he holds his unwarranted self-deceptive belief. Were he in a condition in which he could produce the correct explanation, he would have the sort of self-knowledge Catherine has concerning the maintenance of her unwarranted belief. As he is not in such a condition, the self-deceived individual fails to know why he maintains his unwarranted belief, which is at least part of the failure of self-knowledge Holton rightly notes is present in cases of self-deception. To say this is not to say that a self-deceived person must lack all knowledge of what it is that causes or maintains his self-deceived condition. It is possible for a self-deceived person to have knowledge *about* the cause of his unwarranted self-deceptive belief without knowing *that* this cause figures in the correct explanation of his self-deceived condition. For example, the cuckold might well know that believing that his wife is faithful is pleasant and that believing that she is cheating is unpleasant, without knowing that an aversion to the unpleasantness of believing her to be cheating is the cause of his belief that she is faithful. In this case, the cuckold correctly believes that he has an aversion to believing that his wife is unfaithful; what he fails to believe is that this aversion is the cause of his belief that she is faithful. This failure is a failure to articulate a cause with its effect—it is thus an explanatory failure.

On the view presented here, self-deception does not require the self-deceived individual's unwarranted self-deceptive belief to be false. For example, a concerned husband who has noticed uncharacteristic behavior in his wife that should lead him, in accordance with his own epistemic principles, to conclude that his wife is having an affair. Imagine that like the self-deceived cuckold, this husband believes that his wife is faithful but does not know that he is caused to hold this belief in order to avoid the

---

in which it is correct to characterize the person as believing something while simultaneously avoiding thinking about this belief, but such cases constitute no counterexample to my account.

unpleasantness of believing that she is cheating. Imagine further that, unlike the cuckold's belief about his wife, the concerned husband's belief about his wife turns out to be true—his wife, it turns out, is a spy operating undercover as a call girl who nevertheless is ever faithful to her husband. Despite the truth of his belief about his wife's fidelity, this concerned husband is self-deceived just as the cuckold is because he fails to know the correct causal explanation of his belief. It is the failure of self-explanation and not the representational accuracy of the unwarranted self-deceptive belief that makes both the concerned husband and the cuckold self-deceived.

Not all failures of self-explanation count as cases of self-deception, however; in particular, when such failures do not satisfy any motive on the part of the person, it is wrong to consider them as instances of self-deception. Before proceeding further, I should comment briefly on how the term 'motive' should be understood in this context. Let 'motive' pick out a category that is more general than that constituted by desires—for present purposes, all desires are motives, but not all motives are desires. When a desire causes a person to behave in a certain way, it is standardly in order to bring about a desirable state of affairs. There are motives, however, that typically cause a person to behave in ways whose end is not to bring about what the person would deem a desirable state of affairs. Jealousy is an example of such a motive, and it is one that is relevant to the topic of self-deception. Suppose, for example, that a jealous husband believes on insufficient evidence that his wife is cheating but insists that his jealousy plays no role in causing or sustaining this belief. Surely he is self-deceived, but surely the end that is satisfied—putative confirmation of his jealous suspicion—is not one that he would regard as a desirable state of affairs. The preservation of what might be described as a "core belief" is another such motive. A depressed person, for example, might go on insisting in spite of evidence to the contrary that she is not well-liked just because it is a core belief

of hers that she is intrinsically unlikable. Again, this motive does not result in a desirable state of affairs, yet it clearly causes the depressed person's self-deceived belief in a way that is neither accidental nor the result of some mere cognitive deficiency.[13]

Even on this broad understanding of motives, there are failures of self-explanation that are not the result of motives and hence not instances of self-deception. For an example of such a case, consider the following classic psychological experiment from Norman Maier.[14] Maier put subjects in a room in which there were two cords tied to the ceiling; the subjects' only task was to tie the cords together. The cords were too far apart for a subject to be able to grab one cord and walk with it in hand to the other cord, so subjects had to use various objects in the room (*e.g.*, an extension cord, a pole) to aid in completing the task. Each time a subject completed the task, Maier, who was in the room with the subjects, would ask them to find some other way to bring the cords together. There were four possible solutions; about 40% of the subjects found all four without problem, while the remaining subjects came up with all of the solutions save one, which involved the use of a weight. After these remaining subjects had struggled for ten minutes to find the last solution, Maier would saunter past one of the cords and gently brush against it, causing it to swing. Half of these remaining subjects were then able to determine the last solution, which was to tie the weight to one of the cords and make it into a pendulum; the other half never discovered the last solution.

Of interest to the present discussion is the fact that subjects who managed after seeing the cord swing to discover the pendulum solution typically did not understand their seeing the swinging cord as the source of their idea to use the weight to construct a pendulum. When asked how they came upon the solution, subjects gave answers such as,

[13] Mele calls these sorts of cases "twisted" self-deception. See Mele 2001, ch. 5.
[14] See Maier 1931. The example is cited and discussed at length in Nisbett and Wilson 1977 on pp. 240-41.

"It just dawned on me;" when explicitly asked whether or not seeing the cord swinging caused the idea of a pendulum to come to mind, a majority of the subjects claimed either that seeing the cord played no role in the formation of their plan or that, if it did play some role, it did so unconsciously.[15]

While the subjects' failure to identify the cause of their plan raises a variety of interesting questions about the nature of the mind and about self-knowledge, at present I want only to note that these subjects differ from the self-deceived cuckold in that the subjects' failure to know the relevant cause was not motivated, whereas the cuckold's failure is. The cuckold is motivated to disbelieve the correct causal account that explains his faith in his wife, for were the cuckold to admit that his belief in his wife's fidelity is grounded in nothing more reasonable than his desire that this be true, he would thereby have to admit that his belief is not rationally grounded. Perhaps it is possible that the cuckold could go on believing that his wife is faithful while admitting that the belief is unwarranted, but again, as already noted, were he to do this he would no longer count as self-deceived. By remaining self-deceived, the cuckold does not have to confront the fact that his belief is unwarranted and thereby is able to go on maintaining his self-satisfying belief. Maier's subjects are influenced by no such motives and so are not self-deceived; whatever the cause of their inability to explain correctly the cause of their pendulum-plan, it is not something that allows them to satisfy a motive on their part.[16]

---

[15] The exact results are more striking than this. Of the group who discovered the solution but only after being cued by the swinging cord, 16 reported that the idea of a pendulum occurred fully formed, whereas 7 claimed that the idea came in parts. Of the former group, not a single subject claimed that the pendulum idea was caused by seeing the swinging cord, while in the latter group, all but one mentioned that the seeing the swinging cord played a role in instigating the chain of reasoning that eventually led to tying the weight to the cord.

[16] Lockie 2003 makes this point against any account that claims self-deception is an unmotivated phenomenon. Although not explicitly mentioned in Lockie's paper, Patten 2003 is a clear example of Lockie's target.

2.3 INTENTIONALIST ACCOUNTS OF SELF-DECEPTION

Almost every contemporary theory of self-deception has ignored the fact that the self-deceived individual's condition involves the motivated sort of self-explanatory failure just described.[17] Inasmuch as the failure satisfies a motive of the self-deceived individual, it may seem as if the only way to explain self-deception is to say that the self-deceived person acts on an unconscious intention, which is a distinctly intentionalist sort of explanation.[18] Let us consider this style of account, focusing on the intentionalist theory that deserves to be called the contemporary standard, that of Donald Davidson.

According to Davidson, the self-deceived cuckold acts intentionally on an unconscious desire to relieve the discomfort of the belief that his wife is cheating on him. On Davidson's account of the phenomenon, the cuckold recognizes the evidence for believing that his wife is cheating on him but finds that holding this belief would have unpleasant consequences. The cuckold desires not to experience the unpleasantness caused by the belief, and he believes that were he to acquire evidence against the belief, he could repudiate the belief and not suffer the psychic anguish it would cause. Unconsciously, the desire not to suffer anguish and the belief that finding evidence to the contrary will keep him from holding the pain-causing belief combine in the self-deceived cuckold to form his intention to seek evidence that will confirm his belief in his wife's fidelity. If successful (as he is in the story), the cuckold comes to believe that his wife is faithful, despite the evidence he recognizes as telling against this belief. Davidson says that the self-deceived cuckold is paradoxically irrational because he holds an unwarranted

[17] Barnes 1997 is a noteworthy exception; it shall be addressed at the end of section 2.4.
[18] Intentionalist accounts of self-deception can be found in Davidson 2004b, 2004c, and 2004d, in Pears 1984, in Rorty 1988, and in Talbott 1995.

belief while simultaneously recognizing that he is warranted to hold the opposite belief. Davidson explains away the apparent paradoxicality by partitioning the mind; the unwarranted belief functions in one part of the mind while the evidence that grounds the warranted belief functions *qua* such evidence in another part. Thus separated, the evidence and the unwarranted belief can each go on existing despite the rational incompatibility of the one with the other.[19]

Davidson insists that 'part' is only to be understood functionally here; in doing so, he rejects the model of the mind one might call "psychoanalytic" or "depth psychological."[20] Such a model understands the mind as constituted by a group of dynamically interrelated parts, each with its own motivations and its own strategies and means through which these motivations are achieved. According to this model, in self-deception one of the parts achieves its end by keeping the conscious subject unaware of its activity, and so the activity of the relevant part is described as "unconscious" or "subconscious." Now note that for a "part" of the mind to be able to do what it does on this model, that part must have the motivations and strategies already mentioned as well as a capacity to represent the contents of the conscious agent's mind and, in order to dupe the conscious agent, a capacity for instrumental reasoning. Such "parts" are thus no mere parts of a mind; they are minds unto themselves. On this conception of human psychology, which Mark Johnston has derisively labeled 'homuncularist,' the single person is constituted by a plurality of minds, many of which are at war with one another without the person's knowledge.[21] To account for the self-deceived individual's unconscious intention by saying it belongs to some such part of the mind about which he

[19] For Davidson's account of self-deception, see Davidson 2004b, 2004c, and 2004d.
[20] Such a model is endorsed in Lockie 2003.
[21] See Johnston 1988, esp. pp. 63-67, for the introduction of this term.

lacks knowledge is to make self-deception practically indistinguishable from demonic possession. Davidson wants no truck with such an account.

Davidson believes he avoids any charge of homuncularism by claiming the mind is divided functionally and not "psychoanalytically" as just described. Even so, Davidson's partitioning still leaves intact the curious idea that in self-deception a person intentionally executes an act that she disbelieves she performs; this leaves Davidson with a problem that he cannot get around merely by labeling the partitioning he describes as "functional." This problem is not peculiar to Davidson but affects any theory that understands self-deception as the result of some process intentionally executed by the self-deceived person.[22] The heart of the problem is not that the self-deceived person is unaware of the *process* that generates the putative intention; we might concede to the intentionalist for the sake of argument that intentions can be formed by unconscious acts of practical deliberation, and that just such an unconscious process is that which generates the self-deceived person's intention to find evidence for her preferred belief. Even so, we can still demand that the intentionalist explain how, once formed, such an intention could be acted upon by the self-deceived individual, who disbelieves she acts on the relevant intention.

The only option the intentionalist appears to have here is to say that the relevant intention is "unconscious," but saying this raises its own problems. The intentionalist will have to provide an account of what an unconscious intention is if saying that the intention is unconscious is to solve his explanatory challenge rather than just relabel it. He will have simply relabeled the problem if the only difference between conscious and unconscious intentions is that the person has knowledge of the former but lacks

[22] This holds even for accounts of self-deception that describe the phenomenon as the result of the intentional activity of a wholly non-partitioned mind. For example, Talbot 1995 is liable to this criticism; see Barnes 1997, pp. 88-97 for an extended discussion along these lines of Talbott's failure.

knowledge of the latter. If this is the only difference between conscious and unconscious intentions, then describing an intention as "unconscious" does not help us understand how it is a *bona fide* intention, given that it can be acted on yet denied to exist by the person who acts on it. If the intentionalist responds here that this possibility of denial does not require special explanation because self-knowledge is simply not necessary for intentional action, he will owe us an explanation of why so frequently we do know what we are doing when we intentionally act and why so many thinkers on the topic have taken self-knowledge to be a standard, if not necessary, condition on intentional action. While this project seems hopeless, all we need to see here is that the position the project would defend is intuitively implausible. In order to hold an intentionalist account that is adequate for explaining cases like that of the self-deceived cuckold, then, one must also defend a strong and intuitively implausible position regarding the putative non-necessity of self-knowledge for intentional action.[23]

2.4 NON-INTENTIONALIST ACCOUNTS OF SELF-DECEPTION

Non-intentionalist accounts do not face this sort of problem precisely because they do not describe self-deception as the result of intentional action. The challenge for non-intentionalists is to make sense of the way in which self-deception manages to serve ends of the self-deceived individual despite that individual's lack of self-knowledge. One strategy the non-intentionalist might pursue is to reject my claim that all self-deception is motivated and to characterize the ends of the self-deceived individual in terms of something other than a motive. The clear way—perhaps the only way—to pursue this project is to describe self-deception mechanistically, as a condition that is the product of

[23] Lazar 1999 argues along these lines against intentionalists; see pp. 278-80.

some mechanism. A mechanistic account could characterize the end in terms of the function of the mechanism. If the function of the mechanism could be described without referring to a motive, then the non-intentionalist could explain the lack of self-knowledge feature in terms of the mechanistic production of self-deception.

An initial complaint against any motive-free mechanistic account of self-deception, of course, is that self-deception so often clearly does satisfy a motive of the self-deceived individual, so at very least the account is going to be limited. The problem is worse than this, however, for it is not clear that a motive-free mechanistic account can adequately describe *any* case of self-deception. Consider Doug Patten's account, which describes self-deception as the unmotivated result of applying what he calls "self-schemata" to self-explanations.[24] A self-schema is a collection of general beliefs a person has of her own character. Although Patten does not say this, we can imagine the function of the schema being to maintain cognitive organization; by saying this, we can characterize the function of the schema without referring to any motive of the person. Consider a version of the cuckold example in which the cuckold believes that devoted, hard-working, loyal husbands are not cheated on by their wives and that he is just such a husband. Part of his self-schema is his belief that he is a devoted, hard-working, loyal husband. This schema allows him to organize his explanations of himself in a rationally coherent manner. Patten would describe the cuckold's false belief regarding his wife's fidelity as the simple, unmotivated result of a misapplication of his self-schema.

I agree with Patten that self-deception is often the result of the misapplication of something like a self-schema, but it makes no sense to describe such schemata as functioning to produce self-deception in an unmotivated manner. A self-schema is a malleable set of beliefs: beliefs that constitute it can change, thus altering the schema.

[24] See Patten 2003.

When a person is self-deceived as Patten imagines, he irrationally maintains his self-schema instead of altering it in accordance with the facts. Patten's thought is that the irrational persistence of the schema need not be motivated, and perhaps in cases other than self-deception it is not. If the persistence is not motivated, however, then we should expect the person to correct the irrationality in accordance with the facts when the irrationality is pointed out to him. If the function of the self-schema were not at least in part to satisfy certain motives but instead were entirely to promote a motiveless end like cognitive organization, then we should expect for a person rationally to alter his schema when he is presented with its incorrectness, not for the schema to force the person into false self-explanations. When a person is self-deceived, he does not correct his schema if he is told it is wrong; he denies that he is wrong in order to preserve the schema. The flaw in Patten's account does not seem to be anything particular he says about self-schemata other than his description of them as motive-free mechanisms. The account I will offer in the next chapter does not make this mistake; it recognizes the central role motivation plays in self-deception.

My account recognizes this role without sliding back into intentionalist descriptions of self-deception. This backsliding is a mistake found in accounts similar to the present one, including Mark Johnston's tropistic account of self-deception. A tropism, according to Johnston, is a natural tendency of the mind to proceed in a certain manner without the intentional guidance of the person. In self-deception, Johnston claims, mental tropisms cause the self-deceived individual to turn away from anxiety-provoking thoughts and thus irrationally to believe what, by his own lights, he has better reason not to believe. When Johnston exemplifies his account by quoting a passage from Augustine, however, Johnston backslides into an intentionalist characterization. Augustine, speaking of his iniquity before the eyes of God, says of this iniquity, "I

refused to look at it and put it out of my memory."[25] Johnston glosses this passage in the following way: "Here the self-directed accusation of self-deception is an accusation of mental cowardice, of flight from anxiety (or angst), a failure to contain one's anxiety, a lack of courage in matters epistemic."[26] In this passage, Augustine sees himself for who he is and then turns away, so that when he performs his cowardly retreat, he knows what he is doing. Indeed, this is why Augustine says he must *forget* the truth of his iniquity. By exemplifying his point as he does, Johnston falls short of accomplishing one of his main explanatory goals, *viz*., to understand the end-serving feature of self-deception as the result of some non-intentional process of the mind.

The idea that the mind operates according to various tropisms, however, is fundamental to explaining adequately the failure of self-explanation that occurs in self-deception. As a first step towards seeing why this is the case, consider the shortcomings of Annette Barnes's account, which is the most detailed non-intentionalist account to date that acknowledges the necessary role of a failure of self-explanation in characterizing the phenomenon.[27] Like Johnston, Barnes believes that the cause of the unwarranted belief in any case of self-deception is an anxious desire on the part of the self-deceived individual. She offers the following as individually necessary and jointly sufficient conditions on self-deceiving oneself:

1. One has an anxious desire that *q* which causes one to be biased in favor of beliefs that reduce one's anxiety that not-*q*. This bias or partiality operating in one's acting or thinking or judging or perceiving etc. causes one to believe that *p*.
2. The purpose of one's believing that *p* is to reduce one's anxiety that not-*q*.
3. One is not intentionally biased or partial.

[25] Augustine, *Confessions*, VIII (7-16), as quoted by Johnston 1988, p. 85.
[26] Johnston 1988, p. 85.
[27] See Barnes 1997, esp. chs. 3, 5, 6, and 7.

4. One fails to make a high enough estimate of the causal role that one's anxious desire that *q* plays in one's acquiring the belief that *p*. One believes (wrongly, when condition 1 is met) that one's belief that *p* is justified.[28]

Barnes's fourth condition specifies the self-explanatory failure that is present whenever an individual is self-deceived due to an anxious desire. The self-deceived individual Barnes describes unwarrantedly believes that *p*, but he fails to explain correctly the cause of this belief, which is his anxious desire that *q*. He may acknowledge that he anxiously desires that *q*, but what he fails to acknowledge, as long as he is self-deceived, is the causal role this desire plays in the determination of his unwarranted belief. This failure is precisely the sort of "mistaken belief about the self" Holton identifies as essential to self-deception.

Barnes's account is laudable for including this mistaken belief about the self amongst its criteria for self-deception; it also successfully makes intelligible a way in which anxiety can lead a person non-intentionally into self-deception. The account is inadequate, however, because it fails to incorporate Johnston's general insight that the mind operates according to tropisms. This slip leads Barnes to mischaracterize the cause of the self-deceived individual's unwarranted belief. Barnes claims that this cause is an 'anxious desire,' which, according to her official account, is a desire a person has whenever the person "(1) is uncertain whether *q* and or not-*q* and (2) desires that *q*."[29] As Mele notes in a discussion of Barnes's account, this formulation involves no affective characterization, and so on it one could have an 'anxious desire' without feeling any of the typical affective responses characteristic of anxiety.[30] If this is not what Barnes wants, *i.e.*, if she wants that 'anxious desire' be understood as defining a state that involves affective anxiety, then, as Mele rightly notes, her account of self-deception is

[28] See Barnes 1997, p. 117, four footnotes omitted. Barnes distinguishes between deceiving oneself and self-deceiving oneself; the latter is a species of the former and is the present topic.
[29] Barnes 1997, p. 39.
[30] See Mele 2001, pp. 54-6. Mele claims this makes Barnes's account close to his own.

too narrow, for there are cases of self-deception that involve no feeling of anxiety on the part of the self-deceived individual. Consider the overly proud individual who self-deceptively believes he is well-regarded when he is not; there is no reason to think that a feeling of anxiety must play a causal role in his false self-conception. But neither is there reason to think that any uncertainty causes his unwarranted belief in himself; indeed, his problem may be that he is all too certain of himself. Barnes's official definition of 'anxious desire,' then, in which the relevant idea of anxiety involves no affect but rather simple uncertainty, is subject to the same complaint of over-specificity; it is also inadequate for characterizing all cases of self-deception. On either understanding of 'anxious desire,' the description of self-deception is too restrictive to account for all cases.

The way to resolve this problem, I claim, is not to account for the cause of self-deception in terms of some more general desire but rather to account for the cause strictly in terms of what Johnston calls "tropisms." Johnston's full account is itself liable to the sort of complaint just raised against Barnes, for, as already noted, Johnston accounts for the cause of self-deception narrowly in terms of anxiety. The crucial mistake that both of them commit, however, which is made by most authors on self-deception, is to characterize the unwarranted belief of the self-deceived individual as the solution to some *already existing* problem that the self-deceived individual faces. If one thinks that in all cases of self-deception there is *first* some problem of squaring a desire with evidence to the contrary that *then* leads the self-deceived individual to form a belief unknowingly in accordance with the desire, it is all but inevitable that one will characterize self-deception as, in Johnston's terms, a "flight" from the unhappy truth suggested by the evidence. One is then stuck explaining how this "flight" is either intentional yet unknown to the self-deceived individual (this is the intentionalist's problem) or non-intentional yet still a sort

of "flight" (this is the non-intentionalist's problem). The former explanatory task seems impossible, while the latter task, when successfully accomplished (as, *e.g.*, in Barnes's account), results in account of self-deception that is necessarily inadequate, since there are cases of self-deception that involve no "flight" of the sort under discussion.

In the next chapter, I will provide an account of self-deception that avoids the pitfalls of both sorts of accounts just considered. The key idea will be that self-deception is a sort of confusion: it is a confusion of what I shall call *affective satisfaction* with *rational satisfaction*. This confusion, I will argue, satisfies a motive on the part of the self-deceived individual. Before turning to that, however, I want to close this chapter by considering and critiquing one more contemporary account of self-deception. This account belongs to Akeel Bilgrami. The account attempts to address the relevance of self-deception to discussions of self-knowledge, but it fails, as I will now show, because Bilgrami fails to acknowledge that the failure of self-knowledge in self-deception is motivated. This will help us see the importance of correctly characterizing self-deception when engaging in epistemological reflection, and it will prepare us for seeing the connection between self-deception and virtue.

2.5 MOTIVES, RESPONSIBILITY, AND SELF-DECEPTION

One of the central goals of Bilgrami's book *Self-Knowledge and Resentment* is to argue for what he calls the "transparency thesis," which asserts that, given agency, if a person desires or believes that *p*, then she believes that she desires or believes that *p*.[31] The clause 'given agency' restricts this claim to apply only to desires or beliefs that P that are

[31] Bilgrami 2006. The other central goal is to argue for what he calls the "authority thesis," which asserts that if S believes that she desires or believes that *p*, then she desires or believes that *p*. It is only the transparency thesis that is presently of interest.

objects of "justifiable reactive attitudes" or that are part of a rationalization of such objects.[32] Justified reactive attitudes are evaluative judgments such as blaming, punishing, and resenting. An intentional state is an object of a justifiable reactive attitude if it is one for which its bearer is liable to evaluative judgments such as blame, punishment, or resentment. Bilgrami's thought here is that only when a belief or desire is self-known is that belief or desire, or any action that follows from that belief or desire, something for which the person is directly criticizable.

Bilgrami recognizes that this thesis seems falsified by cases of self-deception. In at least some cases of self-deception, it is at least initially plausible to think that the person has a belief or desire that, despite not knowing that he has the belief or desire, is an object of a justifiable reactive attitude. Bilgrami argues that whenever it seems right to say this, the proper target of criticism is the person's ignorance of his belief or desire and that this ignorance is not an object of a justifiable reactive attitude. If a person self-deceptively does something for which he is liable, then, according to Bilgrami, the proper target of our criticism is not the person's performance of the wrongful deed but rather his ignorance of the fact that what he has done is wrongful.

His argument for this claim is as follows. Assume that standardly a person is responsible for an action if and only if the description of what she is accountable for is a description she would recognize as one of what she is intentionally doing. Call this assumption *the self-knowledge requirement* on responsibility. While the self-knowledge requirement is a standard condition for responsibility, there are cases in which the description a person would provide of what she is intentionally doing is not a description of her wrongdoing, yet she is still responsible for what she does. Suppose that in such cases what we hold the person accountable for is not her ignorance of the truth of the

[32] Bilgrami 2006 discusses justifiable reactive attitudes at length in Chapter 2. He credits Strawson 1962 with originating the notion.

description according to which her action is a sort of wrongdoing; suppose that in such cases what we hold her accountable for is simply her action and its consequences. If we take this line, we are, according to Bilgrami, "backsliding from whatever rationale [we]. . . had for demanding *in general* such things as mens rea and the self-knowledge requirement in the first place."[33] If we backslide from this rationale, the only other option Bilgrami thinks we have is to adopt some form or another of consequentialism. This, however, is repugnant, because the consequentialist worldview is one that "[does] not have any place for respect for individuals and their autonomy."[34] If we are to maintain the self-knowledge requirement generally, then, we must understand the target of blame in the exceptional cases not to be the action itself but rather the person's ignorance that she performs a wrongful action. Therefore, Bilgrami concludes, if a person does not recognize a description of an action as one that applies to what she has done, then neither the action nor the beliefs and desires that cause it are objects of justifiable reactive attitudes.

Bilgrami does not so much as consider that the proper target of criticism in cases in which a person is ignorant of her wrongdoing might be a lack of virtue on the person's part. In the next chapter, I will argue that in the cases of self-deception that are relevant to Bilgrami's—or for that matter, anyone else's—discussion of self-knowledge, what we hold the person accountable for is a lack of virtue. Setting further talk of virtue aside for the moment, I want to argue that in accordance with his own line of thought, Bilgrami cannot legitimately draw the conclusion he does, at least not without further argument. Bilgrami maintains that the capacity to have self-knowledge is a necessary condition on being an agent. When this capacity is properly exercised through an intentional action, part of what the agent knows is the end pursued by the action, and this knowledge is

[33] Bilgrami 2006, p. 106. Bilgrami's emphasis.
[34] Bilgrami 2006, p. 108.

paradigmatically characterized by a description which specifies the action in terms of this end. In this paradigmatic case, there is a unity between an agent and the end pursued by his action that is relevant to the assessment of responsibility, which is paradigmatically constituted by the self-knowledge the agent has of the end being pursued. If, however, this unity between agent and end can be constituted without self-knowledge, it should at very least be an open question for anyone thinking along Bilgrami's lines whether or not the lack of self-knowledge in such a case exculpates the agent from being responsible for the action. If the object of the agent's ignorance includes some end of hers that is satisfied by her action, it is at very least an open question for anyone thinking along Bilgrami's lines whether the proper target of criticism is her ignorance of what she does or her performance of the action itself.

In order to clarify this point, let us consider a pair of examples, the latter of which will be returned to in the next chapter. Consider first a case of wrongdoing in which the description of what the person does in terms of the end he pursues is not a description of the wrongdoing in question. Imagine that an Arab man is the most qualified candidate for a job but that he does not get the job because an intern accidentally throws away his application. While cleaning up the boss's office, this intern notices a disheveled stack of papers on the floor. Assuming that the stack is of papers to be thrown away, the intern bins the whole stack, including the Arab man's application. A good description of the intern's action in terms of the end being pursued is something like 'cleaning up the boss's office.' This description is not one of the wrongdoing, which is something like 'throwing away the best applicant's application.' In this case, it seems clear that the proper target of criticism is the intern's ignorance of his wrongdoing. In this case, Bilgrami's thought regarding the proper target of criticism is correct.

Contrast this with a case of wrongdoing in which the description of what the person does in terms of the end he pursues is a description of the wrongdoing in question. Consider the Arab applicant again, only this time imagine that he fails to get the job because the person making the hiring decision is a self-deceived racist. Suppose that the racist is motivated primarily by his racism to reject the Arab's application but that he denies that this is so. In this case, a good description of the racist's action in terms of the end being pursued will be something like, 'keeping Arabs out of the vicinity.' A good general description of what the racist has done wrong is something like 'acting in a racist manner.' The description of what specifically he has done that is racist, however, is given by the description of his act in terms of its end, 'keeping Arabs out of the vicinity.' Being self-deceived, the racist will deny that the end specified by this description is one he pursues in rejecting the Arab man's application. The racist is thus ignorant of his wrongdoing. Given that the end in question *is* his end, though, we cannot without further argument conclude that just as in the case of the intern, the proper target of criticism here is the racist's ignorance. Indeed, if the rationale guiding our assessment of responsibility is one that emphasizes the unity of an agent with his end in determining responsibility, then we should be led in the opposite direction and blame the racist for his action, not for his ignorance of wrongdoing. By failing to attend to the way in which the racist's self-explanatory failure is motivated, then, Bilgrami fails to note the way in which self-deception is a problematic failure of self-knowledge. In Bilgrami's case, the oversight is severe, for keeps him from seeing a serious flaw in his argument for his transparency thesis.

2.6 SOME PROJECTS TO PURSUE

I have argued in this chapter that common cases of self-deception involve a motivated failure of self-explanation. I have already indicated some of the philosophical projects to which this fact is relevant (*e.g.*, Bilgrami's); I want now to say in a bit more detail which of these projects will be the topic of my focus for the remainder of this dissertation. One is simply the project of providing an account of the nature of the mind that makes clear how self-deception and its characteristic failure are possible. This descriptive project is part of the task of the next chapter. The fact that self-deception involves a *failure* of some sort is relevant to at least two normative philosophical projects, each of which can be posed in the form of a question. What has just been said about Bilgrami's failure to note that self-deception involves a motivated lack of self-knowledge prepares us for the first question, which is this: for what should we criticize the self-deceived individual, given that he has a motivated lack of self-knowledge but does not know he is so motivated? As I have already announced, I think that the answer to this question is that we should criticize him for a lack of virtue. I argue for this claim also in the next chapter. The other normative project is presented by this question: ignoring for the moment that his self-explanatory failure is motivated, is the self-deceived individual epistemically criticizable for lacking explanatory *self*-knowledge in a way that he is not criticizable for lacking knowledge that would help him explain another's thoughts and actions? The answer to this question is, yes; I discuss this in detail in chapter 4.0.

By answering these two questions I will present an account of the rational unity that goes missing in self-deception. As this account develops over the next two chapters, my focus will be primarily on the presence and absence of this unity over the course of producing, maintaining, and revising beliefs, but at times I will remark on the importance

of this unity in the production of bodily action. In the chapter 5.0, I will concern myself exclusively with the role of this unity in producing bodily action, saying how a person is disunified when he acts in a weak-willed way. Once this disunity is described, I will be in a position to say how a person is when her thought and action are rationally unified, when she is, in a word, autonomous. A discussion of autonomy will thus conclude this dissertation.

3.0 VIRTUE AND MENTAL ACTIVITY

Writers on the philosophy of mind and on the epistemology of self-knowledge have recently begun to examine the idea that there are actions of the mind that share fundamental metaphysical and epistemological properties with actions of the body.[35] In the following chapter, I argue that even if it is wrong to think that mental activity is a sort of action, this general approach can help us see that theoretical activity, like practical activity, can be performed either virtuously or unvirtuously. To show this, I will focus on the theoretical activity that results in the epistemically bad conditions of self-deception and epistemic akrasia. I will argue that the activity in question is performed with a lack of courage. While what I will say here may be relevant to virtue epistemology, it is not my goal to forward a thesis on behalf of that philosophical enterprise. My goal is to describe an aspect of the mind in a way that reveals its well-functioning to be a matter of virtue.

3.1 EXPLANATORY CHALLENGES OF SELF-DECEPTION AND EPISTEMIC AKRASIA

I shall begin by characterizing self-deception and epistemic akrasia, focusing on the various explanatory challenges posed by the phenomena. Let us start with some examples, the first an example of self-deception. Recall the racist employer from the last chapter who self-deceptively believes he is egalitarian. This racist is motivated primarily by his racism to reject the Arab job-seeker's application. He believes, however, that something other than his racism is the basis for his hiring decision, and in holding this

[35] Examples of this include Soteriou 2005, Peacocke 2006, and Soteriou and O'Brien forthcoming.

belief he is self-deceived. The racist maintains his particular self-deceptive belief about his hiring decision in accordance with his general conception of himself as egalitarian. This false self-conception is not corrected by the fact that he has acted in a racist way but rather causes him to maintain the unwarranted belief that his action is non-racist.

Like an individual who is self-deceived, an individual who is epistemically akratic maintains a belief she is not warranted to hold. Consider, for example, the parent who believes her son to be innocent of an awful crime of which he has been accused.[36] Imagine that the parent recognizes that there are strong reasons for believing in his guilt and little reason to believe in his innocence. The parent thus recognizes that were she to accord with her general epistemic standards, she would believe that her son is guilty. Because it is her son, however, adopting such a belief would be incredibly painful, and either unable or unwilling to contend with that pain, she goes on believing in his innocence. Like the racist, the parent believes something she is not warranted to believe: in this case, it is that her son is innocent. The epistemically akratic parent differs from the self-deceived racist, however, because unlike the racist she knows that her belief about her son is not in line with her best epistemic practices. If she failed to know this, she would be self-deceived, not epistemically akratic.

Each of these conditions poses its own explanatory challenge. In the case of self-deception, the challenge is to understand how someone like the racist can bring about ends such as not hiring the Arab and thinking well of himself without recognizing that he has the relevant ends. In bringing about the ends, the racist successfully satisfies some of his motives. Typically, one goes about satisfying a motive by knowingly acting to do so. The racist lacks the knowledge that is present in the typical case—he does not know which of his motives cause him to act and to think as he does. The challenge here is to

[36] This example is taken from Hookway 2001, which provides some insightful reflections on epistemic akrasia.

explain how it is that he manages to satisfy these motives, given that he does not know the role they play in causing his action and thought.

In the case of epistemic akrasia, the challenge is to explain how someone like the parent can recognize that she lacks adequate reason for believing her son to be innocent yet go on believing that he is. Typically, when we come to accept that a belief is not adequately grounded in reason, we give the belief up. Typically, this giving-up happens immediately upon accepting that the belief is not sufficiently grounded in reason—there is nothing extra a person must do after discovering a belief to be unwarranted in order to eliminate the belief.[37] The epistemically akratic parent does not change her belief about her child in this way—indeed, she does not change it at all. The challenge here is to explain how she goes on maintaining her belief after she finds it to be unwarranted.

While each of these conditions poses its own unique explanatory challenge, these two conditions are similar in certain fundamental ways. By sorting out the ways in which the conditions are similar and the ways in which they differ, we will position ourselves to understand how a single account of mental activity can answer both of the explanatory challenges just mentioned. I turn, then, to describing the similarities and differences between the two conditions, which I shall do by focusing on three features, all of which may be present in self-deception but only two of which may be present in epistemic akrasia.

[37] This point about the way in which we immediately give up beliefs, in the normal case, when we find them to be inadequately grounded in reason has been made by Burge 1998b (cf. p. 255) and Moran 2001 (cf. p. 131, where Moran discusses what he calls "practical immediacy"). I note this again in the next chapter; cf. fn. 73.

As he has just been described, the self-deceived racist exemplifies two of these features, which I shall call the *end-serving feature* and the *lack of self-knowledge feature*. By being self-deceived, the racist satisfies an interrelated pair of ends, both of which can be considered as constituting the end-serving feature. The racist employer is non-accidentally able to realize his end of keeping Arabs out of the vicinity, and he does so without recognizing that he has such an end. This lack of recognition itself serves an end of the racist, which is to maintain the satisfying conception of himself as egalitarian. When he denies that it is his racism that has led him to reject the Arab man's application, he thus non-accidentally manages to realize both his end of keeping Arabs out of the vicinity and his end of thinking well of himself. These together are the ends that are satisfied by his being self-deceived.

In cases like the racist's, the lack of self-knowledge feature is constituted by a pair of false self-beliefs corresponding to the pair of ends just mentioned. The racist believes that the cause of his not hiring the Arab applicant is something that it is not. The cause of his not hiring the Arab applicant is his racist motivation to keep Arabs out of the vicinity, but he does not believe that this is the cause. Call whatever belief he holds about the cause of his not hiring the Arab applicant *B1*. *B1* is not the only false belief he holds in being self-deceived; in line with what we said about the cuckold in the last chapter, the racist also at least tacitly holds that *B1* is warranted, which it is not. He at least tacitly believes that he holds *B1* because *B1* is true, but in fact he holds *B1* in order to realize the end of maintaining the favorable conception of himself as egalitarian. Were he to recognize that he holds *B1* only in order to realize the end of thinking well of himself, he would thereby have to recognize his lack of warrant for holding *B1*. As long

as he is self-deceived, he will not recognize this lack of warrant. If his end of taking himself to be egalitarian is to be realized, then, he must believe whatever is needed to explain his behavior as non-racist and at least tacitly believe that these beliefs are warranted. If either of these beliefs were true, it would constitute self-knowledge; as neither are true, both are failures of self-knowledge.

In cases of self-deception like the racist's, then, there are a pair of false self-beliefs the person holds and a pair of ends that these false beliefs allow him to realize. I will not attempt to argue here that either of these pairs is necessary for self-deception, but it should be clear that despite their apparent complexity, they are common features of self-deception. The higher-order belief of the lack of self-knowledge feature is simply the belief, which perhaps is usually tacit, that the first-order self-deceived belief is warranted. If the self-deceived individual did not at least tacitly hold this higher-order belief, then he would freely admit—as the epistemically akratic individual does but the self-deceived individual does not—that he should not hold his unwarranted belief. Similarly, at least in cases where bodily action is needed to realize an end, it is common that the individual's ends include both the thing that he actively pursues and, we might figuratively say, keeping himself in the dark about what he pursues. When self-deception manifests itself exclusively in belief, without action, this structure may be lacking.[38] When, however, the self-deceived individual's ends include not just believing something but also realizing some result in the external world, he will commonly have this structured pair of ends.

Epistemic akrasia involves the first but not the second of these features. As we have already noted, the epistemically akratic parent knows that her belief in her son's

[38] An example of this actionless self-deception is the self-deceived cuckold whose only relevant end is to believe that his wife is faithful. One can imagine a case in which this cuckold completely realizes this self-deceptive end solely by maintaining his unwarranted belief in her fidelity. In such a case there would be no self-deceptive ends that require bodily action for realization.

innocence is not warranted. While there may be cases of epistemic akrasia in which the person does not know why she maintains her unwarranted belief but knows only that it is unwarranted, we can imagine that this self-knowledge is present in the case of the parent: she knows that it would be emotionally devastating to believe her son to be guilty and that this is why she goes on believing he is innocent. Unlike the self-deceived racist, the akratic parent thus knows both that her belief is unwarranted and why in spite of this lack of warrant she maintains the belief. What she has in common with the self-deceived racist, however, is an end that is realized by maintaining an unwarranted belief. In her case, the end is the avoidance of the emotional devastation that would result from accepting her son's guilt. Like the self-deceived individual, the epistemic akratic realizes some end by maintaining an unwarranted belief, but unlike the self-deceived individual, the epistemically akratic individual does not have the further end of keeping herself in the dark about this.

The third feature, which may be present in either self-deception or epistemic akrasia, is the role that habit can play in causing either condition. Call this feature the *role of habit feature*, and focus here just on self-deception. It is easy to ignore this role if one thinks of self-deception narrowly as a reaction to some problem posed by the truth. In the last chapter, I noted that it is quite common to conceive of self-deception narrowly as reactive; I made the remark while discussing Johnston and Barnes, with whose accounts of self-deception I have some sympathy. If one thinks of self-deception narrowly as reactive, one must presuppose that the problem posed by the truth is first somehow recognized by the self-deceived individual and then, once recognized, responded to self-deceptively. It is wrong, however, to characterize all self-deception as a response to a recognized problem. Consider once more the self-deceived racist. He has the standing false self-belief that he is not a racist, so he has the habit of explaining his

racist thoughts and actions in ways that do not involve racist characterizations. To acknowledge the truth that he is a racist would indeed pose a problem for him. But the explanations that he gives that allow him to keep from acknowledging the problematic truth need not be the result of responding to the problem of the truth. Rather, they can simply be the result of his habit of explaining his actions in non-racist ways.

Epistemic akrasia may also result from habit. If a parent is in the habit of always thinking best of her child, she may habitually form and maintain beliefs about her child that she knows are not warranted. By considering this last feature as it pertains to either condition, we position ourselves to see that there *are* habits of the mind, which must be established if we are to understand self-deception and epistemic akrasia as the failures of virtue that they are. Towards showing that these failures are failures of virtue, let us now turn to the causal account of this pair of epistemically flawed conditions. This account describes the conditions as the result of mental activity; we shall approach the account through some general remarks on the recently popular topic of mental action.

3.3 MENTAL ACTIVITY

The recent literature on mental action takes as its starting point the simple thought that many sorts of mental episodes are actions. If this thought is correct, then bodily actions are just a species of the more general category of actions, and bodily actions and mental actions have whatever metaphysical and epistemological characteristics actions in general have. A common goal of writers on this topic is to reject a purely passive conception of the mind, which is committed to understanding the epistemology of self-knowledge as a sort of inner observation. By establishing that there are episodes of the mind that are actions and that we know those episodes in the same non-observational way that we

know our bodily actions, these writers hope to show that self-knowledge is not an observational sort of knowledge.[39]

While there are many who agree that any purely passive conception of the mind must be wrong and that the mind is in some sense active, it is a matter of controversy whether this activity is best described in terms of action. One form this controversy can take is as a worry about believing at will. Those who advocate the existence of mental actions tend to include judgments among the sorts of mental episodes that are mental actions. Actions typically, if not necessarily, are the sorts of things that are done intentionally and that can be done willfully. If theoretical judgments, which result in beliefs, are actions, then it seems to follow that theoretical judgments can be made willfully and that, thus, beliefs can be formed at will. But almost everyone agrees that beliefs cannot be formed at will.[40] If it is correct that beliefs cannot be formed at will, then the advocate of mental action needs to explain how we are to conceive of judgments as actions without at once offensively conceiving of beliefs as possible products of the will.[41]

At present, I shall be mute on this issue and remain neutral regarding whether the activity involved in making judgments is best understood as a sort of intentional action. While remaining neutral on this narrow issue, I want to endorse what I take to be the general insight of this literature on mental action: mental activity is indeed a sort of activity that, as such, shares metaphysical characteristics with other sorts of activity, including intentional action. Our activity, both mental and bodily, is a product of our

---

[39] This is a goal of Soteriou 2005 and Peacocke 2006. Wilson 2004 reads Moran 2001 as also pursuing this strategy. It seems to me that Moran relies upon the activity/action distinction I discuss in the following paragraphs and that he does not hold the position Wilson ascribes to him, but I shall not here defend this reading of Moran.

[40] While there is general agreement that one cannot form beliefs at will, there is little agreement as to why this is true. On this issue, see Williams 1973b, Bennett 1990, Shah and Velleman 2005, and Hieronymi 2006.

[41] This concern is raised in Soteriou 2005; see pp. 91 ff.

nature as living beings. As living beings, we have a general tendency to do what brings pleasure and satisfaction and not to do what brings displeasure and dissatisfaction. This tendency can lead us to act in a weak-willed manner, as we do when knowingly against our commitments we indulge in sexual, gastronomic, or drug-induced pleasure or avoid arduous or fearful situations. In some instances, acting on this tendency involves being overcome by an overwhelming feeling of lust or fear and thus acting out of character; in other instances, acting on this tendency involves simply acting in accordance with one's bad habits.

This tendency affects the goings-on of the mind just as it does the goings-on of the body. In the same way in which we have a tendency against putting ourselves in genuinely fearful or sad situations, we have a tendency against thinking genuinely fearful or sad thoughts.[42] This tendency can lead to self-deception. To understand how it can lead to self-deception, however, one must confront an explanatory challenge mentioned in the section 3.1, *viz.*, to explain how the self-deceived individual's mind is affected by this tendency even though he may positively deny that it is so affected. Typically, when one avoids a situation that would be unpleasant, one does so knowingly. Sometimes this knowledge is manifested by a deliberate choice; other times, it is manifested by the recognition that an emotion like fear is guiding one's action in a manner that does not involve deliberate choice. Neither sort of knowledge is present with the self-deceived individual. The racist is perhaps aware that it would be unpleasantly shameful to believe that he is a racist, but he does not know that it is this possible unpleasantness that leads him to believe that he is egalitarian. If told that it is only to avoid shame that he considers himself egalitarian, he denies the accusation. This denial is no straightforward

[42] These tendencies of thought are what Johnston 1988 calls "mental tropisms." In discussing these general tendencies away from what is fearful or sad, I do not rule out that we sometimes gravitate towards what is non-genuinely fearful or sad, like roller coasters and horror movies. Nor do I rule out that the mind sometimes fails to operate in accordance with these tendencies, as can happen when someone is depressed.

lie—he does not through his verbal act of denial misrepresent what he takes to be true. How, though, can he fail to know that it is in accordance with a general tendency towards pleasure and away from displeasure that he holds his belief that he is egalitarian?

Let us reformulate this explanatory challenge a bit more precisely. When a person is self-deceived, he holds a belief because it is affectively satisfying. A belief is affectively satisfying only if it satisfies some motive on the part of the person in question.[43] The characteristic way for a rational agent to realize affective satisfaction is to succeed in pursuing it, which characteristically requires knowing that the satisfaction is an end being pursued. When a person is self-deceived he does not know that he has the end that he manages to realize—in fact, he believes he lacks the end. How can he realize some end that he believes himself not to have?

3.4 CONFUSION AND SELF-DECEPTION

My answer is that his end is realized through a confusion on his part. The state of confusion I have in mind here is one in which a person takes one thing to be something that it is not. This sort of confusion may be between instances of a common kind (as happens when, *e.g.*, I confuse a person with her twin) or between instances of different kinds (as happens when, *e.g.*, I confuse aluminum with molybdenum). One may be confused because one lacks the ability to discriminate between two discriminable items, but one may also be confused because one has a discriminative ability yet fails to utilize it properly on some occasion. One may be confused without knowing that one is

[43] To be still more precise, note that not all motives are desires, and not all that is affectively satisfying is pleasant or desirable. For example, a parent who has lost her child to leukemia may self-deceptively believe that she is responsible for her child's death. Such a belief, though in no ordinary sense desirable, satisfies some motive of the parent (in all likelihood, the motive of finding some sense of control in the face of irreparable loss). On this example, Knight 1988, esp. pp. 182 ff.

confused, and so confusion should be distinguished from the affective condition of being perplexed. If, for example, I do not know there is anything called 'molybdenum', then I will not know when I confuse aluminum and molybdenum that I am confusing the two. Even if I know that they are two distinct sorts of metals, I may on occasion confuse the one for the other without having any idea on a given occasion that I am confusing the two.

When a person is self-deceived in the manner of the self-deceived racist, I claim, he fails to exercise the ability to distinguish between two sorts of satisfaction, one affective, the other rational, and thus confuses them. A thought may be satisfying even though the person thinking the thought knows that it would be false were it formulated as a belief. For example, I may find the thought that I will never die satisfying even though I know that any belief to that effect would be false. But a thought may be satisfying just because it is the result of good epistemic practices, which in the paradigmatic case result in knowledge and understanding. When a thought is satisfying in the former sort of way, it is affectively satisfying; when it is satisfying in the latter sort of way, it is rationally satisfying.[44]

The idea of rational satisfaction may at first sound odd, but it is found in a wide variety of everyday phenomena. Consider, for example, the philosopher who wonders how properly to explain why we cannot form beliefs at will. When a person is troubled by this topic, her perplexity about the matter is accompanied by a distinct sort of rational tension, and she seeks to resolve this tension with a solution that satisfies her. Or consider the person who loves detective stories; this person takes pleasure in the explanatory tension posed by a good mystery and seeks to resolve this tension with a

[44] In calling the one sort of satisfaction 'rational' and the other 'affective', I do not mean to imply that one cannot rationally pursue affective satisfaction. It is often practically rational to do something for the sake of affective satisfaction. The present topic is not practical rationality, however, but theoretical rationality, and it is never theoretically rational to form or to maintain a belief for the sake of affective satisfaction.

satisfying resolution to the mystery. Sometimes this person finds the conclusion of the story a satisfying resolution, other times not. Note that it is not always for the sake of pleasure that people seek to satisfy explanatory tensions; consider the person who is betrayed by her friend and who wants, among other things, an explanation that makes sense of her friend's betrayal. It is wrong to think this person aims at a sort of pleasure in understanding the cause of her friend's betrayal, yet she still wants an explanation that makes intelligible what her friend has done. In each of these cases, an explanatory tension persists in the person until a belief is found that satisfactorily resolves the tension. The satisfaction in each case is rational, for it is satisfaction at having arrived at understanding, at the truth.[45]

My thesis, again, is that the person who is self-deceived like the self-deceived racist confuses the affective satisfaction he finds in some belief or explanation with the rational satisfaction that would be present if the belief or explanation were warranted. In standard cases, like that of the racist, the confusion is made possible in part by the rationalization the self-deceived individual can provide if prompted to do so. A rationalization is so-called because it explains a belief or action in a rationally comprehensible (but false) manner. Because it is rationally comprehensible, a rationalization provides some rational satisfaction. The racist might rationalize his choice not to hire the Arab candidate by saying some other candidate shows more promise for the job; he will not rationalize the choice by saying that whales are mammals. The fact that whales are mammals provides no rationally satisfying grounds for hiring one or another candidate, but the fact, even if only putative, that one candidate shows more promise than another does provide rationally satisfying grounds for hiring that candidate. Because the rationalization provides him with some rational satisfaction, the racist

[45] Lear 1988 discusses the desire for understanding that is fulfilled by rational satisfaction; adapting a term from Melanie Klein, he calls this *epistemophilia* (cf. pp. 3-10).

confuses the affective satisfaction of his unwarranted beliefs with the rational satisfaction of believing what is warranted.

The source of the racist's confusion is his conception of himself as egalitarian. He takes it that this self-belief is satisfying because it is true, but in fact it is satisfying because it allows him to maintain a prideful, not shameful, conception of himself. Were he not confused about the sort of satisfaction he finds in this belief, he would recognize that he lacks rational grounds for maintaining the belief that he is egalitarian. This he will not do. Because he is confused about this, he is further confused about the satisfaction he finds in whatever belief he has regarding the cause of his not hiring the Arab candidate, which in section 3.2 we called *B1*. That belief is satisfying because it allows the racist to maintain his satisfying self-conception, not because it is true. As with the belief about his character, the racist takes his belief about his motive for not hiring the Arab candidate to be rationally satisfying because it is true, but he is wrong about this.[46]

3.5 THE FEATURES OF SELF-DECEPTION AND EPISTEMIC AKRASIA REVISITED

Because it describes the self-deceived individual as confused about two sorts of satisfaction, the present account is able to explain the lack of self-knowledge and end-serving features of self-deception. Because he is confused, the self-deceived individual lacks self-knowledge. The racist's confusion regarding the sort of satisfaction he takes in conceiving of himself as egalitarian explains his failure to know both generally why he considers himself not to be a racist and specifically why he does not hire the Arab

[46] The inspiration for the account presented in this section is Freud 1911. It is here that Freud makes clear how we should understand the mind as capable of conforming either with the 'pleasure principle' so as to achieve affective satisfaction or with the 'reality principle' so as to achieve rational satisfaction. I have sought in this section to explicate this insight of Freud's without endorsing any of the unacceptable pictures of the mind one finds in his writings.

candidate. He can thus satisfy his end of keeping Arabs out of the vicinity without having to endure the affective dissatisfaction of considering himself a racist. Should he ponder the possibility that he is a racist, he will confuse the affective dissatisfaction of that thought for the rational dissatisfaction of finding a belief to be lacking in warrant.

Were he not confused, the racist would know that these unwarranted beliefs are not warranted. In this, he differs from the epistemically akratic individual, who knows that her unwarranted belief is not warranted. She knows this because, unlike the self-deceived person, she is not confused about the satisfaction that causes her to hold the relevant belief. In both cases, however, the end of generating and maintaining an affectively satisfying state is realized by holding an unwarranted belief. In both cases, the tendency to avoid what is affectively unsatisfying causes an affectively satisfying yet unwarranted belief to be held.

In saying this, we have just described the cause of the epistemically akratic person's condition, which raises the explanatory challenge of epistemic akrasia mentioned in section 3.1. That challenge, it will be recalled, was to make sense of the way in which someone could know she lacks adequate reason to hold a certain belief yet go on maintaining it anyway. This situation poses an explanatory challenge, however, only if one focuses narrowly on the way in which the mind is maintained and altered in accordance with norms of rationality and thereby fails to see that the mind can be maintained and altered in other ways. When a person is fully rational, she maintains and alters her mind exclusively in accordance with norms of rationality. But precisely because there is another tendency of the mind to generate and to maintain affectively satisfying states, persons are not always fully rational. It is a worthwhile explanatory project to describe exactly how a person represents the rational commitment she recognizes but fails to live up to when she is akratic, either epistemically or practically,

but that is not a project to be pursued at present. All that is to be done here is to make sense of the possibility that knowledge of the rational commitment might not immediately lead the person's mind to be adjusted in accordance with the commitment. That possibility is explained by the simple fact that there is a tendency of mental activity, like any activity, to be directed at bringing about affective satisfaction. When a person's thinking is directed by this tendency, it should not be surprising if her knowledge of the tendency's efficacy is not adequate for counteracting the tendency.

The present account can also describe cases that are between self-deception and epistemic akrasia in which the person has obscure knowledge that his belief is not fully backed by good reason. For example, imagine the racist says that although it is not his only reason, one of the reasons he cannot believe himself to be a racist is that doing so would destroy himself. One might wonder what exactly he means when he speaks of such self-destruction, but it would seem to refer at least in part to a drastic and unpleasant adjustment to his self-conception.[47] To offer this as a reason why he believes he is not a racist is to admit that it is not the truth alone that grounds this self-belief. In this case, the racist has some awareness of the sort of satisfaction that causes him to believe what he does, though unlike the epistemically akratic parent, he believes this is not the primary reason he maintains his unwarranted belief. The present account can describe this intermediate case as well, for it allows that a mental state might at once be both rationally satisfying and affectively satisfying. If I do something praiseworthy, I may in reflecting upon my action find both rational satisfaction in believing what is true and affective satisfaction in thinking well of myself. In the present example, the racist acknowledges the affectively unsatisfying aspect of thinking that he is a racist, but he confuses this

[47] Velleman 2002 discusses the relation between this sort of "self-destruction" and selfhood. The discussion is a critique of Harry Frankfurt's thoughts on these matters. For a place where Frankfurt himself directly discusses the relevance of self-deception to these sorts of issues, see Frankfurt 1998b, pp. 106-07.

dissatisfaction with the combined dissatisfaction of thinking what is both unpleasant and false. Confused in this way, he acknowledges that it would be painful to conceive of himself as a racist but denies that this is his primary motivation for conceiving of himself as egalitarian.[48]

The account can also describe the third feature mentioned in section 3.2, that of the role of habit in causing self-deception or epistemic akrasia. Focus again on self-deception. So far it may have seemed as if the account describes all self-deception as including some phenomenally felt tension that is then resolved through self-deceptively false self-explanations. Surely this process is involved in some cases of self-deception, but in other cases affective satisfaction is realized without there being any phenomenally felt tension that is then resolved. In general, we achieve all sorts of affective satisfaction and avoid all sorts of affective dissatisfaction simply by acting in accordance with habit. Hunger brings displeasure, but I can avoid hunger without ever feeling hungry by maintaining a habitual eating schedule. By maintaining such a habitual schedule, I solve the problem of the dissatisfaction of hunger by never allowing it to arise.

There are habits other than those of daily routine that cause one to avoid dissatisfaction by not allowing it to arise. For example, suppose that I have found fast food unpleasant in the past. Suppose that one evening you give me the option of picking the kind of cuisine we will have for dinner, and suppose that the idea of eating fast food does not occur to me at all, not even as an option to rule out. The fact that I do not choose to eat fast food is part of my dietary habits. It is a behavioral regularity on my part, such that if I am presented with the option of eating fast food, I will explicitly reject

[48] It is thus that the present account describes various degrees of self-deception. On the idea that self-deception comes in degrees, see Lockie 2003, esp. pp. 142-43. For an account of self-deception that (I think, wrongly) requires a person to be between a state of self-deception and epistemic akrasia to count as self-deceived, see Scott-Kakures 1996.

it. If I am not given the choice, however, I implicitly reject fast food simply by not considering it at all.

We can imagine the racist avoiding the affective dissatisfaction he would find in believing himself to be a racist in the same way I avoid the affective dissatisfaction of eating fast food. The racist never allows the dissatisfaction to arise because he simply, from habit, conceives of himself as egalitarian.[49] Should a question arise regarding why he has done something, the thought that it might be due to racist motives simply does not occur to him, just as the thought of choosing fast food simply does not occur to me when the question of what to eat arises. *If* the thought that his action is racist crosses his mind, he will reject it, confusing affective satisfaction with rational satisfaction as I have described above. The point presently being made is that he can realize this confused state of satisfaction without the thought that his action is racist ever crossing his mind, just by thinking of himself as he habitually does.[50]

The racist's habit here is an epistemically bad one, for it causes him to form beliefs irrespectively of what is warranted. The good habit to have here—which, along with the racist, the epistemically akratic parent also lacks—would be the character trait of forming beliefs without affective satisfaction playing any role in the process. To do this, as we shall now see, is to be virtuous as the self-deceived and epistemically akratic individuals are not.

---

[49] The discussion of self-deception in Barnes 1997 is limited to cases in which self-deception is caused by an "anxious desire" and so fails to address the sort of case I am presently discussing. This limited scope is the major flaw of what is otherwise an illuminating inquiry into self-deception.

[50] My discussion here of self-deceptive explanatory habits bears some similarity to that found in Patten 2003. Patten describes self-deception as the result of a misapplied self-schema, which is a body of beliefs about one's general tendencies that one draws upon to explain one's own behavior. In cases like the racist's, this seems like an accurate description. Patten goes wrong, however, when he claims that self-schemata cause self-deception in unmotivated ways. If the role of these self-schemata were non-motivational, then we would not find the angry defensiveness self-deceived individuals commonly exhibit upon being accused of being self-deceived.

3.6 THE VIRTUE OF EPISTEMIC COURAGE

In claiming that the self-deceived and akratic individuals commit some failure of virtue, I claim that each is to be criticized for some failure of character. A person's character is composed by different character traits, which are the habits that we cite in explaining the tendency of a person to act in similar ways in similar situations. A trait is a virtue if it is one that causes a person to act in the correct way in correct circumstances, and to the extent that a trait fails to live up to this standard, it is one for which a person can be accused of being unvirtuous.[51]

What is distinctive about thinking of correctness in terms of character is best brought out by contrasting the ethics of virtue with other accounts of ethical correctness. Traditional consequentialist ethics, for example, take the consequences of action as the focal point of moral assessment. According to the virtue ethicist, an action is not good just because it produces good consequences; in order to be good, the action must also be produced in the right way, *viz.*, by virtue. The virtue ethicist claims that in order to be a good act, it is not enough that, *e.g.*, a consequence of the act is a charitable distribution of resources; in order to be good, it must be done because the person who performs the act has the virtue of charity. This also distinguishes virtue ethics from some traditional deontic ethical theories, which take a person's relation to a duty as the focal point of moral assessment. According to these theories, an act is good if it accords with a duty and the acting person performs the act for the sake of the duty. The virtue ethicist faults these sorts of theories for failing to acknowledge the importance of the attitude a person bears towards her duties, which itself can be better or worse and is relevant to moral assessment. There is a difference, the virtue ethicist notes, between the selfish person

[51] I will draw in the next few paragraphs from Hursthouse 2002, which provides an excellent introduction to virtue ethics.

who, out of a sense of obligation, forces himself to give to charity and the benevolent person who, acting in accordance with her character, gives to charity. Traditional deontic ethical theories do not mark this morally relevant distinction.

Virtue ethicists thus take themselves to be able to describe the full range of features we take to be relevant to moral assessment. They are often criticized, however, for not being able to give accounts that provide moral guidance. Their accounts might describe the variety of features we take to be relevant to assessing what someone has done, but, the complaint might be put, they are of no use for telling us what to do in morally complicated situations. Their accounts recommend that we do what the virtuous person would do, but if we are not virtuous, how are we to know what that is? Traditional consequentialist theories face no such problem; they guide us to act so as to produce the best consequences. Traditional deontic theories can also guide action—they tell us to follow our duties, which the deontic theorist presumes can be codified by a table of laws. Virtue ethicists have resisted this charge and have argued that appealing to what the virtuous person would do is in fact adequate for guiding action. Whatever action-guidance this is, it is not that of performing a utility calculation nor that of following a rigidly codified law. As we shall see, reflection on the sort of guidance virtue-theoretical accounts can offer will provide a reason for conceiving of self-deception and epistemic akrasia as failures of virtue. This is so because neither the self-deceived individual nor the epistemically akratic individual is in a position to be guided by the sorts of rules the consequentialist or deontic theorist has to offer.

Before saying more about rules and guidance, however, we should further specify the epistemic flaw common to both self-deception and epistemic akrasia. It might be thought that the common flaw is that in both conditions a person holds an unwarranted belief. If this is correct, then perhaps what fundamentally unites these phenomena is a

sort of irrationality, of a person irrationally holding an unwarranted belief. But a person can irrationally hold an unwarranted belief and be neither self-deceived nor epistemically akratic. She might explicitly hold one belief that is incompatible with other commitments of hers, but she may not be aware of the inconsistency simply because she has never considered the belief and the commitments at the same time. Call this person the *unreflectively inconsistent* person. This person holds a belief that is inconsistent with her other commitments and therefore is not fully rational, but she is also clearly unlike the self-deceived and epistemically akratic individuals. Even if it is true that both of these latter individuals irrationally believe what is unwarranted, something else distinguishes the error common to them from the unreflectively inconsistent person's error.

The difference between the unreflectively inconsistent person and the self-deceived and epistemically akratic individuals is that the mental activity that produces the former condition but not the latter conditions aims at rational satisfaction. Since the conclusion of all of their activity is belief formation or maintenance, the unreflectively inconsistent individual's activity is of the right sort, whereas the self-deceived and epistemic akratic individuals' activities are of the wrong sort. Given that the beliefs of the latter individuals result from activity that is in some sense wrongful, we should find it initially plausible that there is some ethical failure for which they but not the unreflectively inconsistent person are culpable.[52] If there is any common ethical wrongdoing in their cases, however, it is one that the consequentialist will find difficult to describe, for there are sorts of self-deception whose consequences can be good both for the self-deceived individual and for those affected by her being self-deceived. A cancer patient, for example, may out of optimism believe she will beat her disease in the face of

[52] This is not to say that the passively inconsistent person is not culpable for some other wrongdoing—perhaps she is epistemically negligent, or perhaps she fails to be epistemically virtuous for not doing more to bring consistency to her beliefs. If she is culpable for some wrongdoing, it is still not that of the self-deceived individual or the epistemic akratic.

contradictory statistical evidence, and this confidence may actually improve and extend her life.[53] In such a case, the consequences of being self-deceived benefit the life of the patient, her family, and her friends; in such a case, it is hard to see for whom there are any negative consequences. It is not true, then, that in all cases of self-deception or epistemic akrasia the consequences of being in the relevant condition are worse than if the person were not in the condition.

Attempts to describe the common wrongness of these conditions in deontic terms face their own problems. First, it is not clear that there is a duty violated by both the self-deceived and the epistemically akratic individuals but not by the unreflectively inconsistent person. If there is such a duty, it is not clear how it is to be codified as a single law other than 'Do not be self-deceived or epistemically akratic'. If this is the law that codifies the common wrongdoing, then if it is not to be ad-hoc it needs to be a specification of a more general law that embodies a general principle. The more general law cannot be 'Do not believe what you should not believe' or, more narrowly 'Do not maintain irrational commitments', for the unreflectively inconsistent person, who again is neither self-deceived not epistemically akratic, violates both of these laws. Nor can the more general law be 'Always know what you believe and why you believe it', for the epistemically akratic person knows what he believes and why he believes it. It is not clear that a deontic account can provide the needed law here in a manner that is not ad-hoc.

Even if this problem can be overcome and there is a principled law that singles out what is commonly violated by both and only the self-deceived individual and the epistemic akratic, it is impossible that such a law could serve the guiding function that laws are supposed to serve. It is at very least odd to conceive of the epistemically akratic

[53] For empirical investigations into the positive effects on health self-deception can have, see Taylor 1991.

person correcting his epistemic wrongdoing by checking some law that tells him not to be epistemically akratic, which in turn guides him to epistemic well-being. It is odd, because his epistemic akrasia is constituted by the fact that he already knows both that and why he believes what he should not. If it is possible in the case of the epistemically akratic person for such a law to provide guidance, though, it is impossible in the case of the self-deceived individual, for the self-deceived individual does not believe himself to be self-deceived and so cannot be guided to be otherwise by any law. If the self-deceived racist were to encounter some law commanding him not to be self-deceived, he would take himself already to have satisfied the law and not thereby be changed from his self-deceived condition. Similarly, it seems odd in the case of the epistemic akratic and is impossible in the case of the self-deceived individual to conceive of a utilitarian calculus guiding either person away from his mistake. If consequentialist and deontic ethical theories are supposed to have the benefit of providing guidance, then the fact that epistemic akrasia and self-deception are conditions away from which one cannot be straightforwardly guided counts against thinking of the failure commonly involved between the two as one properly characterized by consequentialist or deontic theories.

What makes the self-deceived individual and the epistemically akratic individual commonly bad is their shared failure of virtue. What would count as virtuous activity in both cases would be believing only what is warranted and not being swayed by the affective satisfaction of unwarranted beliefs. The character trait these individuals lack, which, if they had it, would make them virtuous in this regard, is well-described as the epistemic courage to believe only what is warranted.[54] This courage is sometimes

[54] I believe that what I am here calling "epistemic courage" is more general than what is described by Montmarquet 1987 as "intellectual courage" (cf. pp. 484 ff.). Montmarquet is interested in discussing the courage involved in considering alternatives to popularly held beliefs or, especially if one is a theoretician, to one's theories and hypotheses. To be courageous in the way Montmarquet describes is to have epistemic courage, but it is not the only way to have this courage. A person's mental activity is epistemically

figuratively described as the courage to face the truth. In cases of epistemic akrasia, the person knows she lacks this courage, for she knows that she believes what she is not warranted to believe. In cases of self-deception, the person confusedly believes himself to have this courage, but he lacks it. In either case, the virtuous trait of forming and maintaining beliefs only if the beliefs are warranted is lacking.

(I intend here only to describe the failure common between self-deception and epistemic akrasia. There are other culpable failures present in one but not the other case, most notably the self-deceived individual's lack of self-knowledge. This topic is taken up in the next chapter.)

To understand this courage properly as a virtue, we must keep clearly in mind that it is not the sort of courage we ascribe to someone when he overcomes some fear and does what duty demands. Courage as a character trait is not manifested by overcoming the obstacle of fear but rather by never having fear arise as an obstacle to overcome. The epistemically courageous person does not realize what he should believe, grow timid in the face of this belief, and then muster the strength to believe what he should. This sort of person is best described as epistemically continent. This sort of person is similar to the way I would be if I craved fast food yet, rightly believing it to be unhealthy, fought off the urge to eat it whenever the urge arose. I am not such a person, however, for it is part of my character that I have no such urges. The epistemically courageous person is similar to me, for she has no urges to believe anything other than the truth. The epistemically courageous person simply believes what she should because of her epistemically courageous character. [55]

---

courageous whenever it is such that, were the person to lack the courage, it would be guided by affective satisfaction.

[55] For an explanation of the importance of this difference between continence and virtue, see McDowell 1998b, esp. pp. 46-49.

This difference between epistemic continence and epistemic virtue can be easy to miss if one does not attend to the role that habits can play in mental activity. Those who write on epistemic akrasia, for example, tend to describe the virtue that the epistemic akratic lacks as epistemic continence.[56] They are led to do so, I think, because they tend to consider the problem that the truth poses to the epistemic akratic on the reactive model of self-deception described at the end of section 3.2. To see the truth as a threat in the first place, however, is already to lack courage; to see the truth as a threat is, in Augustine's words, "a morbid condition of the mind which, when it is lifted up by the truth, does not unreservedly rise to it but is weighed down by habit."[57] This bad habit of the mind is one that leads its activity away from the truth, either through epistemic akrasia or self-deception. Only if a person is completely free of the effects of this habit is she virtuous in her mental activity.

We thus see the importance of considering mental activity in the course of discussing self-deception and epistemic akrasia. By raising these topics in the context of mental activity, we are able to consider both of them as the possible result of habits of the mind. By discussing habit, we are led to think of these epistemically erroneous phenomena in terms of character and virtue. By reflecting on the lack of virtue in the bad case, we are able to see the good case—that of epistemic courage—as a virtue. Only by keeping habits firmly in mind do we see this virtue of epistemic courage for what it is: a constancy of character. One has this constancy when one is led by nothing other than the truth in forming beliefs.

[56] See, *e.g.*, Hookway 2001 and Heil 1984. It is worth noting here that the topic of virtue is not explicitly addressed anywhere in the literature on self-deception.
[57] *Confessions*, VIII.ix (21) (p. 148 in the Chadwick translation).

3.7 APPENDIX: THE "PARADOXES" OF SELF-DECEPTION

Most of the contemporary literature on self-deception is concerned with answering one or another of the various "paradoxes" that the phenomenon appears to pose. It has seemed to some philosophers that when a person is self-deceived, he must simultaneously yet paradoxically hold both a belief and that belief's negation; call this the doxastic paradox of self-deception. It has seemed to some that the self-deceived individual must will himself to believe what he wishes were true, thereby achieving a belief at will; call this the paradox of wishful thinking. It has seemed to some that the self-deceived person must repress some belief, which he can only successfully manage by paradoxically being both aware and unaware of the repressed belief; call this the paradox of repression. Finally, it has seemed to some that in order to arrive at his condition, the self-deceived individual must dupe himself into believing something without realizing he is being duped; call this the strategy paradox.[58]

These so-called paradoxes only arise, however, if the phenomenon of self-deception is improperly characterized: none of them arise on the account currently being presented. The doxastic paradox does not arise, for on the present account the self-deceived person is not understood as holding a contradictory pair of beliefs. He *should* believe what he does not, and to that extent he is irrational, but it is precisely because he does not believe what by his own lights he should that he is self-deceived and not in some doxastically paradoxical condition. The self-deceived cuckold, for example, should believe that his wife is cheating on him, but he does not; if you ask him why she did not come home last night, he will construct an explanation compatible with the belief that she

[58] Johnston 1988 calls the first of these purported paradoxes "the surface paradox" (p. 63); it is at the heart of what Mele 1987 calls the "paradox of belief" (cf. ch. 9). I take the titles for the second and third purported paradoxes from Johnston 1988 (cf. pp. 70, 76); I take the title of the fourth from Mele 1987 (cf. ch. 10).

is faithful and incompatible with the belief that she is cheating. He also arguably should know that the cause of his irrational belief is his irrational motivation to hold the belief true and not a desire for the truth; this too he fails flatly to believe, in no doxastic tension with any belief to the contrary.[59]

The paradox of wishful thinking does not arise because the present account does not understand the self-deceived individual as performing any mysterious act of believing what he wants at will. We might characterize his condition as involving a sort of wishful believing, but because he confuses what is affectively satisfying with what is true, he does not willfully believe what he wishes were true but rather mistakenly believes that his wishes are true. The paradox of repression does not arise on the present account, for it does not understand the self-deceived individual as performing any mysterious act of simultaneously acknowledging and not acknowledging the existence of some repressed belief that he must struggle to hold at bay. The tension that is often discussed under the head of 'repression' is just the tension of rational acceptability tending the mind towards one belief while affective acceptability confusedly tends the mind towards an incompatible belief. Nor does the strategy paradox arise on the present account, for it does not understand the self-deceived person as mysteriously acting on a strategy to deceive himself that can succeed only if he lacks knowledge of said strategy. His epistemically irrational interest in avoiding the affective dissatisfaction of certain beliefs regardless of their truth is served simply by mistaking one sort of acceptability for another.

Some (*e.g.*, Donald Davidson and Mark Johnston) have thought that the deepest paradox involved in self-deception is none of these just mentioned, but rather the apparently paradoxical fact that in self-deception, a mental state or event causes a belief

[59] I say "arguably" here, but I believe that the self-deceived person should, in fact, know the cause of his unwarranted belief. I argue for this claim in the next chapter.

for which the former is not a reason—call this the paradox of irrationality.[60] If we assume that one mental state or event can cause a belief within the same mind only if the cause is a reason for the effect, and if we assume that a mental state or event can be a reason for a belief only if the former is relevant to the truth of the latter, then forming a self-deceptive belief is paradoxical. The cause of the resulting belief is not relevant to the truth of the resulting belief, which is incompatible with the assumptions just presented.

While a self-deceived individual is liable to charges of irrationality, on the present account that irrationality is not understood as paradoxical in the manner just described. There is a sense in which the self-deceived person's condition is amenable to explanation in terms of reasons, if we consider the affectively unsatisfying beliefs and explanations he manages to avoid generally as mental items, not specifically as doxastic or epistemic items. Thinking this way, we can say that the avoided mental items are potential sources of dissatisfaction for the self-deceived individual, and so to the extent that it is practically reasonable for him not to suffer undue dissatisfaction, he has a reason to reject these mental items. In this sense, an interpreter might make sense of the self-deceived individual by ascribing to him the practical reason of not enduring avoidable dissatisfaction, although the self-deceived person will of course reject this ascription as incorrectly explaining the self-deceptive attitudes he holds.

In characterizing the self-deceived person in this way, the interpreter accuses that person of irrationality, but not of paradoxical irrationality. Because the self-deceived individual believes something for which he lacks warrant, it is right to say that he is being epistemically irrational. This irrationality appears paradoxical, however, only if we further assume that the only mental states or events that can in one and the same mind cause a belief must bear on the truth of the resulting belief. In self-deception there are no

[60] On this, see Davidson 2004b, esp. pp. 179-181. See also Johnston 1988, pp. 79ff., which addresses Davidson's thought on these matters.

states or events that bear on the truth of the relevant resulting belief, which on the current assumptions implies that the resulting belief is not caused by a reason, which implies it is not mentally caused, which implies it is not caused at all. The claim in this chain of implications to give up, I suggest, is that the only possible intramental causes for a given belief are events or states that are relevant to the truth of the resulting beliefs. These may be the only epistemically rational causes, but again, believing and explaining are mental activities and as such are describable in the language that is generally appropriate for action, which can explain an act as occurring simply because the act causes affective dissatisfaction to be avoided.

4.0 UNDERSTANDING AND SELF-KNOWLEDGE

In the last chapter, I argued that the wrongdoing that is common to self-deception and epistemic akrasia should be understood as a lack of epistemic courage on the part of self-deceived and epistemically akratic individuals. In saying this, I have executed one of the projects I set for myself in section 2.6. That project, recall, was put forward by the question, for what should we criticize the self-deceived individual, given that he has a *motivated* lack of self-knowledge? The answer is that we should criticize him—as well as the epistemically akratic individual—for a lack of virtue. We should wonder, however, whether the self-deceived individual is criticizable for something further, *viz.*, his lack of self-knowledge. In saying this, I call to mind the second project announced in section 2.6, which is to describe the epistemic wrongdoing involved in the self-deceived individual's failure of self-knowledge. This project is worth pursuing because the only way to say what is epistemically wrong with the self-deceived individual's failure of self-knowledge is to give an account of the way in which self-knowledge is epistemically unique. That is the task of this chapter.

I want to pursue this task by first mentioning some of the different conclusions about self-knowledge one might draw on the basis of the existence of self-deception. Anyone who writes on self-knowledge should acknowledge the fact that self-deception happens and should accordingly not argue that the knowledge one has of one's own mind is infallible. It is commonplace to recognize this.[61] Many who write on self-knowledge—at least in the philosophical literature—go on to assert that even if self-knowledge is not infallible, there is still something epistemically unique about it. To take

[61] Cf. section 2.0.

this position is to give oneself the task of explaining what this epistemic uniqueness is and why it is a facet of self-knowledge. One might, however, argue that the existence of self-deception demonstrates that, far from being infallible, self-knowledge is not even epistemically unique. To take this position is to maintain that the epistemic wrongdoing of the self-deceived individual does not differ in kind from that of a person who has false beliefs about the reasons, intentions, desires, or beliefs of another. Let us imagine someone holding this position and call him *the Sceptic*, for he is sceptical of the idea that there is anything epistemically unique about self-knowledge. The person who believes that self-knowledge is epistemically unique cannot agree with what the Sceptic maintains about self-deception. Since the former person believes that self-knowledge is epistemically unique, she must also hold that a failure of self-knowledge, such as that which occurs in self-deception, is an epistemically unique sort of failure.

I hold the view just described, and in this chapter I am going explain the epistemic uniqueness of self-knowledge by showing why the Sceptic must be wrong. I will argue that the account of the mind to which the Sceptic is committed cannot describe what it is to understand one's own reasons. To argue this, I will describe in detail what I have in mind when I speak of understanding one's own reasons. A person has this sort of understanding when her faculty for productive reasoning functions well, and as I will base my account of the epistemic uniqueness of self-knowledge on this fact, my account is well-described as rationalistic. As such, it is similar in important respects to the recent rationalistic accounts of Tyler Burge and Richard Moran. The arguments that they put forward for their accounts, however, are not adequate for showing the Sceptic to be wrong. My project here may thus be understood as further developing the rationalistic sort of accounts put forward separately by Burge and Moran.

My strategy here of arguing against the Sceptic my appear to be a needlessly complicated way of arriving at my goal. I have chosen to proceed in this manner because I think it is a good way of homing in on what is at stake in saying that self-knowledge is *epistemically* unique. As I will now show, there is much that has been recently written on self-knowledge that the Sceptic need not deny. Since he need not deny it yet can go on being sceptical in his way, the accounts to be presently discussed do not describe what is epistemically unique about self-knowledge. Showing this will help us focus on what it means to say that self-knowledge is epistemically unique and what, in turn, is needed to establish that it is so unique.

4.1 THE SCEPTIC AND LINGUISTIC APPROACHES TO SELF-KNOWLEDGE

In maintaining that self-knowledge is not epistemically unique, the Sceptic is committed to the following thesis: self-knowledge is not grounded in a sort of entitlement that is uniquely first-personal. Being committed to this thesis does not force the Sceptic to deny that there are certain objects of knowledge that are uniquely knowable from a first-person position. The Sceptic may grant that a person is uniquely positioned to know of states like her own euphoria and dysphoria. A person can know that she is in a euphoric or dysphoric condition by feeling it, and no one else can feel her pleasure or pain. The Sceptic also may grant that a person is uniquely positioned to know her own occurrent thoughts. When, for example, a person daydreams, she alone knows the details of her fantasy. The Sceptic is committed to denying, however, that from the fact that a person is uniquely positioned to have these sorts of self-knowledge, it follows that she is entitled in a uniquely first-personal way to these or other sorts of self-knowledge. The Sceptic may grant that the epistemology of these sorts of self-knowledge is either that of inner

observation, or of comprehending inner testimony, or of making quick unconscious inferences. The Sceptic must insist, however, that neither the internality of the observation or testimony nor the unconscious quickness of the inferences makes the entitlement in question distinct in kind from that which one has to normal, external observation or testimony or to normal, conscious inferences. A good account of the way in which observation, testimony, and inference allow us to know the minds of others will be fully adequate, according to the Sceptic, for accounting for the knowledge we have of our own minds.[62]

The Sceptic thus need not deny the claims made by one linguistic approach to self-knowledge, which I shall call the *semantic-authoritative approach*, because this approach does not purport to establish that we are uniquely entitled to knowledge of our own mental states and events. The goal of this approach to self-knowledge is simply to establish that we have unique first-person knowledge of what we are saying when we speak, which cannot be threatened by accounts of mental and semantic content that might make it seem otherwise. Consider, for example, Donald Davidson's account of self-knowledge. Davidson claims that whenever a speaker makes an utterance she knows what her words mean in a way in which her interlocutor cannot. An interpreter can always wonder whether a the speaker's meaning is best captured by the syntactically identical sentence of the metalanguage or whether some other metalinguistic sentence better characterizes the speaker's meaning. For example, an interpreter can always wonder whether "'Wagner died happy' is true if and only if Wagner died happy" successfully specifies the speaker's meaning or whether some other sentence of the metalanguage better captures the speaker's meaning. While an interpreter can wonder about this, a speaker cannot. Davidson says that she cannot improve on sentences like

[62] The Sceptic finds much agreement with the positions and arguments of Ryle 1949.

the one quoted above in understanding what her words mean. A speaker non-interpretively knows what her words mean, and because she non-interpretively knows what they mean her knowledge of their meaning is uniquely first-personal.[63]

Precisely because this is the sort of authority Davidson's account provides for, it is of no worry to the Sceptic. The Sceptic has no problem granting that a person enjoys some sort of authority over what her words mean. He holds no position regarding the basis of this semantic authority: it might derive from the person's idiolect, or it might derive from her upbringing in a particular linguistic community, or it might derive from something else. Whatever the basis is for this authority, it is no basis for the person's knowing when she claims to hold a belief that, in fact, she holds that belief. Davidson's account can only show that a speaker enjoys a certain authority concerning *what* belief is averred when she speaks. It cannot show that the belief that is averred is, in fact, one that the speaker holds. If failures of self-knowledge like self-deception are not epistemically abnormal, then the authority a person has over what her words mean only guarantees that she enjoys a unique sort of knowledge regarding what she *takes herself* to believe. Whether she *actually* believes what she takes herself to believe is an entirely different matter.

The inability of Davidson's account to answer the Sceptic's concern is the result of Davidson's conceiving of the issue of self-knowledge within the context of his interpretivist account of mental and semantic content. Unsurprisingly, this same inability is present with other accounts that are motivated in the first place by threats to semantic authority posed by otherwise laudable theories of mental and semantic content. The literature concerning self-knowledge and semantic externalism or semantic anti-

[63] This account is presented in Davidson 2001b.

individualism is full of such accounts.[64] The single goal of this literature is to explain whether and how a person knows *what* he is thinking when he thinks, but to establish that is to fall short of showing that the content considered while thinking is in fact the content of some belief, desire, or intention that the person holds. Even if an account of this sort meets its explanatory goals, it will not be adequate for showing the Sceptic to be wrong.

The thought that the semantic-authoritative approach fails to provide all we might want from an account of self-knowledge is not novel. Crispin Wright, for example, agrees that this approach to self-knowledge cannot explain the presumption that a person authoritatively knows not just the content of her utterances but her beliefs themselves.[65] To explain this presumption of authority, Wright tentatively forwards an account that he calls the 'Default View.' This view is an instance of another sort of linguistic approach to self-knowledge, which I shall call the *pragmatic-authoritative* approach. Accounts that take this approach focus on the speech act of avowal and seek to explain the pragmatic authority we afford to such speech acts in conversation. Some pragmatic-authoritative accounts go no further than this, while others attempt to derive epistemological consequences for self-knowledge from their accounts.

Wright's Default View is an instance of the former sort of account. According to this view, it is not because of any unique first-person entitlement that we are inclined to treat a person's avowals as authoritative. Rather, Wright says, according to the view "it is just primitively constitutive of the acceptability of psychological claims that, save in cases whose justification would involve active self-interpretation, a subject's opinions about herself are default-authoritative and default-limitative."[66] This view focuses on what it is that makes a psychological ascription acceptable, and it holds that it is simply

[64] See, for example, the essays in Ludlow and Martin 1998.
[65] Wright 2001c discusses this approach to self-knowledge specifically as it occurs in Davidson's and Burge's writings.
[66] Wright 2001c, 369-70.

constitutive of psychological concepts that when they are deployed in avowals, their status is acceptable by default. By calling this fact 'constitutive,' Wright means that the view rules out both that it can be explained and that it so much as needs to be explained. In particular, the fact does not need to be explained by any constructive epistemology, which would account for the fact in terms of a special entitlement a person has to her avowable attitudes. Wright characterizes this view as in line with the later Wittgenstein's hostility to needless philosophical explanation.[67]

Wright himself worries that this view might just be a "merely an unphilosophical turning of the back."[68] The Sceptic, however, does not conceive of it as an unphilosophical turning of the back but rather applauds it as a justifiable staring through an illusion. The illusion is that of a unique entitlement to self-knowledge underpinning the pragmatic authority of avowals. It may be true that as a matter of custom we treat avowals as correct by default, and it may even be true that we are right to do so because avowals are, more often than not, accurate. That this is true, the Sceptic maintains in agreement with the Default View, is not grounds for saying that there is a unique first-person entitlement in virtue of which avowals have this default status. As far as the Sceptic is concerned, this aspect of the Default View is correct.

Not all pragmatic-authoritative accounts are as epistemically modest as the Default View. Consider the description of the 'middle road' to self-knowledge discussed

[67] For an alternative reading of the relevant parts of Wittgenstein, see McDowell 1998d. McDowell argues that Wright misidentifies the target conception that concerns Wittgenstein: while Wright takes it that it is the epistemic asymmetries between first-person knowledge and other sorts of knowledge that need to be "deconstructed" as not calling for explanation, McDowell claims that Wittgenstein's target is the idea of an inner world conceived of as "a good place for a stand against the encroachments of a pernicious idealism" (p. 61). On McDowell's reading, the explanatory demand that Wright believes Wittgenstein rejects is not articulated by Wittgenstein as a demand in the first place. The Sceptic can ignore this debate. He is not anxious about the encroachments of a pernicious idealism.

[68] Wright 2001c, p. 369. Again, as a reading of Wittgenstein, McDowell 1998d finds the suggestion that Wittgenstein is performing some back-turning here to be wrong. According to McDowell, the demand to which Wright would have Wittgenstein turning his back is simply not there in the first place for Wittgenstein to turn his back on.

in Dorit Bar-On's book *Speaking My Mind*.[69] According to this account, a person is uniquely entitled to knowledge of any mental state that he can express through an avowal. He is so entitled in part because of an epistemic immunity to error through misascription that he enjoys when he makes an avowal.[70] Bar-On presents this immunity by asking us to consider how odd it is for a person to say the following sorts of things: "I am feeling *something*, but is it thirst?", "I am mad at *someone*, but is it *you*?", "I'm in *some* state, but is it being mad?".[71] She does not claim that such questions are non-sensical, but that when a person asks such questions he undertakes to discover something that he is normally in a position to know without having to conduct an investigation. When a person is in the normal position, he can simply say, *e.g.*, "I am thirsty," "I am mad at you," or "I am wondering whether it's time to leave." Because no investigation is required in order to acceptably self-ascribe these states and attitudes, they are immune to ascriptive error.

Bar-On builds her account of the privileged status of self-knowledge on the basis of this immunity to ascriptive error of avowals. The details of this construction are irrelevant to the Sceptic, however, who demands to be shown why the thesis of immunity to error through misascription is based on anything other than a mere custom. Bar-On provides no answer to this demand. She offers a test for determining whether a component of an ascription rests on a recognitional judgment: append to the ascription "It appears (to me) that . . ." and see if what results is anomalous.[72] If it is not anomalous,

---

[69] Bar-On 2004. Bar-On describes this account of self-knowledge as the 'middle road' between the 'low road' of reliabilism and the 'high road' pursued by Burge and Richard Moran, which we will soon discuss.
[70] According to Bar-On, his entitlement also partially rests on an immunity to error though misidentification. This immunity can be characterized as follows: when a subject makes an avowal, he cannot know that the content of the avowal is true but wonder of whom it is true. This topic has been discussed at length by, among others, Strawson 1966, Shoemaker 1968, and Evans 1982. The Sceptic can accept that we enjoy this immunity to error through misidentification—it is the immunity to error through misascription of which he is sceptical.
[71] Bar-On 2004, p. 193.
[72] Bar-On 2004, p. 194. She credits the test to Ram Neta.

the ascription is based on some sort of investigation, but if it is anomalous, then the ascription is not based on an investigation and so is immune to ascriptive error. To the Sceptic, this test is only useful for probing our linguistic intuitions. This falls short of establishing that these intuitions are based on the existence of a unique first-person entitlement as opposed to mere custom. Bar-On has thus not shown what she needs in order to answer the Sceptic.

The Sceptic's response to Bar-On here is of a sort that he will take against any account of privileged self-knowledge that is built out of facts about speech acts. For the Sceptic, descriptions of the ways we talk provide no grounds for drawing epistemological conclusions. What does provide such grounds are psychological descriptions, which are explicitly ignored by pragmatic-authoritative accounts. Asymmetries between avowals and other sorts of ascriptions are to the Sceptic merely evidence of our linguistic customs and conventions; they are no grounds for showing that there are genuine epistemic asymmetries of the sort we presume to exist.

The Sceptic, then, is not disarmed by either of the linguistic approaches to self-knowledge just described. The semantic-authoritative approach does not answer his demand, for it only can show that we have special knowledge of what our words mean when we speak. The pragmatic-authoritative approach does not answer his demand, for it cannot establish that there are uniquely first-personal epistemic grounds for adhering to our custom of treating avowals as authoritative. These sorts of accounts thus do not establish that there is a unique first-person entitlement that grounds self-knowledge.

4.2 THE RATIONALISTIC APPROACH TO SELF-KNOWLEDGE

The next sort of approach we shall consider—which I shall call the *rationalistic approach*—has the resources to answer the Sceptic's demands, but the accounts of this sort put forward to date have yet to say what is needed to refute him. The accounts I shall discuss belong to Burge and to Moran.[73] Both approach the topic of self-knowledge by investigating the conditions under which explicit acts of reasoning are possible.[74] Both argue that such explicit acts are only possible if a person can be uniquely entitled to knowledge of her reasons, or at least to knowledge of those reasons that occur while reasoning. They differ on the conditions under which this entitlement is enjoyed, but both of their accounts develop from the common thought that the knowledge a person normally has of her own reasons is epistemically unique.

Both Burge and Moran conceive of this epistemic uniqueness as a consequence of the rational unity of the person. Burge describes this unity as the unity of a point of view. When a person assesses the rational credentials of another's belief, there are two points of view: the one to which the belief under review belongs, and the other from which the rational review is conducted. In contrast, when a person engages in critical reasoning and assesses the credentials of her own belief, the belief being reviewed and the judgments involved in conducting the review belong to one and the same point of view. In the latter case, but not the former, there is what Burge calls an *immediate* relation of rational

[73] See Burge 1998a and Moran 2001. The thought that a person's unique entitlement to self-knowledge stems from the person's capacity to reason, either by deliberating or by responding to a demand to justify one's attitudes with reasons, can also be found in Shoemaker 1988, Bilgrami 2006, Gallois 2004, and O'Brien 2005. The response that will soon be presented on behalf of the Sceptic against Burge and Moran suffices as a sceptical response to any of these other authors.

[74] Burge focuses exclusively on critical reasoning, the reasoning a person performs when she assesses the rational credentials of some belief she holds. Moran discusses this sort of reasoning as well, but he also discusses two other sorts of activity in which a person considers her reasons. The first of these is the sort of reasoning a person performs over the course of making up her mind about what to believe or to do. The second is the process a person engages in when she is asked to provide reasons for one of her actions or beliefs.

relevance of the reviewing judgments to the belief being reviewed. It is in virtue of this rationally immediate relation that one's entitlement to self-knowledge, as such knowledge is involved in critical reasoning, is based on one's rational capacities and is uniquely first-personal.[75]

An example will help clarify the connection Burge finds between this immediate rational relation and the unique entitlement the critical reasoner has to self-knowledge. Suppose that we are both trying to solve a crime and that while I think Alberto did it, you think Karl did it. The evidence I have provides better grounds for believing that Alberto did it, and so I judge that Alberto did it. I believe that we have the same evidence, and because of this I judge your belief that Karl did it to be unreasonable. Suppose, however, that unbeknownst to me your belief is based on a trustworthy article you have read that strongly makes the case that Karl did it. Suppose further that my ignorance of the basis of your belief is something for which I am not rationally culpable—suppose, *e.g.*, I have not had the opportunity to ask you why you think Karl did it. My judgment that your belief is not reasonable is rationally irrelevant to whether or not you should believe that Karl did it. Such a judgment of mine is only ever rationally relevant to you if I am not ignorant of all of the beliefs that ground your conclusion, and this will only ever be a contingent matter. In the case under consideration, this contingency does not obtain

[75] In Burge 1998a and Burge 1998b, the term 'immediate' and its cognates are used in at least three distinct ways. It is used to refer to the rational immediacy presently to be described, which is its primary usage in Burge's argument. This rational immediacy is normative and obtains whether or not a person makes a judgment on the basis of it. Burge also describes the judgment that a person might make on the basis of this rational immediacy as 'immediate'—the term is used here to describe the act of judgment as non-inferential. In the good case, by making a judgment regarding what is most reasonable to believe the person thereby comes to adopt the relevant belief immediately. Saying this exemplifies the third way in which Burge uses the term, which is meant to contrast with the way in which judgments of reasonability can affect the beliefs of others. If a person judges that another has most reason to believe something, the person must adopt some means if she is to get the other to believe what the judging person deems most reasonable. No such means are needed in the first-person case. (Moran 2001 calls this last sort of immediacy 'practical immediacy'—cf. p. 131.)

because I fail to know your reasons for believing that Karl did it. Again, as we have imagined the case, this failure is not one for which I am rationally culpable.

With judgments about the reasonableness of one's own attitudes, no such contingency mediates the rational relevance of the evaluative judgment to the belief being evaluated. In cases where it is one's own attitudes that are under rational review, the reviewing point of view and the reviewed point of view are one and the same. In this case, the considerations that ground the belief under critical evaluation are only those that the person entertains towards making the evaluative judgment, so the rational relevance of the evaluative judgment to the evaluated belief is immediate, not contingent. The way Burge conceives of things, these grounds are the beliefs that constitute the person's reasons for holding the belief being evaluated.[76] Since in critical reasoning these grounding beliefs are one's own, any knowledge one has of those beliefs is self-knowledge. If one fails to know one's own reasons while attempting to reason critically, then whatever one is doing is not critical reasoning. To try but fail to reason critically is to commit a rational failure. To fail to know one's own reasons while attempting to reason critically is thus to commit a rational failure. Since such a failure is a failure of the well-functioning of one's rational capacities, success in this regard—*i.e.*, having knowledge of one's own reasons—is grounded in the well-functioning of these rational capacities. The entitlement one has to self-knowledge in critical reasoning thus derives from one's own rational capacities. So, Burge concludes, this entitlement is a uniquely rational first-person entitlement.[77]

---

[76] One is not required to think that a consideration or belief grounds some other belief by the former itself being the grounds for the latter. As we shall see, Moran thinks that the ground for holding a belief is not some other belief but rather the way the world is. On Moran's conception, then, a belief grounds another belief not by being the ground but rather, we might figuratively say, by putting the person in touch with the ground, which is the way the world is. This difference in thinking about reasons and beliefs might make Burge's and Moran's accounts appear to contrast with each other in ways that they do not; cf. fn. 77.

[77] I have added a step to the argument here that is not in Burge 1998a. The added step is the claim that to fail to know one's own reasons when reasoning critically is to try but to fail to reason critically, which is a

Moran's argument for the uniquely first-person entitlement involved in self-knowledge hinges upon his account of what he calls a "transparent relation" between the question of what one believes and the question of how one judges the world to be. Consider the example of Alberto and Karl again. If I want to settle the question of who you think committed the crime, I need to investigate you. This investigation might be as simple as asking you what you believe, but since for me your belief is a fact about you to be discovered, I must investigate you in order to discover it. In my own case, I normally do not have to investigate myself in order to settle the question of what I believe, for my belief about the matter is constituted by the judgment I make about the world. In order to figure out whether I believe Alberto or Karl did it, I think about the evidence and whether, based on it, I should judge Alberto or Karl to be guilty. Indeed, what I figure out here is not so much what I *do* believe as what *to* believe. Because of the transparency in my own case between the questions of how the world is and how I believe the world to be, I can answer the latter question by answering the former.

This relation of transparency depends upon the rational unity of the person. I am able to settle the question of what I believe by settling the question of how the world is because of a certain unity of reason I must recognize if I am to be rational. The unity in question is that of the reasons that justify my belief with those that psychologically explain its existence. I can be fully rational yet fail to find such a unity in your reasons: if you draw a conclusion from a belief I know to be false, I can psychologically explain the existence of your conclusion in terms of your false belief without thereby taking the conclusion to be one I could justifiably hold. In my own case, if I am being rational, there can be no such distinction between beliefs whose existence I can psychologically

---

rational failure. Burge asserts that the failure of self-knowledge here is a rational failure without saying it is a failure to accomplish an attempted act of reasoning. I intend the added step to be in the spirit of Burge's original argument.

explain and beliefs I can justifiably hold. If I take some belief of mine to be unjustified, I cannot, while being rational, ignore the fact that it is not justified by providing reasons that psychologically explain its existence. If I am to be rational, I must acknowledge what Moran describes as the priority of justification over mere psychological explanation in considering my own reasons. Because I must acknowledge this priority if I am to be rational, I can know the reasons that explain my belief by reflecting on what I take to be the justification for the belief. Moran does not claim that we can know all of our reasons in this way: we can be self-deceived about why we hold some belief, in which case the reasons we take to justify the belief do not correctly explain why we hold it. We can only conceive of self-deception as abnormal, however, if we recognize it as a deviation from our standard epistemic condition, in which we know the reasons that explain our belief through reflection on those reasons that justify holding it.[78]

Moran argues that this holds for any attitude that is subject to rational criticism, not just for those attitudes actually considered in the course of reasoning. This holds for beliefs, for intentions, and for any desire about whose object an interlocutor might reasonably ask, "Why do you want that?". In any case in which a person is rationally answerable for her attitude, the relation she bears to her attitude is one that involves the priority just described of justification over mere psychological explanation. This relation of priority is itself prior to any judgment or justification a person might make or provide on its basis. The knowledge a person has of any attitude for which she can present a

[78] Moran's claims about transparency and the priority of justification over mere psychological explanation can be constructed out of various things said in Burge 1998a. First, Burge describes critical reasoning as "reasoning guided by an appreciation, use, and assessment of reasons and reasoning as such" (p. 246). He then claims that "it is arguable that use of *therefore* in reasoning—deductive or otherwise—constitutes an exercise of this meta-cognitive ability" (p. 246). If this is so, then Burge's topic can be understood to include any act of the mind whereby a person knowingly draws a conclusion on the basis of reasons. This will include conclusions about what one believes drawn by answering the transparent question of how the world is. When one reasons in this way, one, in Burge's terms, "appreciat[es] the force and relevance of reasons to attitudes as such" (p. 246, fn. 3). To appreciate this force just is to prioritize justification over mere explanation in the first-person case.

justification is thus knowledge that is attainable through reflecting on and making a judgment about what there is in the world, what to do in the world, what in the world is desirable. None of these is a judgment about the way things are psychologically with oneself. The knowledge that is had through making these worldly judgments is uniquely first-personal and is entitled by the well-functioning of the person's rational capacities. The entitlement one has to this self-knowledge is thus a distinctly rational first-person entitlement.

Here we see the fundamental point of agreement between Burge and Moran. It is because we have knowledge of our own reasons that is uniquely entitled by our rational capacities that we enjoy a more general entitlement to self-knowledge that is uniquely first-personal. The knowledge we have of our reasons when we are being rational is not like the knowledge we have of others' reasons. Because of the unity of personhood, it is knowledge based on the rationally immediate relation between the reasons and the attitudes they are reasons for. Because of the unity of personhood, it is knowledge based on the primacy of justification over mere psychological explanation. Such rational bases can only entitle self-knowledge; they cannot entitle knowledge of another's reasons. To say this is not to say that all self-knowledge is based on this unique sort of entitlement, for it sometimes happens that we gain self-knowledge by the same investigative means that we gain knowledge about others. Still, when we are rational, we are uniquely entitled by the well-functioning of our rational capacities to knowledge of our own reasons and to knowledge of the attitudes grounded by those reasons.

4.3 THE SCEPTIC'S RESPONSE TO THE RATIONALISTIC APPROACH

The Sceptic's complaint against Burge and Moran is that they both illegitimately derive an epistemological conclusion from premises about norms of reasoning. The Sceptic need not doubt that there are such norms; he need not doubt that some reasons are better than others. He denies, however, that the existence of any such norms provides grounds for concluding that a person is uniquely entitled to knowledge of her own reasons. Nothing that Burge says convinces the Sceptic that he is required to draw the epistemic conclusion Burge draws from the fact of rational immediacy. Against Burge, he says that even if failure to know one's own reasons while attempting to reason critically constitutes a failure of rationality, it does not follow that knowledge of one's own reasons when it is had during critical reasoning rests on any unique first-person entitlement. We may be subject to certain norms when we reason critically, and we may have to know our own reasons in order to satisfy these norms, but from these facts we need not conclude that this knowledge is grounded by an entitlement that is uniquely first-personal.

Unsurprisingly, Burge asserts that the facts just mentioned do force us to conclude that self-knowledge in critical reasoning requires a unique entitlement. Burge's argumentative move here is transcendental: he asserts that were there no such unique entitlement, there would not exist the immediate rational relevance of the evaluating judgment to the judgment under evaluation that is constitutive of critical reasoning. Since critical reasoning exists, Burge claims, the immediate rational relevance that partially constitutes this reasoning and the entitlement that conditions the existence of the immediate relation must also exist.[79] What Burge puts forward as a transcendental condition, however, the Sceptic sees as a question-begging assumption. In order to

[79] See Burge 1998a, p. 256 for this argument.

silence the Sceptic, Burge needs to provide some reason for accepting that the entitlement he discusses is necessary for critical reasoning, but he provides no such reason. The Sceptic thus finds no reason to grant what Burge insists must be true.

The Sceptic similarly finds nothing in Moran's account that requires him to draw a conclusion about the epistemic uniqueness of self-knowledge from the priority of justification over mere psychological explanation in considering one's own reasons. The Sceptic can grant that if we are to be rational, we must give justifications priority over mere psychological explanations when providing reasons for our beliefs and actions. He can grant that in order to do this our judgments must be aimed primarily at the world and not at our own psychological states when we consider our own reasons. He will insist, however, that whenever a person prioritizes justifications over mere explanations in her thought, she can only ever be thinking about the reasons for which something *should* be done or believed, which on any given occasion may not be those that explain what *is* done or believed. To accept the priority of justification over mere psychological explanation is only to accept that a certain movement of thought—of a person providing a mere psychological explanation for what she thinks or does when justification is called for—is illegitimate. One can accept this while maintaining that any justificatory movement of this sort, be it legitimate or illegitimate, is at best a symptom for the movement of thought that correctly explains why the person thinks or acts as she does. If this is so, then even if she has some special sort of knowledge of her own justifications, such knowledge can serve at best as a basis for inferring how things really are with her psychologically. Moran seems to think that were we in such a condition we would be radically alienated from our own minds but that, as we can tell simply by reflecting on our quotidian lives, we know we are not so alienated. The Sceptic, being a sceptic, is not

impressed with this argument and thinks that this condition, even if it arouses anxieties of alienation, is simply how things are with us.

If challenged to offer an alternative account of the epistemology of knowledge of one's own reasons, the Sceptic might proceed as follows. First, he divides his account into two parts: one concerns the knowledge we have of reasons that do not occur in explicit acts of reasoning, the other concerns such knowledge as it occurs in explicit acts of reasoning. Developing the first part, he claims that when a person explains the cause of her action she is simply applying a general theory about the sort of cause that would explain the sort of action performed. Any self-knowledge gained by the correct self-application of this general theory rests on the same sort of entitlement that is enjoyed when the theory is correctly applied to explain the action of others. The Sceptic cites the research of psychologists like Daryl Bem, Richard Nisbett, Lee Ross, and Timothy Wilson as having demonstrated that despite what we may think, it is normal for us to lack this self-knowledge because we fail to know the cause of our own behaviors.[80] He generalizes this thought to apply also to the cause of our beliefs, saying that it is also normal for the justifications we give for our own beliefs to fail to accord with the proper psychological explanation of those beliefs. If a person is able to identify the cause of a given action or belief reliably, it is only because she is particularly adept at applying general psychological explanations to herself. The Sceptic thus concludes that there is nothing uniquely first-personal about a person's knowledge of her own reasons, at least for knowledge of those reasons that do not occur in explicit, overt acts of reasoning.

To explain a person's psychological and epistemic condition when she is explicitly engaged in reasoning, the Sceptic might say the following. Reasoning takes place in highly sophisticated mental modules. Whether or not the reasoning-module

[80] Canonical works by these authors include Bem 1978, Nisbett and Wilson 1977, and Nisbett and Ross 1980.

functions rationally is a matter that is liable to rational criticism. When a person has knowledge of an act of reasoning, what she is knowledgeable of is the activity of her sophisticated module of rationality. The Sceptic admits that a person is in a special position to know of the functioning of her own rational module. This knowledge, however, is just an instance of a sort of knowledge the Sceptic has allowed for since the outset, *viz*., knowledge of her own occurrent thoughts. Again, the Sceptic might describe the unique epistemic position that provides for this knowledge as giving rise to inner observation or to comprehension of inner testimony. In either case, however, the inner character of this knowledge does not make it of a different epistemic sort than normal observation or normal testimony is of. If a person fails to know the workings of her highly sophisticated rational modules, the failure does not differ epistemically from the failure to know of the workings of another person's mind due to a normal mistake of observation or comprehension.

I do not commit the Sceptic to this particular account of the knowledge one has of one's own reasons; if there is a better account to be had while remaining a Sceptic, then he may have it. I present this particular account only to make it seem plausible that the Sceptic may be sceptical of the rationalistic accounts of self-knowledge as put forward by Burge and Moran. His scepticism against these accounts cannot be sustained, however, once he is forced to consider what it is for someone to understand her reasons for belief or for action. Let us see why this is so.

4.4 UNDERSTANDING AND REFUTING THE SCEPTIC

To understand one's own reasons is to understand why a given reason counts as a reason for believing what one does or for acting as one acts. I will focus exclusively on the

theoretical case, but I intend what I say here to apply to the practical case as well. It needs to be clear at the outset that the sort of understanding that is to be discussed is distinguishable from semantic understanding, for one can understand the meaning of a sentence yet not understand how what the sentence says counts as a reason for some conclusion. For example, I might understand the meaning of 'The leaves on the plant are turning brown' and the meaning of 'The plant is getting too much water' without knowing that a plant's leaves turning brown is an indication that it is getting too much water. If I do not know this, then I do not understand why the plant's leaves turning brown gives me a reason to believe that it is getting too much water. For the remainder of this chapter, the term 'understanding' will be used in the manner of the last sentence and not as it is used when talking of understanding a sentence's or term's meaning.[81]

Understanding is a sort of knowledge. When one understands a subject matter, one is able through the activity of one's own rational faculty to produce and to explain knowledgeable attitudes regarding the subject matter. Suppose that you know that a given kind of plant's leaves turning brown is an indication that it is getting too much water, and suppose you know that my plant is of the relevant kind and that its leaves are turning brown. You, unlike I, can understand why the leaves are turning brown, because you, unlike I, are able to produce a good explanation for why the leaves are turning brown. Understanding comes in degrees: a botanist who understands at a chemical level why overwatering produces brownness in the leaves of a plant of this kind may be able to produce explanations that neither of us can produce regarding the browning of my plant's leaves. The botanist understands what is going on with my plant better than you do, whereas I do not understand at all.

[81] The usage here of 'understand' and its cognates thus differs from that found in Burge 1988 and Burge 1998a.

It is important to note that one can make knowledgeable judgments about some subject matter without having any understanding of the matter. Consider, for example, two logic students, one who understands the relevance of counterexamples to demonstrating the invalidity of arguments, and one who does not. The latter student—call him the dim student—can know that a certain example demonstrates the invalidity of an argument without understanding why. Suppose, for example, that the dim student is taking an exam and that he has stolen the answer key to the exam. When asked to present counterexamples that show arguments to be invalid, he reliably gives correct answers. Moreover, because he knows the answer key to be correct, his answers carry the justification of authority, and so he knows that the answers he gives are correct. The dim student's knowledge differs from that of the former student, whom we shall call the bright student. The bright student understands the relevance of counterexamples to demonstrating the invalidity of arguments and so produces correct answers on the exam as a result of her understanding. Her knowledge is not justified by the authority of an answer key but by her rational ability to produce correct answers. The bright student's understanding distinguishes her both rationally and epistemically from the dim student. She has a rational capability that the dim student lacks, and because of this capability her answers on the logic exam are justified by reason, not by the authority of an answer key.

It might be tempting here to construe the bright student's epistemic superiority as resting on a sort of internal authority that the dim student lacks. On this line of thought, in either case the entitlement to knowledge rests on authority, but in the dim student's case that authority rests in the answer key while in the bright student's case the authority resides in herself. If we only distinguish these sorts of authority in this external/internal manner, however, we leave open the possibility than the bright student's knowledge does not rest on a sort of entitlement that is distinct from the dim student's. To say merely that

the bright student's authority resides in herself is to fail to distinguish the way in which her authority differs from that of an expert chicken-sexer. Purportedly, an expert chicken-sexer can reliably determine the sex of a chicken without knowing to what feature or features of a chicken he attends when determining that chicken's sex. Such a person knows that he is an authority when it comes to determining the sex of a chicken, but he cannot explain what it is about him or his relation to chickens that makes him such an authority.[82]

If we only say of the bright student that she has an internal sort of authority that the dim student lacks, we leave open the possibility that the bright student's authority is like that of the chicken-sexer. If we only say that the bright student's authority is internal, then we leave open the possibility that the student knows that her logic judgments are reliably correct but does not understand why. On this conception of the bright student, she does not understand why any of her answers to the exam are correct, including the answers she gives to questions that ask her to justify her reasoning. When she provides justifications, she reliably gives correct answers, but she knows that they are correct only because she knows that she is an expert on logic. If the bright student were like this, her entitlement to all of her logic judgments would be of a similar sort as the entitlement the chicken-sexer has to his chicken-sexing judgments. If this were so, then the bright student's entitlement would also be of a similar sort as that of the dim student to his answers on the logic exam. The relevant similarity would be this: in all three cases, there would be knowledge that the knower knows to be justified without understanding the justification. But the bright student is not similar to the dim student and to the chicken sexer in this regard. If we only say, then, that the bright student's authority is internal and thus fail to distinguish it from the chicken-sexer's authority, we thereby fail

[82] The example of the chicken-sexer is put to a different use in Brandom 1998. I have been told that chicken-sexers do not exist. Whether or not this is true is of no relevance to the point I am making here.

to distinguish the sort of entitlement the bright student has to her logic judgments from the sort of entitlement the dim student has to his logic judgments.

The bright student is bright because her authority is rational authority, not chicken-sexer authority. The bright student not only knows that her answers are correct; she understands why they are correct. She does not relate to her rational faculty as the dim student relates to his answer key, which is the way the chicken-sexer relates to his capacity for reliably making correct chicken-sexing judgments. The bright student's entitlement rests not merely on the fact that she, or some faculty of hers, successfully produces the correct judgments. Her entitlement rests on the fact that she can produce a line of reasoning and, on the basis of having produced that line of reasoning, correctly draw a conclusion that follows from the line of reasoning. Because of her capacity to do this, she understands her answers and why the justifications she puts forth in defense of them count as justifications. The entitlement she has to her knowledge is distinctly rational, as the entitlement the dim student and the chicken-sexer have to what each knows is not.

The productivity of one's rational faculties not only entitles one to authoritative judgments like the bright student's logic judgments; it also entitles one to self-knowledge. Towards arguing for this claim, let us consider what the Sceptic may say in response to the line of thought pursued thus far, for it is against his responses that the argument will develop. The central claim of this line of thought is that the bright student is entitled to her logic judgments in a way in which the chicken sexer is not to his chicken-sexing judgments. Since this claim will serve as the basis for arguing that self-knowledge can be grounded in a unique first-person entitlement, the Sceptic must take one of two positions with respect to it. His first option is to accept the claim but to insist that doing so does not commit him to accepting that there is a unique first-person

entitlement that grounds self-knowledge. Otherwise, in order to deny that there is a unique first-person entitlement that grounds self-knowledge, he must deny that the bright student has a sort of entitlement to her logic judgments that is distinct from the sort of entitlement the chicken-sexer has to his chicken-sexing judgments.

Suppose the Sceptic chooses the latter of these options and asserts that the sort of entitlement the bright student has to her logic judgments is of the same sort as that which the chicken sexer has to his chicken-sexing judgments. To assert this is to deny the distinction between rational authority and chicken-sexer authority that was drawn a few paragraphs back, for it was by distinguishing these sorts of authority that we distinguished between the two sorts of entitlement. As we have already noted, to assert that the bright student has chicken-sexer authority over her logic judgments is to say that she has this sort of authority over *all* of her logic judgments, including those that she offers as justifications for other of her logic judgments. If the Sceptic believes that this is the sort of authority the bright student in particular can have over logic judgments that she offers as justifications, then he is committed to believing that this is the sort of authority anyone might have in providing any justification for any subject matter. So if the Sceptic asserts that that the sort of entitlement the bright student has to her logic judgments is of the same sort as that which the chicken sexer has to his chicken-sexing judgments, he must hold that the authority on which any authoritative justification rests is chicken-sexer authority.

If this is the case, then a given reason counts as a justification only because the person who has provided the reason counts as an authority on the relevant subject matter. If this is the case, then a putative reason considered without regard for the source of the reason cannot be understood as counting for or against holding some attitude, because it is only in virtue of the source of the reason that it counts as a good or bad reason. On this

conception of justification, reasons themselves do not justify anything. On this conception of justification, the bright student does not count as an authority on logic because she is able on her own to produce good justifications on her exam; rather, her justifications count as good because she is an authority. This is clearly backwards. A line of reasoning does not justify holding an attitude because the source of that reasoning is authoritative; a line of reasoning justifies holding an attitude because it itself is a good, rational line of reasoning. A line of reasoning is a good, rational line of reasoning not because of its source but because it accords with rational norms. If we are to recognize the existence of rational norms as such, then we must accept that when reasons justify holding an attitude, it is because they are good reasons and not because they are issued from an authoritative source. If we are to recognize the existence of rational norms as such, then we cannot accept that the bright student's authority over her logic judgments is of the same sort as that which the chicken sexer has to his chicken-sexing judgments. The Sceptic would be wrong, then, to claim that the sort of entitlement the bright student has to her logic judgments is of the same sort as that which the chicken sexer has to his chicken-sexing judgments.

The Sceptic can deny this, but he can only do so by denying the existence of rational norms as such and by sceptically maintaining that the authority of reason is conventional. If the Sceptic is to deny that the bright student's reasons justify her answers not because she is an authority on logic but because they are good reasons *per se*, he will have to say that there is no such thing as good reasons *per se* and that what we take to be such reasons are merely conventionally determined. The Sceptic can only hold this if he holds that the norms of rationality according to which we judge reasons to be bad or good are themselves merely conventional. To hold this is to be sceptical of the existence of *bona fide* rational norms. I shall not presently argue against this stronger

sort of scepticism, nor do I need to. I have been addressing the Sceptic in order to show that everyday failures of self-knowledge like those that are involved in self-deception are uniquely first-personal sorts of epistemic wrongdoings. If the only way in which the Sceptic can maintain that such failures are not uniquely first-personal sorts of epistemic wrongdoings is to maintain that the authority of reason is merely conventional, then he will have to accept that his own position can only be defended on the basis of conventional authority. Accepting this may simply be self-defeating. Even if it is not, the Sceptic's acceptance of this more radical scepticism provides grounds for simply ignoring him at this point.

If he is not to become a more radical sceptic, then, the Sceptic must hold that while there is some difference of entitlement between the bright student and the chicken-sexer, the existence of this sort of entitlement does not demonstrate anything uniquely first-personal about one's entitlement to self-knowledge. In defense of this position, he might note that so far we have only shown that the entitlement the bright student enjoys to her logic judgments rests on her rational ability to make those judgments, through which she has knowledge of logic. But what, he might ask, does this have to do with *self*-knowledge?

In responding to this question, let us return to an idea that can be found in both Burge's and Moran's accounts, *viz*., that a person has *self*-knowledge in rationally producing her own reasons because *she*, through the well-functioning of *her* rational capacities, is the one who produces them. When we encountered this idea earlier, we saw that neither Burge nor Moran succeeds in exploiting this idea in an argument that is adequate for undermining the Sceptic. Burge claims that uniquely entitled self-knowledge of the attitudes involved in the production of reasons is a necessary condition on the possibility of this rational production, but, as we saw, the Sceptic can complain

that Burge's claim here is no more than a question-begging assertion. Moran similarly claims that the knowledge a person has of the justifications she might rationally produce for her beliefs, desires, and intentions provide her with uniquely first-personal grounds for knowing what her attitudes are; here, we saw the Sceptic can complain that Moran illegitimately draws an epistemological conclusion from premises that only concern how we ought to justify our own attitudes. What is missing from both of their accounts is the recognition that their topic is not just a uniquely first-personal sort of entitlement or ground for knowledge, but a uniquely first-personal sort of knowledge: *understanding* one's own attitudes and reasons.[83] The entitlement or ground is uniquely first-personal because the knowledge itself is of a uniquely first-personal sort. When a person does not just merely know her own attitudes and reasons but understands them, her understanding unites the productivity of her rational capacities with knowledge of this productivity. The Sceptic forces us to realize that we can conceive of the knowledge of one's own rational productivity and that productivity itself as related in a way that is not uniquely first-personal; it is for this reason that the Sceptic is not disarmed by anything Burge or Moran explicitly says. When we conceive of self-knowledge and rational productivity in this way, however, we are not conceiving of the sort of self-knowledge one has when one understands one's own reasons. It is understanding that Burge and Moran mean to be talking about, and once we see that the account of the mind to which the Sceptic is committed cannot describe what it is to understand one's own reasons, we see that Burge and Moran are right to assert that uniquely entitled self-knowledge is a condition on the possibility of the well-functioning of one's own faculty of productive reasoning.

[83] Lear 1988 makes what I think is the same distinction between mere knowledge and understanding. He asserts that according to Aristotle it is part of our rational nature to be driven not just to have mere knowledge, but understanding; cf. pp. 6-7.

With this in mind, let us now see why the understanding the logic student has of her logic judgments involves self-knowledge to which she is entitled in a uniquely first-personal way. When the bright student produces her judgments on the logic exam and justifies them with reasons, the judgments and reasons are both uniquely hers because *she* has rationally produced them. This production is not like that of the chicken-sexer, because the bright student's answers but not the chicken-sexer's judgments are produced along with what we might call a rationale for her answers. The bright student's rationale for any given answer is constituted by the battery of reasons that she would give to defend that answer. The bright student does not need to make this rationale explicit, but the fact that she can make it explicit if asked shows that it is part of what is produced when she produces her answers. Because her production of this rationale is made with understanding, she relates to her rationale in a manner that is distinct from the way in which the dim student relates to the answer key. The bright student understands why her rationale counts as a justification for the judgment she makes on its basis, whereas the dim student cannot grasp why the answers on the answer key are good answers to the exam questions. The understanding she has of her rationale and of her judgments, then, is self-knowledge, for it is an understanding both of *her* beliefs about logic and *her* rationale for holding those beliefs. This self-knowledge rests on an entitlement that is uniquely first-personal, for the rational activity that manifests a given person's understanding can only be produced by that person herself.[84]

I should be clear here that just because a person is entitled to self-knowledge by her understanding, it does not follow that her belief and its rationale satisfy every norm of

[84] To be sure, a person can understand another's beliefs and reasons. When a person understands another, however, the former person's rational capacities do not produce the other's beliefs and reasons but rather produce beliefs *about* the other's beliefs and reasons. In the first-person case, like that of the bright student, both the state of understanding and the beliefs and reasons that are understood are produced by the same rational capacities.

rationality. One may be entitled to self-knowledge by the productive activity of one's rational capacities even though the capacities are defectively productive. For example, a person might have a bad reason for believing something yet not know that her reason is bad. In this case, she is entitled by the productivity of her rational capacities to her knowledge both of her belief and of her reason for holding the belief, but she is rationally criticizable for the badness of her reason. A person might also have inadequate reasons for believing something, know that the reasons are inadequate, yet go on maintaining the unwarranted belief because he believes that giving up the belief would be extremely painful. When this happens, the person is epistemically akratic as described in the last chapter. In such a case, the person knowingly fails to respect the priority of justification over mere psychological explanation in maintaining his belief. Still, he understands why he believes what he does, and because of this understanding he is entitled in a uniquely first-personal way to his knowledge both of his unwarranted belief and of the reasons he holds it.

4.5 SELF-DECEPTION REVISITED

There is still more to say on the role of understanding in mental life. For example, we should want to know the place that understanding has in the accounts of autonomy; I take this up in the next chapter. Presently, however, I want to conclude by returning to where I started, with the topic of self-deception. I claimed at the outset that self-deception seems to involve a uniquely first-personal sort of epistemic failure because the self-deceived individual has a criticizable lack not just of knowledge but of self-knowledge. Recall now what this failure involves. As we saw in the last chapter, in standard cases of self-deception the self-deceived individual denies that the correct explanation of why he

maintains some belief or why he behaves in some way is in fact correct. In such cases, the self-deceived individual lacks self-knowledge because he falsely believes his denial. It is not uncommon for the self-deceived individual instead to maintain a false rationalization of his belief, desire, or intention, which he wrongly takes to be the correct explanation of his thought or deed. In such cases, the self-deceived individual both lacks self-knowledge regarding the reasons for his thought or action, and, worse, has a false belief regarding said reasons.

We can now see how this condition involves a uniquely first-personal sort of epistemic wrongdoing. In order to be fully rational, one's rational capacities must be well-functioning. When one's rational capacities are well-functioning, one understands one's own reasons. This self-understanding is a sort of self-knowledge that rests on a unique first-person entitlement. In failing to have this self-knowledge, the self-deceived individual is criticizable for a uniquely first-personal sort of epistemic failure. Now to be sure, a person may be criticizable for this sort of failure without being in as epistemically bad a condition as the self-deceived individual is in. A person might not know why she believes something, saying, if pressed, "I don't why, but I just believe it;" a person might not know why she has done something, saying, if pressed, "I don't know why, I just did it."[85] Like the self-deceived individual, these individuals also fail to understand themselves, but they know that they are self-ignorant. The self-deceived individual is epistemically worse off, for not only does he not know that he lacks this self-understanding, but he also wrongly believes that he has this self-understanding. When

[85] For a discussion of this sort of self-knowledge of self-ignorance in the practical domain, see sections 17-20 of Anscombe 1963. Anscombe is interested in such cases as intermediaries in which her famous question "Why?" both does and does not have application: "it has application in the sense that it is admitted as an appropriate question; it lacks it in the sense that the answer is that there is no answer" (p. 26). When this is the case, according to Anscombe, a person has acted in a voluntary but non-intentional manner.

we say, then, that the self-deceived individual is at fault for his lack of self-knowledge, we are accusing him of a uniquely first-personal failure of rationality.

This chapter has argued that when a person's rational faculties are well-functioning, her understanding unites the productivity of her rational capacities with her knowledge of this productivity. The discussion here has focused almost exclusively on an epistemic sort of unity. In the next and final chapter I turn to a different but related kind of unity, *viz*., the unity of the self in autonomous action. Once that unity is described, I will be in a position to say how a person is when both her thought and action are rationally unified. That discussion—on the nature of autonomy—will conclude this dissertation.

5.0 SOME REMARKS ON AUTONOMY

So far in this dissertation, I have described the sorts of philosophical projects that are raised by the phenomenon of self-deception, and I have completed two of those projects. I have said what is wrong with the self-deceived individual holding a motivated belief, and I have said what is wrong with the self-deceived individual's lack of self-knowledge. In virtue of these wrongdoings, we might describe the self-deceived individual as lacking, to borrow a term from Harry Frankfurt, wholeheartedness.[86] If the self-deceived individual were wholehearted, then he would have the epistemic courage I discussed in section 3.6, and his beliefs would not be irrationally formed in accordance with what affectively satisfies him. In discussing wholeheartedness, however, Frankfurt intends to talk not only of the rational unity that is present in properly forming and maintaining beliefs; he also means to talk of the rational unity that is present when a person acts autonomously. When a person acts autonomously, there is a unity of knowledge, motivation, evaluation, and action; to act with this unity is to act wholeheartedly. Much contemporary work in practical philosophy has gone into describing how a person is when she acts in this wholehearted way.

The standard way of pursuing this project is to analyze how a person is when she acts in a weak-willed way, for when a person acts in a weak-willed way she lacks the unity in question. The hope is that by understanding the difference between the weak-

[86] See, *e.g.*, Frankfurt 1988d and 1998b. Frankfurt 1998b has this to say explicitly about self-deception and wholeheartedness: "Now someone who is engaged in self-deception in a matter concerning what he is or what he is doing is conceding thereby that he is not satisfied with himself. Like everyone else, of course, he would like to be wholehearted; as all of us do, he wants to love himself. Indeed, this is his motive for self-deception. It is his desire to love himself that leads him to replace an unsatisfying truth about himself, which he cannot wholeheartedly accept, with a belief he can accept without ambivalence" (p. 106). I am largely in agreement with this description of the self-deceived individual whose self-deception concerns only himself (as, *e.g.*, the self-deceived cuckold's does not), but I would clarify that the unsatisfying truth that he cannot accept is *affectively* unsatisfying.

willed person's condition and the autonomous person's condition one will have thereby discovered what is necessary for an action to be autonomously performed. This difference is usually taken to be a difference in the way in which each person relates to the desire that causes her action; the project is then to provide an adequate description of the way in which each person relates to her respective desire. In a recent paper on the topic, Michael Bratman has described this difference in terms of "ownership": the autonomous person "owns" the desire that causes her action, whereas the weak-willed person "disowns" the desire that causes her action.[87] Now in some sense, the weak-willed person clearly owns her desire; the desire in question belongs to her and not to anyone else, for it is the cause of her action, not of anyone else's.[88] Moreover, the weak-willed person knows the role of her desire in causing her action—she is not self-deceived about this. Still, her action is weak-willed and not autonomous because, at a minimum, she thinks it would be better if she were not under the causal influence of the relevant desire. Because she is under the control of the desire, there is a sense in which it is rogue, a desire that she does not own.

Thinking of autonomy and ownership in this way, the project is further refined. Suppose someone thinks a certain desire is not good to act on. She is autonomous if, because of this thought, she does not thereby act on the desire; she is weak-willed if, in spite of this thought, she goes on and acts on the desire. The project then is to explain how the thought can be causally efficacious as it is in the autonomous case and what a person is like when the thought fails to be so efficacious, as in the weak-willed case.

---

[87] Bratman 2003, p. 221.

[88] More accurately, it is not the immediate cause of anyone else's action. If I am a weak-willed drinker who is trying to abstain from drinking, my desire to drink might cause my friend to do various things, *e.g.*, to refrain from offering me an alcoholic beverage. In this case, my desire figures in the causal explanation of why he does not offer me a drink, but the immediate cause of his action is still a desire of *his*, *e.g.*, to help me abstain from drinking.

Once the nature of this thought has been explained, ownership, and in turn autonomy, will be understood.

Frankfurt and Bratman, amongst others, pursue this refined project by describing the sort of thought in question as a higher-order conative attitude. Their theories are thus appropriately called hierarchical theories of autonomy. In the first part of this chapter, I will criticize these hierarchical theories. My goal in doing so will be to show that we should explain autonomy not in terms of the ownership of desires but instead in terms of the ownership of actions themselves. I will go on to show how the attitude in virtue of which an action is autonomous is both evaluative and causally efficacious. Finally, I will close this chapter and this dissertation with some broad remarks on the nature of autonomy.

5.1 HIERARCHICAL ACCOUNTS OF AUTONOMY

Let us start with an example of a weak-willed action. Consider Augustine as he presents himself in Book VIII of *The Confessions*. He has come to believe that his habits of lust are self-destructive because they turn his will away from God, yet he cannot overcome their force. The desires that underpin these habits are in a sense external to Augustine—they seem to be in control of him, against his will. And yet, when he acts on these desires, he does what he does intentionally, even though, as we might put it, a part of him wishes he would do otherwise. Actions that Augustine performs in accordance with these desires are thus not autonomous actions. The actions are the expression of a disunity in Augustine, for the desires that cause his lustful action seem to spring against his will from beyond the bounds of his self.

Frankfurt has attempted to clarify just what this boundary of selfhood is by presenting a variety of related but distinguishable descriptions. He has written of "identification," "volitional necessities," and, as I indicated earlier, "wholeheartedness," all in an effort to distinguish desires that feature in autonomous actions from those that do not.[89] In all of these descriptions, however, Frankfurt maintains that the boundary is to be defined in accordance with a hierarchically conceived motivational structure. Frankfurt's thinking here proceeds along the following lines. When a person acts intentionally yet non-autonomously, she fails to embrace the desire that motivates her action. This metaphor of embracing a desire is to be cashed out in terms of higher-order desires. The person desires that the motivating desire does not motivate her as it does; she does not desire the motivating desire. In contrast, when she acts autonomously, she embraces the desire that motivates her action. In this case the motivating desire is itself one that the person desires. As such, the motivating desire is the object of a higher-order desire, which is part of a whole framework of higher-order desires that concern which first-order desires are desirable to act on and which are undesirable to act on. A first-order desire that fails to be the object of one of these second-order desires is thus in the relevant sense external to the person, while a first-order desire that accords with an appropriate second-order desire is internal to her.

Bratman agrees with Frankfurt that in order to explain autonomy we must describe the ownership of desires in terms of a hierarchical motivational structure. Bratman identifies a quintet of features that are characteristic of hierarchical accounts of autonomy, including both his and Frankfurt's. The first feature is simply the assertion of the existence of a higher-order attitude for every first-order desire that causes autonomous action. The second is a description of what sort of attitude this is: it is

[89] On "identification," see Frankfurt 1988b and 1988d; on "volitional necessities," see Frankfurt 1988c; on "wholeheartedness," see Frankfurt 1988d.

conative. The third feature is a description of the attitude's temporal orientation. It is "forward-looking," in the sense that the content of the attitude pertains to the future. The way in which the attitude pertains to the future is directive, guiding the activity of the first-order desire that is its object. Accordingly, the fourth feature is a description of the higher-order attitude's function, which is to guide the relevant first-order desire's activity. The fifth feature, Bratman says, is the assertion that the attitude "is to constitute—at least in part, and given relevant background conditions—a commitment on the part of the agent concerning the role of the target desire in her own agency: the agent is appropriately settled on this."[90] A theory of agency is thus hierarchical just in case it describes the first-order desires on which a person autonomously acts as objects of forward-looking, guiding, higher-order conative attitudes that in some sense constitute commitments on the part of the agent.

5.2 SOME CHALLENGES TO HIERARCHICAL ACCOUNTS

The fifth feature, the assertion that higher-order attitudes constitute practical commitments of the agent, is meant to respond to a challenge Gary Watson has posed to hierarchical theories. Watson points out that the fact that a lower-order desire is the object of some higher-order desire does not by itself make it the case that the action that results from the lower-order desire is autonomous. A higher-order desire is still a higher-order *desire*, and, as such, may not be one that the person owns. Watson has described higher-order desires that do not produce autonomy as "wanton" and has argued that any first-order desire that is the object of a wanton second-order desire is not one that can

[90] Bratman 2003, p. 224.

feature in autonomous action.[91] The hierarchical theorist's intuition is that we have some first-order desires that we second-order desire not to have, and so only the first-order desires that are themselves objects of a second-order desire are internal to the self. Watson points out that without further specification, this model of autonomy fails to distinguish between second-order desires that the subject finds desirable and second-order desires that the subject does not find desirable. If a second-order desire is itself not one that the subject finds desirable, *i.e.*, if it is wanton, then the first-order desire that is its object is not internal to the self, and action in accordance with this desire is not autonomous action. Just being a second-order desire by itself does not suffice to make it a desire internal to the self, then—the second-order desire needs to be held non-wantonly.

The challenge posed by Watson's explanatory demand might lead one to consider whether the hierarchical theorist's approach to autonomy is fundamentally wrong. One might wonder if there is not a far simpler account of autonomy to be given that does not face the problem of wantonness. Bratman imagines such an account, which he describes as a "Platonic challenge" to hierarchical accounts.[92] A person who holds this simpler account believes that actions are autonomous just in case they are consistent with the evaluative judgments of the acting person. According to this account, an action is autonomous as long as the agent does not judge that performing the action is out of line with his reasons for action. If the agent does judge that performing the action is out of line with his reasons, then should he perform the action, he does so non-autonomously. In this case, the desire that causes the action is external to the agent. This account raises no worries of wantonness. A person owns a desire just in case the action that desire might cause is evaluated by her as good; otherwise, she does not own the desire. The

[91] See Watson 1975.
[92] This challenge is also addressed in Watson 1975. Bratman credits Watson for this.

person who holds this "Platonic" account of autonomy challenges the hierarchical theorist to show why the latter's more complex account is needed.

In response to this challenge, Bratman provides an extended argument whose goal is to show that evaluative judgments are not sufficient for autonomy. Bratman thinks that in order to argue decisively against the Platonic challenge he must show that autonomy is not necessarily based on a comparative evaluative judgment. In order to establish that autonomy is not necessarily based on a comparative evaluative judgment, Bratman presents a case in which a person cannot make such a comparative judgment because he finds two competing courses of action to be equally worthwhile. His example is the decision whether or not to enlist in the military as considered by someone who sees both enlisting and not enlisting as equally good choices. Since, as Bratman says, "life must go on"—or, as Jean-Paul Sartre says, we are "condemned to be free"—this person must commit one way or the other.[93] Since the person judges that either course of action would be equally good, whatever commitment he adopts cannot be made on the basis of a judgment that one is better than the other. The conclusion Bratman draws is that "the agent's value judgments by themselves underdetermine his stance in response to the practical issues raised about how he is to live."[94]

Bratman's example here does not establish what he wants it to establish. It establishes that there are cases in which there is no single course of action that is judged to be better than every relevant alternative. In those cases there will be no comparative judgment about the single best course of action, so whatever action is taken, it cannot be

[93] Bratman 2003, p. 230; Sartre 1956, p. 567. In this section, Sartre goes on to say of this condemnation, "This means that no limits to my freedom can be found except freedom itself or, if you prefer, that we are not free to cease being free." Bratman's example here is of a sort that can help us get a grip on Sartre's often baffling and hyperbolic language.

[94] Bratman 2003, p. 231. In a footnote, Bratman cites Hampshire 1977, Wiggins 1976, Wolf 2002, Holton 1999, Nozick 1981, and Lehrer 1997 as allies in favor of the idea that evaluative judgments underdetermine practical commitments. To this list one can add Moran 2001, esp. sections 3.4 and 4.7. While I disagree with the specific point Bratman makes, it should be clear that I agree that the sort of evaluative judgment Bratman has in mind is to be distinguished from one's practical commitments.

taken because it is the best of all relevant alternatives. In any such case, however, there will be some competing courses of action judged to be equally good, but there will also be other courses of action that the person judges to be worse than the former alternatives. It is open to the Platonic challenger to say that this is all he needs to be the case in order to defend his account of autonomy. As long as for every action there are alternatives, some better, some worse, an action is autonomous just in case the person does not take an alternative that he judges to be worse than any other alternative. In Bratman's example, a failure to enlist in the military will be autonomous only if what the person does instead of enlisting is judged to be no worse than enlisting.

Indeed, I think it is right for the Platonic challenger to insist on this, for a comparative evaluative attitude of the sort just mentioned is necessary for an act to be autonomous. We will return to this point towards the end of this chapter. For now let us set it aside, for there is a more fundamental criticism of hierarchical accounts invited by Bratman's response to the Platonic challenge. The easiest way to present this criticism is by responding to another of Bratman's examples, to which we now turn.

5.3 THE MISTAKE OF THE HIERARCHICAL STRATEGY

One of the central examples in Bratman's extended argument compares a pair of individuals who both make the same evaluative judgment but who do not both thereby commit to act in accordance with the judgment. In the example, each of two people routinely drink alcohol, and each judges that it would be beneficial to abstain from drinking alcohol. Only one of these people goes on to commit to a life of abstinence, however, so only one rejects the desire to drink when it arises. Bratman says that such a rejection "consists in a higher-order, conative attitude. . . that partly constitutes a

commitment not to build that desire and its targeted activity into [the abstainer's] life."[95] Such an attitude is lacking in the person who, despite judging a life of abstinence to be good, does not abstain from drinking. This difference, according to Bratman, supports the conclusion that evaluative judgments by themselves do not "fully constitute the basis of ownership and rejection."[96]

I want to draw attention to what according to hierarchical theories is purportedly rejected by the abstainer but not by the drinker. The abstainer is autonomous, these theories claim, because he rejects his desire to drink alcohol and thereby does not act on the rejected desire. This is not the only way, however, of describing what the abstainer rejects. Instead of saying he rejects his *desire* to drink, we might say he rejects the *act* of drinking. On this alternative conception, for which I shall now argue, it is only because an action is owned or rejected that any desire that might cause it is owned or rejected. The difference in wording is not trivial. In asserting that it is primarily the action and not the relevant desire that the autonomous agent owns, I am also asserting that in order to explain what it means for a desire to be owned we must first explain what it means for an action to be owned. The hierarchical theorist does not agree with this order of explanation. His whole goal in developing a hierarchical account of motivation is to use that account to explain the ownership of action. Having raised the possibility of an alternative order of explanation, we can present him with yet another challenge, which may be put as follows: why explain the ownership of action in terms of the ownership of desires instead of vice versa?

If we reflect on the way in which the hierarchical theorist approaches the topic of autonomy, we can see what his answer here must be. The hierarchical theorist thinks that we have an unproblematic grip on the idea that an action is caused by an appropriate

[95] Bratman 2003, p. 228.
[96] Bratman 2003, p. 229.

belief-desire pair combining in an actor's mind. Given that we have a grip on this basic account of how actions are caused, the only problem we face in distinguishing between autonomous and non-autonomous actions is figuring out how to distinguish between the different sorts of desires that cause each. What we need to do, then, is add something to the basic account that allows us to draw the needed distinction. Now the way to add something to the account, the hierarchical theorist thinks, is to imagine taking a being that has only first-order beliefs and desires, which as such is presumably capable of non-autonomous action, and then adding to it a structure of states or capacities that would make it capable of performing autonomous actions. Sometimes this strategy is pursued without being considered as a strategy; other times it is explicitly recognized as such, often under the head of "creature construction."[97]

Once this strategy is adopted, the decision to explain autonomous action in terms of something like a higher-order desire comes naturally, for the reasons that motivate Frankfurt's initial account. The weak-willed drinker is under the control of a desire to drink, but he does not desire that the desire so control him. Were he not weak-willed, his desire to withstand the desire to drink would be fulfilled; that is, if he were autonomous, he would successfully act on his desire to withstand the desire to drink. It thus appears to be fundamental for the very possibility of acting autonomously that one have a higher-order desire of this sort. So, the hierarchical theorist concludes, something like a structure of higher-order desires must be the thing to add to the simple creature to make it into an autonomous creature.[98]

---

[97] The source of the title "creature construction"—and no doubt a source of this strategy—is Grice 1974-75. Bratman 2000a, Bratman 2000b, and Velleman 2001b all explicitly follow Grice in pursuing the strategy.

[98] It is worth noting here that while Velleman 2001b engages in this method of creature construction, he argues that what needs to be added is not a hierarchical motivational structure but the constitutive aim of self-knowledge. I shall not develop the following complaint here, but I think Velleman's account of self-knowledge is one that the Sceptic of the last chapter can easily accept and thus is subject to the criticisms of that chapter.

Whatever explanatory advantage this method might seem to provide is lost, however, once the hierarchical theorist is asked to describe precisely what these higher-order attitudes are. Whatever they are, these attitudes that are the centerpiece of hierarchical accounts cannot be full-fledged desires. First-order desires can be and characteristically are satisfied through intentional actions that aim non-instrumentally at bringing about a desired state or event. Second-order desire-like attitudes cannot be satisfied in this way, because one cannot non-instrumentally bring about the state or event specified by the content of such a second-order attitude. Because a second-order desire-like attitude cannot be satisfied in the way that is characteristic of a first-order desire, the former sort of attitude cannot be a full-fledged desire.

To see this, suppose I find myself desiring a drink, and suppose I want to refrain from acting on that desire. On the belief-desire model of action to which all hierarchical theories subscribe, what I need if I am to keep from acting on my desire to drink is that some other desire cause me to act in a way that does not involve drinking. Suppose that I successfully refrain from drinking. How might my higher-order desire-like attitude have affected the desire that causes my teetotal action? It cannot bring the first-order desire into existence in the way that desires normally bring things into existence, for normally a desire brings something into existence because of the intentional activity of the person to whom the desire belongs. To bring a first-order desire intentionally into existence in this way would be to form a desire at will, which is no more possible than bringing a belief into existence at will.[99] My higher-order desire-like attitude also cannot either strengthen the desire that causes my action or weaken my desire to drink. The idea of

[99] I assert that one cannot believe or desire at will without argument. There is a growing literature on the topic of believing at will, the majority of which aims to show why one cannot believe at will. It is presumed by most, however, that one cannot believe at will. (An exception is Velleman and Shah 2005.) I will help myself to this presumption without showing why it is true—that is a task for another occasion.

intentionally strengthening or weakening a desire is just as obscure as intentionally bringing a desire into existence.

The hierarchical theorist can avoid these problems if he claims that the immediate effect of my higher-order attitude is to focus my attention. The following bit of practical reasoning is at least *prima facie* intelligible: I do not want to act on my desire to drink; if I do not focus on some other desire, I will act on my desire to drink; if I focus on my desire to act admirably, I will not act on my desire to drink; so, I will focus on my desire to act admirably. The problem here is that the higher-order part of the account is doing no work. My reasoning is in no significant way different if it proceeds as follows: I do not want to drink; if I do not focus on doing something else, I will drink; if I focus on acting admirably, I will not drink; so, I will focus on acting admirably. If the causal role to be ascribed to these higher-order attitudes is attention-focusing, then there is no need for the attitudes to be higher-order at all.

This shows, I think, that any higher-order attitude of the sort discussed by hierarchical theorists cannot play the causal role that full-fledged desires play in a person's life. Because its causal role cannot be that of a normal desire, the higher-order attitude cannot be a normal desire. This fact ruins the account of any hierarchical theorist who holds, as Bratman does, that the relevant higher-order attitude "is itself just one more desire in, as we might say, the psychic stew."[100] We can imagine, though, that some other hierarchical theorist will respond to this charge by saying that higher-order conative attitudes are not like normal first-order desires precisely because they are higher-order. If he says this, however, then he is committed to making autonomy out to be either a sort of wish fulfillment, if the higher-order attitude plays no causal role, or, if the attitude does play some causal role, a condition that besets the person. Consider the first option.

[100] Bratman 2000a, p. 37.

Wishes are desire-like attitudes that lack the causal efficacy of desires. If the relevant higher-order conative attitude plays no causal role, then an action is autonomous just in case the desires that the agent wishes would cause the action do in fact cause the action. This is not what the hierarchical theorist wants. What he wants is that the higher-order attitudes are causally efficacious but are not the sort of attitudes on which a person can intentionally act. If they cannot be acted on intentionally, however, then their effects are something that happens to the person. When an action is autonomous, however, it is not because something has happened to the person any more than it is because a wish of that person has been fulfilled. Neither of these options will do as accounts of autonomy.

By challenging the hierarchical theorist's explanatory strategy, then, we see that no account of autonomy in terms of the sort of attitudes that are the centerpiece of hierarchical accounts can succeed. The correct response to this is to quit pursuing the hierarchical theorist's explanatory strategy. Abandoning this strategy opens us up to the possibility of describing autonomy in terms of some positive attitude the actor has towards his action, not towards the desire that causes his action. I now want to pursue this possibility. Drawing on the work of the previous section, I claim that the attitude in virtue of which an action is autonomous is one that evaluates the action as no worse than any competing alternative. If this is all that is said about the attitude, however, the way in which the attitude is causally efficacious is not yet evident. As the hierarchical theorist is right to emphasize, the attitude plays some causal role in autonomous action. We must be careful in describing this role, however, for as the hierarchical theorist also rightly emphasizes, the attitude can fail to be causally efficacious, as happens when a person is weak-willed. I believe that the way to understand the causal features of this attitude is by first explicating some of the relevant parts of G. E. M. Anscombe's account of practical rationality. So to that.

5.4 ANSCOMBE AND PRACTICAL RATIONALITY

If we adopt an Anscombean approach to the various topics of this chapter, we are not inclined to explain weakness of will and autonomy in terms of internal and external desires. Indeed, we are not inclined to address matters of practical philosophy in terms of psychological states of desire at all. One of Anscombe's major goals in *Intention* is to turn analytic philosophy away from trying to explain intention and action in terms of psychological states and toward understanding these practical phenomena in terms of reasoning and rational explanation. This reorientation is meant to take us from thinking of intention and action in terms of brute psychic forces battling within us to thinking about these topics in terms of the reasons we can and do give for what we do. If we follow Anscombe, the starting point for an account of autonomy will not be a characterization of a class of desires as internal to the self; rather, it will be a description of a good act of practical rationality, which results in autonomous action.

Anscombe insists that if we are to understand wherein an act of rationality can be distinctly practical, we must accept that such an act concludes not with some psychological state but with an action itself. This is at least part of what she has in mind when she claims that practical rationality has a logical form that differs from that of theoretical rationality. When a person performs an act of theoretical rationality, her goal is to know what is the case, and as a result of her act she forms a belief that in the good case is both true and justified. When a person is reasoning practically, in contrast, she is not discovering what to believe but deciding what to do. If we attempt to comprehend what it is to make a decision narrowly on the model of theoretical reasoning, we are likely to think that the result of a decision is a psychological state, perhaps a belief about

what one ought to do or a self-prediction concerning what one will do.[101] If we conceive of decisions in this way, however, we will miss what is distinctly practical about practical reasoning.

Anscombe's argument for this position is no shorter than the entire length of *Intention*, which is more than can be dealt with at present.[102] This need not, however, prevent us from explicating those parts of Anscombe's account that will help us understand the causal efficacy of the evaluation that accompanies autonomous action. Towards the end of *Intention*, Anscombe notes that it is rare for a person to reason explicitly from a practical first premise to a practical conclusion, so no good account of practical rationality would assert that a person must explicitly go through such reasoning in order to act intentionally or, for that matter, autonomously.[103] The rational order that such reasoning proceeds along, however, is present in any action and can be discovered by asking him Anscombe's famous question "Why?". [104] When we ask such a question, we discover the chain of practical reasons that explain the person's action, which can be reconstructed into a line of practical reasoning whose result is the person's action.

To see this, consider a version of one of Anscombe's examples. Suppose we meet a farmer in the Hereford market and ask him why he is there. We can imagine him telling us, "Because they have good Jersey cows here." He thus presents us with a reason for being in the market, a reason that is part of his chain of reasons for being there. This

---

[101] It appears that Anscombe thinks various moral theorists of her time hold the former view. For an example of someone who holds the latter view, see Velleman 2001b and 2001c.

[102] Moran makes a similar claim about Anscombe's argument for the non-observationality of self-knowledge in action; see Moran 2001, p. 126.

[103] Anscombe 1963, section 42, p. 79.

[104] Asking someone why they are doing something will result in an explanation of their reasons for action only if certain conditions are met. First, we cannot discover a person's reasons if she is self-deceived. Second, there are cases of bodily movement in which no answer the person gives will provide a reason-explanation, as when, *e.g.*, a person is unaware of her bodily movements, or she can only know these movements by observation, or the only way to explain the movements is by citing some fact of neurobiology (as one might, *e.g.*, explain why one has twitched just before falling asleep). See Anscombe 1963, sections 5-17, pp. 9-26, on these various cases.

might be a reason we do not understand, perhaps because we do not know that he is a farmer. If so, we can repeat the question-form, this time asking him what he wants with a Jersey. If he says that it suits his needs, he thereby expresses what, according to Anscombe, functions as the first premise in his chain of reasons. Should we again repeat the question-form, asking why it would suit his needs, we will not learn anything more about his reasons; instead, we learn something about *him*, presumably, that he is a farmer with such-and-such a farm. What we learn up to that point, however, are his reasons for doing what he does. It makes no difference to their status as his reasons whether he considers these reasons through an overt act of reasoning performed before going to the market or whether he explicitly considers them for the first time only after we query him.

The reasons in the example rationalize the act of getting a cow, not the formation of a psychological state that, should the farmer do something else, would contribute to his getting a cow. If this is not obvious, consider the case in which the farmer engages in overt reasoning before heading out to the market. Given that for the farmer a Jersey cow is desirable, *i.e.*, is something it would be good for him to have, it is clearly more rational for him as a result of his reasoning to set about getting the cow rather than just to form a belief about the benefits of such activity. It is not a meiosis to say this, for putting the point this way allows us to understand why a good act of practical reasoning concludes with an action. If we do not acknowledge that good acts of practical reasoning conclude with actions, we will then need to explain what else the farmer needs to do once his reasoning is done to bring himself to go after a cow. Whatever this extra act is, it cannot be rational if it is to stand between the conclusion of reasoning and the action of getting a cow. If the extra act were rational, then it would be part of a rational chain that starts with the farmer's reflections on cows and ends with him getting a cow. If it is part of this chain, then it is internal to the act of reasoning, and thus its effect—getting a cow—is the

conclusion of the act of reasoning. If the extra act is truly extra, then it must not be part of this chain of reasoning. In this case, the extra act is not rational. If it is not rational, though, it makes a mystery of the obvious rational connection between an act of practical reasoning and the action that results from that reasoning.

The reader not convinced of Anscombe's position may not find this argument decisive, but it requires him to tell a story that intelligibly connects whatever he thinks concludes an act of practical reasoning with the action that that reasoning rationalizes. The Anscombean does not see how this can be done. If the Anscombean is right on this count, then it is wrong to think that desires, considered as psychological states, figure properly in action-explanations. If a given action is to be explained by the reasons the person might give for them, then it is ultimately for the sake of doing or achieving something desirable that the person acts. Consider the farmer again. It is not because of the presence of some occurrent psychological state that it is rational for him to set off to the Hereford market. Rather, it is because a Jersey cow is useful to him that it is rational for him to try to get one. To say this is not to deny that the farmer may undergo various psychological states during the process of getting the cow: he may be excited to get one, anxious that they may all be sold out, and then relieved when he finally purchases one and turns to head back home. The presence of these various affective states, however, does not explain why it is rational for him to get the cow, for the action will still be rational even if he does not undergo any of these states. A given bit of behavior counts as an intentional action, then, not because of the presence of some psychological state but because a certain form of explanation applies to that activity. Anscombe puts this point by saying, "In fact the term 'intentional' has reference to a *form* of description of events.

What is essential to this form is displayed by the results of our enquiries into the question 'Why?'".[105]

With Anscombe's account in place, we can now say how the sort of evaluation that is present in autonomous action can be motivationally efficacious. Consider the case of the autonomous abstainer refusing a glass of wine that has been offered to him. If asked, he would agree that his choice to refuse the glass of wine no worse than any relevant alternative; indeed, if asked, he would agree that his choice is better than the relevant alternative, which here is to accept the glass of wine. This evaluation is part of the chain of reasons he might give if asked Anscombe's question "Why?" about his refusal. Even if he never explicitly considers the evaluation, it is part of the reasoning that explains his action, and so it is part of what causes his action. Just as the desirability of a Jersey cow rationally causes the farmer to set out towards the Hereford market, the desirability of abstaining rationally causes the abstainer to refuse the glass of wine. To properly understand the causal role of the evaluation, then, we should not consider it as some psychological state in the abstainer's mind. Rather, we should focus on its place in the abstainer's assessment of the desirability of the goal of his decision.

To fully account for the nature of this evaluation, however, we must not only make clear how it can be causally efficacious but also how it can fail to be so efficacious, as in the case of weak-willed action. Before pursuing this task we should recall something that was stated at the outset: the weak-willed person knows what he does and why he does it. He is not self-deceived. The failed abstainer, just like the successful abstainer, can answer Anscombe's question "Why?" if it is addressed to his decision to drink a glass of wine. There is no problem, then, in conceiving of what the failed abstainer does as action. If there is a problem, it can only be that our description of the

[105] Anscombe 1963, section 47, p. 84.

evaluation that is present in autonomous action renders failure to act in accordance with it mysterious. I want now to diffuse whatever mysteriousness might be lurking here.

First, note that if there is a mystery here, it will be one that we face in attempting to understand not only weak-willed action but also continent action.[106] We should find weakness of will mysterious only if we are inclined to think that there is some immediate or necessary connection between evaluating one course of action as better than another and thereby deciding to do the former rather than the latter. If this inclination makes us find weakness of will mysterious, however, we should find continent action just as mysterious. When a person acts continently, she evaluates some course of action to be what she has reason to do, but then there is a temporal gap between her making this evaluation and her adopting the intention to act accordingly. The continent person, just like the weak-willed person, thus judges that something is to be done but does not thereby immediately resolve to perform the action. This being so, we may, when it suits our purposes, switch our focus from weakness of will to continent action in explaining how the evaluation we are investigating can fail to cause action in accordance with it.

Depending on what else one thinks about the distinction between theoretical and practical cognition, one might go any of several ways in giving the needed explanation. If one thinks that a practical judgment, being practical, must immediately result in either an intention to act or the action itself, then the evaluation in question cannot be a practical judgment, since both the weak-willed and the continent individual evaluate a course of action as to be done but do not immediately therein form the intention to perform the action. If this is one's view, then the evaluation is understood to be a theoretical judgment about the goodness of the relevant action. When a person is weak-willed, this

[106] In making this point I have drawn heavily on John McDowell's description of the issue that weakness of will and continence both raise in Aristotle's account of practical rationality. See McDowell 1998b (esp. p. 47) and McDowell 1998c (esp. p. 55).

theoretical judgment fails to correspond with the person's actions; when a person is continent, this theoretical judgment is made before the corresponding practical judgment and action. When a person simply does what she thinks is best, this theoretical judgment simply accompanies the relevant practical judgment. The evaluation is still causally efficacious if it is part of the practical reasoning the person would give for her action, but it need not be so efficacious because, considered merely as an evaluation, it is a theoretical judgment.[107]

Alternatively, if one thinks that the evaluation must be a practical judgment because the reasons that back it are practical, one might describe the judgment as one that leaves a chain of practical reasoning incomplete. On this view, a chain of practical reasoning, being practical, is necessarily completed by an action. If a person evaluates an action as right to perform but, being continent, does not therein form the intention to perform the action, then according to this view the person is in the midst of a chain of reasons whose conclusion she recognizes but does not yet draw. Because it is a recognition based on a chain of practical reasons, it is a practical judgment, but it is not the practical judgment that completes the chain of reasoning, for that judgment results in an intention to act. According to this view, the good, non-continent case is distinguished from the continent case because in the former but not the latter the judgment that completes the chain of reasoning is the only one wherein what is to be concluded is recognized. Weakness of will is then understood as involving a practical judgment regarding what to do that leaves the relevant chain of reasoning forever incomplete. On

[107] As I understand it, this is Pamela Hieronymi's view of weakness of will.

this view, the evaluation is causally efficacious only if it is part of a chain of practical reasoning that is eventually completed.[108]

One might complain that while the first view is wrong to conceive of the evaluation as a theoretical judgment, the second view is wrong to think that there are practical judgments that are not involved in any action. In light of these complaints, one might offer this third view of the evaluation of goodness present in continent and weak-willed action. According to this view, when the weak-willed abstainer drinks, he is at once drinking and doing something else that is in tension with his drinking. In this particular example, the other thing he is doing is trying to live a healthy life. To see this, assume that the day after this man has drunk his wine he reflects on his action and judges once again that he should, for the sake of his health, abstain from drinking. If he is being sincere, then both before and after his weak-willed hours of drinking he is engaged in the activity of trying to live a healthy life. According to the present view, the right way to think of being engaged in such an ongoing activity is as something that one is *constantly* doing. A person is engaged in the activity of living healthily both when his bodily movements are bringing about health—*e.g.*, when he is exercising—and when they are not—*e.g.*, when he is drinking wine. His evaluation that he should not drink wine is thus practical because it is part of something he *is* doing—trying to live a healthy life—even as he drinks the wine.[109]

Having considered these various views, it is difficult to see why the occasional failure of the evaluation to be causally efficacious might seem mysterious. If one is comfortable with conceiving of the evaluation as a theoretical judgment, then there is clearly no mystery. If one insists that the evaluation is a practical judgment, then the

[108] It is possible that the difference between the two views just mentioned is merely verbal. I have chosen to discuss both, however, just in case the difference is substantive. I shall not at present attempt to sort out whether the dispute is substantive or verbal.
[109] As I understand his account of weakness of will, Sebastian Rödl holds something like this view.

mystery is eliminated if one accepts either of the two views just presented. If one's commitments do not allow one to accept any of these views but still leave the evaluation's occasional lack of causal efficacy a mystery, then it seems right to suppose that there must be something wrong with one's commitments.

5.5 UNITY AND AUTONOMY

When a person owns a desire, then, it is because she owns the action that is caused by it. To own an action is to evaluate it as no worse to perform than any of the relevant alternatives. If a person performs an action that she evaluates to be worse than some relevant alternative, the action and desire that causes it are ones that the person disowns. Because she knows what she is pursuing and why she is pursuing it, her activity is intentional. It is not fully hers, however, because she recognizes that she has better reason to pursue an alternative course of action.

Having said this, we can now make some general remarks on autonomy. Depending on how one conceives of autonomy, continent actions, which were discussed in the last section, may or may not count as autonomous. Given the way they talk about autonomy, it appears that Frankfurt and those who pursue the topic in his manner believe that continent actions are autonomous actions. On their line, it appears that as long as I act on a desire that is internal in the relevant sense, my action is autonomous—it does not matter whether I immediately act on the desire when circumstances are appropriate or whether I must force myself to act on the desire. In either case, I have exerted some self-control, and if autonomous action only requires some self-control, then continent actions will be autonomous. A virtue ethicist may disagree here, claiming that if circumstances are appropriate for acting on a practical commitment and the person does not thereby

immediately act then the action is not autonomously performed. Either way, it seems clear both that either case involves a kind of self-control that is lacking in weakness of will but that there is a difference between forcing oneself to do what one thinks one should do and simply doing what one thinks is to be done. Whether cases of continence do or do not merit the label 'autonomy' appears to be little more than a matter of nomenclature.

In determining whether or not an action is autonomous, we must also assess whether the acting person has made a mistake in judging what he has best reason to do. There are several ways in which a person could make such a mistake. One way, which is a source of the debate surrounding "internal" and "external" reasons, is to reason explicitly from a belief that, while not completely unwarranted, is false. If I think that a glass of petrol is a glass of gin and think that drinking gin would be pleasant, I may try to get the glass of petrol in order to drink it, which is clearly not a reasonable thing to do.[110] Given that it is not reasonable, one may want to say that I have no reason to pursue taking a gulp from the glass. Even if this is the thing to say, it should be clear that were there gin in the glass, my action would be reasonable, and one can easily imagine circumstances in which the belief I form when I judge that the substance in the glass is gin (or is at least potable) is warranted. Deciding to drink what is in the glass, then, does not seem to be a failure of autonomy, especially if, as is likely, I have no reason to expect that the glass is filled with petrol.

This case should be contrasted with the case in which a person does something unreasonable but the lack of reasonability is not the result of a warranted false belief. Suppose, for example, that I decide to vote for a political candidate because his manner of speaking is more like mine and my neighbors' than is the manner of speaking of the

[110] See Williams 1981b for the example.

competing candidates. I may know that the way in which one speaks is not a reliable indicator of the choices they will make. I may think this is particularly true of politicians, and this may cause me generally to distrust the way in which they speak. I may know that the candidate is of a socio-economic class far more wealthy than mine and my neighbors', and I may believe that politicians tend to make decisions to benefit their own class. Still, I may not take these various considerations as grounds for voting against this particular candidate and instead vote for him. My choice of whom to vote for, it seems clear, has not been reasonably made. But is my choice thereby non-autonomous?

To answer this question we need to say more about my attitude towards my reason for voting for the candidate. I may think that, in this particular case, my preference for the way in which the candidate speaks provides me with adequate reason to dismiss my standing distrust of the way in which politicians talk. I thus understand fully why I have chosen to ignore this standing distrust, and I think I have better reason to vote for this candidate than for any other. If this is so, then while I may be criticized for being less than fully rational in making my decision, I cannot be criticized for having non-autonomously made my decision. I fully understand why I am doing what I am doing, and I do not think there is some alternative course of action that I should but do not pursue.

On the other hand, I may not fully understand why this candidate inclines me to disregard my standing distrust of the way in which politicians speak. Let us suppose that the candidate's speeches have resulted in my fearing anyone but him being elected but that I do not recognize this fear as such—all I know is that the thought of having him elected gives me a better feeling than does the thought of anyone else being elected. If this is how I am, I lack the self-understanding that was the topic of the last chapter. When I go to cast my vote for the candidate, I understand myself well-enough to know

that I am voting for him because of the feeling he gives me, but I do not understand myself well-enough to know *why* I have this feeling. To use the language of the chapter 3.0, I know that I am voting for him because I find it affectively satisfying to do so, but I do not know that the affective dissatisfaction I find at the thought of voting for anyone else is a fear that the candidate's speeches have caused me to associate with that thought. Let us suppose that the fear is unwarranted and that, were I to reflect on it, I would realize it to be so. In this case, the lack of understanding I have regarding my lack of rationality leaves me criticizable both for being less than fully rational and for making a non-autonomous decision. My decision in this case is non-autonomous because in failing to understand my reasons, I leave myself unable to adjust my thought or action through rational reflection on those reasons.

In contrast to all of these cases is the case in which I fail to understand my reasons because I am self-deceived. If I am driven by fear to vote for this candidate but, because I find it affectively dissatisfying to think that I am so motivated by fear, explicitly deny that this is so, then I am self-deceived. In this case too my action cannot be autonomous. If I cannot understand my reasons because I find it affectively unsatisfying to think of them as my reasons, then much as in the last case I cannot adjust my thought or action through rational assessment. Self-deception, however, renders a person less autonomous than she is when she simply fails to have a full understanding of her reasons for action. If there is a simple failure of full understanding, a person does not need to work against her own motives in order to bring the order of reason to her thought and action. When, by contrast, a person is self-deceived, her own motives prevent her from bringing about this rational order. In the worst case, the person's self-deceived condition causes her to do something that, were she to recognize the end being satisfied by her deed, she would deem to be worse than some relevant alternative. This is how I am if, *e.g.*, I am self-

deceived about the fact that my voting decision is guided by fear and, were I to accept this, I would find my decision to vote for this candidate worse than some relevant alternative.

To end up in such a condition is a rational calamity. But it has not been my goal here to spell out all that has gone wrong when a person is so self-divided. My aim, rather, has been to investigate the nature of the self-deceived individual's failures in order to learn in what the unity of the self consists and why it is of importance to us as humans. If I have been successful, it will be clear that the unity in question is the unity of rationality, which is of importance to us because we are rational beings. This unity, which is present with us when we are at our best, is a condition on us being who we are in the first place.

BIBLIOGRAPHY

Anscombe, G. E. M. (1963). *Intention*. (Cambridge, MA: Harvard UP)

Augustine (1991). *Confessions*. (H. Chadwick (Trans.). Oxford: Oxford UP)

Bach, K. (1981). An Analysis of Self-Deception. *Philosophy and Phenomenological Research, 41:3*, 351-370

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

Barnes, A. (1997). *Seeing Through Self-Deception*. (Cambridge: Cambridge UP)

Bem, D. (1978). Self-Perception Theory. (In Berkowitz 1978 (pp. 221-282).)

Bennett, J. (1990). Why is Belief Involuntary? *Analysis*, *50*, 87-107

Berkowitz, L. (1978). *Cognitive Theories in Social Psychology*. (New York: Academic Press)

Bilgrami, A. (2006). *Self-Knowledge and Resentment*. (Cambridge, MA: Harvard UP)

Brandom, R. (1998). Insights and Blindspots of Reliabilism. *The Monist, 81*, 371-92

Bratman, M. (2003). A Desire of One's Own. *Journal of Philosophy, 100*, 5, 221-242

Bratman, M. (2000a). Reflection, Planning, and Temporally Extended Agency. *Philosophical Review, 109*, 1, 35-61

Bratman, M. (2000b). Valuing and the Will. *Philosophical Perspectives, 14*, 249-65

Burge, T. (1998a). Our Entitlement to Self-Knowledge. (In Ludlow and Martin 1998 (pp. 239-63).)

Burge, T. (1998b). Reason and the First Person. (In Smith, Wright, and MacDonald 1998 (pp. 243-270).)

Burge, T. (1988). Individualism and Self-Knowledge. *Journal of Philosophy, 85:11*, 649-663

Buss, S. and Overton, L. (Eds.) (2002). *Contours of Agency: Essays on Themes from Harry Frankfurt*. (Cambridge, MA: MIT Press)

Cohen, J. and McLaughlin, B. (Eds.) (2006). *Contemporary Debates in the Philosophy of Mind* (Oxford: Blackwell Press)

Davidson, D. (2004a). *Problems of Rationality*. (Oxford: Clarendon Press)

Davidson, D. (2004b). Paradoxes of Irrationality. (In Davidson 2004a (pp. 169-188).)

Davidson, D. (2004c). Incoherence and Irrationality. (In Davidson 2004a (pp. 189-197).)

Davidson, D. (2004d). Deception and Division. (In Davidson 2004a (pp. 199-212).)

Davidson, D. (2001a). *Subjective, Intersubjective, Objective*. (Oxford: Clarendon Press)

Davidson, D. (2001b). First Person Authority. (In Davidson 2001a (pp. 3-14).)

Davidson, D. (2001c). Knowing One's Own Mind. (In Davidson 2001a (pp. 15-38).)

Evans, G. (1982). *The Varieties of Reference*. (Oxford: Oxford UP)

Fairweather, A. and Zagzebski, L. (Eds). (2001). *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. (Oxford: Oxford UP)

Frankfurt, H. (1998a). *Necessity, Volition, and Love*. (Cambridge: Cambridge UP)

Frankfurt, H. (1998b). The Faintest Passion. In Frankfurt 1998a, pp. 95-107.

Frankfurt, H. (1988a). *The Importance of What We Care About*. (Cambridge: Cambridge UP)

Frankfurt, H. (1988b). Identification and Externality. In Frankfurt 1988a, 58-68

Frankfurt, H. (1988c). The Importance of What We Care About. In Frankfurt 1988a, 80-94

Frankfurt, H. (1988d). Identification and Wholeheartedness. In Frankfurt 1988a, 159-176

Freud, S. (1966). *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. (J. Strachey (Ed. and Trans.), in collaboration with A. Freud. London: Hogarth Press)

Freud, S. (1911). Formulations on the Two Principles of Mental Functioning. (In Freud 1966 (vol. XII, pp. 218-226).)

Gallois, A. (2004). *The World Without, The Mind Within: An Essay on First-Person Authority*. (Cambridge: Cambridge UP)

Grice, P. (1974-75). Method in Philosophical Psychology (From the Banal to the Bizarre). *Proceedings of the American Philosophical Association*, 23-53

Hampshire, S. (1977). *Two Theories of Morality*. (Oxford: Oxford UP)

Heil, J. (1984). Doxastic Incontinence. *Mind*, *93*, 56-70

Hieronymi, P. (2006). Controlling Attitudes. *Pacific Philosophical Quarterly, 87*, 45-74

Holton, R. (2000). What is the Role of the Self in Self-Deception? *Proceedings of the Aristotelian Society, 101*, 53-69

Holton, R. (1999). Intention and Weakness of Will. *Journal of Philosophy, 96*, 5, 241-62

Hookway, C. (2001). Epistemic *Akrasia* and Epistemic Virtue. (In A. Fairweather and L. Zagzebski 2001 (pp. 178-199).)

Hursthouse, R. (2002). *On Virtue Ethics*. (Oxford: Oxford UP)

Johnston, M. (1988). Self-Deception and the Nature of Mind. (In B. McLaughlin and A. Rorty 1988 (pp. 63-91).)

Knight, M. (1988). Cognitive and Motivational Bases for Self-Deception: Commentary on Mele's *Irrationality*. *Philosophical Psychology, 1:2*, 179-188

Lazar, A. (1999). Deceiving Oneself or Self-Deceived? On the Formation of Beliefs "Under the Influence." *Mind, 108*, 265-290

Lear, J. (1988). *Aristotle and the Desire to Understand*. (Cambridge: Cambridge UP)

Lehrer, K. (1997). *Self-Trust: A Study of Reason, Knowledge, and Autonomy*. (Oxford: Oxford UP)

Lockie, R. (2003). Depth Psychology and Self-Deception. *Philosophical Psychology, 16:1*, 127-148

Ludlow, P. and Martin, N. (Eds.) (1998). *Externalism and Self-Knowledge*. (Palo Alto: CSLI Publications)

Maier, N. R. F. (1931). Reasoning in Humans: II. The Solution of a Problem and its Appearance in Consciousness. *Journal of Comparative Psychology, 12*, 181-194

McDowell, J. (1998a). *Mind, Value, and Reality*. (Cambridge, Mass.: Harvard UP)

McDowell, J. (1998b). Some Issues in Aristotle's Moral Psychology. In McDowell 1998a, 23-49

McDowell, J. (1998c). Virtue and Reason. In McDowell 1998a, 50-73

McLaughlin, B. and Rorty, A. (Eds.) (1988). *Perspectives on Self-Deception*. (Berkeley: University of California Press)

Mele, A. (2001). *Self-Deception Unmasked*. (Princeton: Princeton UP)

Mele, A. (1997). Real Self-Deception. *Behavioral and Brain Sciences, 20*, 91-102

Mele, A. (1987). *Irrationality: An Essay on Akrasia, Self-Deception and Self-Control*. (Oxford: Oxford UP)

Montmarquet, J. (1987). Epistemic Virtue. *Mind, 96*, 482-497

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. (Cambridge, MA: Harvard University Press)

Nisbett, R. and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. and Wilson, T. (1977). Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review, 84:3*, 231-259

Noordhof, P. (2003). Self-Deception, Interpretation, and Consciousness. *Philosophical and Phenomenological Research, 67:1*, 75-100

Nozick, R. (1981). *Philosophical Explanations*. (Cambridge, MA: Harvard UP)

O'Brien, L. (2005). Self-Knowledge, Agency and Force. *Philosophy and Phenomenological Research, 71*, 580-601

Patten, D. (2003). How Do We Deceive Ourselves? *Philosophical Psychology, 16:2*, 229-246

Peacocke, C. (2006). Mental Action and Self-Awareness (I). (In Cohen and McLaughlin 2006 (pp. 358-376).)

Pears, D. (1984). *Motivated Irrationality*. (Oxford: Oxford UP)

Rorty, A. (1988). The Deceptive Self: Liars, Layers, and Lairs. (In McLaughlin and Rorty 1988 (pp. 11-28).)

Ryle, G. (1949). *The Concept of Mind*. (Chicago: University of Chicago Press)

Sartre, J. P. (1956). *Being and Nothingness*. (H. Barnes (Trans.). New York: Washington Square Press)

Scott-Kakures, D. (1996). Self-Deception and Internal Irrationality. *Philosophical and Phenomenological Research, 56:1*, 31-56

Shah, N. and Velleman, J. D. (2005). Doxastic Deliberation. *The Philosophical Review, 114*, 497-534

Shoemaker, S. (1968). Self-Reference and Self-Awareness. *Journal of Philosophy, 65*, 555-567

Shoemaker, S. (1988). On Knowing One's Own Mind. *Philosophical Perspectives, Epistemology, 2*, 183-209

Smith, B., Wright, C., and MacDonald, C. (Eds.) (1998). *Knowing Our Own Minds: Essays on Self-Knowledge*. (Oxford: Clarendon Press)

Soteriou, M. (2005). Mental Action and the Epistemology of Mind. *Nous, 39:1*, 83-105

Soteriou, M. and O'Brien, L. (Eds.) (Forthcoming). *Mental Action*. (Oxford: Oxford UP)

Strawson, P. F. (1966). *The Bounds of Sense*. *(*New York: Routledge Press)

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy, 48*, 1–25

Talbott, W. J. (1995). Intentional Self-Deception in Single Coherent Self. *Philosophy and Phenomenological Research, 55:1*, 27-74

Taylor, S. (1991). *Positive Illusions: Creative Self-Deception and the Healthy Mind* (New York: Basic Books)

Velleman, J. D. (2002). Identification and Identity. (In Buss and Overton 2002 (pp. 91-123).)

Velleman, J. D. (2001a). *The Possibility of Practical Reason*. (Oxford: Oxford UP)

Velleman, J. D. (2001b). Introduction to *The Possibility of Practical Reason*. In Velleman 2001a, 1-31

Velleman, J. D. (2001c). Epistemic Freedom. In Velleman 2001a, 32-55

Velleman, J. D. and Shah, N. (2005). Doxastic Deliberation. *The Philosophical Review, 114*, 497-534

Watson, G. (1975). Free Agency. *Journal of Philosophy, 72*, 8, 205-20

Wiggins, D. (1976). Truth, Invention and the Meaning of Life. *Proceedings of the British Academy, 62*, 331–78

Williams, B. (1981a). *Moral Luck*. (Cambridge: Cambridge UP)

Williams, B. (1981b). Internal and External Reasons. In Williams 1981a, 101-13.

Williams, B. (1973a). *The Problems of the Self*. (Cambridge: Cambridge UP)

Williams, B. (1973b). Deciding to Believe. In Williams, B., 1973a, 136-51.

Wilson, G. (2004). Comments on *Authority and Estrangement*. *Philosophy and Phenomenological Research, 69*, 440-447

Wolf, S. (2002). The True, the Good, and the Loveable: Frankfurt's Avoidance of Objectivity. In Buss and Overton 2002, 227-44

Wright, C. (2001a). *Rails to Infinity: Essays on Themes from Wittgenstein.* (Cambridge, MA: Harvard UP)

Wright, C. (2001b). The Problem of Self-Knowledge (I). (In Wright 2001a (pp. 319-344).)

Wright, C. (2001c). The Problem of Self-Knowledge (II). (In Wright 2001a (pp. 345-373).)