UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Evaluators, Explainers, Planners: The Importance of Basic Conceptions of What We Are Like as Agents

Permalink

https://escholarship.org/uc/item/1pn7f2m7

Author

Mitchell-Yellin, Benjamin

Publication Date

2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Evaluators, Explainers, Planners:
The Importance of Basic Conceptions of What We Are Like as Agents

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Benjamin Mitchell-Yellin

June 2012

Dissertation Committee:
Dr. Andrews Reath, Chairperson
Dr. John Martin Fischer
Dr. Agnieszka Jaworska
Dr. Michael Nelson
Dr. Gary Watson

Copyright by
Benjamin Mitchell-Yellin
2012

The Dissertation of Benjamin Mitchell-Yellin is approved:

_____

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

I would be remiss if I did not acknowledge my immense
debt to all of my Committee Members.
Each one of you has contributed a great deal to this project,
and I hope that you find some sign of my respect for you in these pages.

For Christie, Simon and Miles,
who keep me focused on what's really important.

ABSTRACT OF THE DISSERTATION

Evaluators, Explainers, Planners:
The Importance of Basic Conceptions of What We Are Like as Agents

by

Benjamin Mitchell-Yellin

Doctor of Philosophy, Graduate Program in Philosophy
University of California, Riverside, June 2012
Dr. Andrews Reath, Chairperson

This dissertation focuses on two questions: What is right/wrong action? What is self-governed action? I argue that prominent, contemporary answers to these questions—moral theories and theories of self-governance, respectively—are grounded in basic conceptions of what we are like as agents. The substance of these theories can be explained, ultimately, in terms of a conception of what is fundamental to human agency. And I show how this observation has significant implications for how we should understand debates about which of these theories is best.

I focus on three basic conceptions of what we are like as agents: the evaluator, explainer and planning conceptions. I show that three of the leading theories of self-governance are each grounded in a different one of these

conceptions and how this affects the debate about which theory is best. This way of looking at the dialectic tells against a common way of arguing against rival views. Arguments that appeal to intuitions about cases in order to generate counterexamples are not apt to be fair or persuasive because they too easily invoke contentious conceptions of what we are like as agents. But it suggests an alternative way to proceed. One can muster a holistic argument in favor of one's preferred basic conception, showing that it can ground independently plausible philosophical theories of various kinds. I suggest how such an argument might go for the evaluator conception, considering its merits as grounds for both a comprehensive theory of human agency and a moral theory.

Table of Contents

Introduction

You order water instead of beer because you are the designated driver. This action speaks for you; it expresses where you stand; you are fully behind it. We think of ourselves as capable of authoring our behavior in ways that license these metaphors. In a word, we think our actions may be self-governed. Our capacity to exhibit this robust form of agency is a reason we take ourselves to be the kinds of agents who can be morally responsible for what we do and who have the moral standing to make claims on others. You order water because you promised to give your friends a ride home. It would be wrong to order beer and get drunk. It is a deep feature of how we understand ourselves and our relations to others that we take ourselves to be capable of self-governance and subject to the demands of morality.

In this dissertation, I focus on two questions that arise given how we understand ourselves and our relations to others: What is right/wrong action? What is self-governed action? I consider answers to these two questions—moral theories and theories of self-governance, respectively—and argue that prominent contemporary accounts fail to adequately attend to the background conceptions of human agency animating their own and rival views.

My argument has two main strands. On the one hand, I show that by paying attention to the background conceptions of human agency in the contemporary literature we can identify infelicitous arguments and tensions internal to particular views. We can see why the contemporary debate has hit an impasse and why particular views should be revised in certain ways. On the other hand, I suggest a way to move the debate forward by providing a certain kind of holistic argument directly in favor of or against a particular background conception of human agency. The idea is to show that one's preferred basic conception can be seen to ground (or to be properly related to conceptions that ground) independently attractive philosophical theories of various kinds. I focus here on two kinds of theory—theories of self-governance and moral theories—but I think the project could be promisingly expanded to encompass other kinds of theory. Perhaps, for example, there are accounts of the proper function of punishment that can be seen to be grounded in some of the basic conceptions of what we are like as agents identified here (or ones suitably related to them). Insofar as these accounts are independently attractive, they may figure in holistic arguments in favor of these basic conceptions.

Both of these strands run throughout the chapters that follow. Let me briefly summarize each chapter in turn. Chapter One focuses on three influential, contemporary theories of self-governance developed and defended by Michael Bratman, J. David Velleman and Gary Watson. I argue that these theories, though

substantively quite distinct, all share a common structure. And I argue that we can explain the substantive differences between these theories in terms of their being grounded in distinct basic conceptions of what we are like as agents. Bratman's view is grounded in the planning conception, according to which extending one's agency over time by means of plans is fundamental to human agency. Velleman's view is grounded in the explainer conception, according to which understanding the causes of one's own behavior is fundamental to human agency. And Watson's view is grounded in the evaluator conception, according to which justifying one's actions in terms of values is fundamental to human agency.

In Chapter Two, I show that other kinds of theory may also be grounded in one or another of these basic conceptions. In particular, I show that Christine Korsgaard's Kantian moral theory is grounded in the evaluator conception, the conception behind Watson's theory of self-governance. And I argue that this allows that certain arguments may be more forceful than they at first seem to be. I take up as a working example an argument of Bratman's against Watson's view. The argument appeals to a case of putatively self-governed, weak-willed action that is supposed to provide a counterexample to Watson's theory of self-governance. Importantly, we can see that the trouble raised by the case is supposed to stem from the evaluator conception in which the theory is grounded. I argue that, if successful, this counterexample would not only

provide a consideration against Watson's theory of self-governance, but also Korsgaard's moral theory, which is grounded in the same basic conception.

In Chapter Three, however, I argue that the kind of argument from counterexample that looks so promising given the discussion of Chapter Two is not apt to be fair or persuasive. It is difficult to mount a convincing argument of this form against a rival view without invoking a background conception of human agency that is contentious in the relevant dialectical context. I show in detail that this is the case with respect to Bratman's argument. And I sketch a way of responding to this argument, such that the purported counterexample ends up supporting the view it was supposed to discredit. This involves suggesting a way of expanding Watson's theory of self-governance into a more comprehensive account of human agency, including, in particular, an (incomplete) account of moral responsibility that remains grounded in the evaluator conception. This is a first step in suggesting how one might go about providing a holistic argument in favor of a particular basic conception of what we are like as agents.

In Chapter Four, I turn my focus to moral theories. I argue that, as in the literature on the philosophy of action, we find arguments in the literature on moral philosophy that are not apt to be fair or persuasive because they invoke a contentious conception of human agency. I focus, in particular, on Velleman's version of the criticism that Korsgaard's Kantianism does not entail that immoral

action is always irrational and show that his argument for this claim presupposes his own background conception of human agency, which Korsgaard does not share. This discussion helps to further the project of sketching a holistic argument for a particular basic conception. In Chapter Three, I sketch a way of expanding a theory of self-governance grounded in the evaluator conception into a more comprehensive account of human agency. In this chapter, I defend a moral theory grounded in this same conception. Thus, the evaluator conception may be seen to ground both an attractive account of human agency and an independently attractive moral theory.

In Chapter Five, I argue that T. M. Scanlon's contractualist moral theory should be revised. His claim that psychopaths can be properly judged morally blameworthy and properly morally blamed is inconsistent with central aspects of his overall view. In particular, it is inconsistent with the contractualist conception—justification to others in terms of reasons is fundamental to moral agency—in the background of his view. This is interesting, in part, because it shows that we have a further reason to attend to basic conceptions of agency. Doing so can help us to identify tensions internal to particular views. It is also interesting because it suggests a second way of furthering the holistic argument in favor of the evaluator conception. Though they are not identical, the contractualist conception and the evaluator conception are related in such a way that the former may be seen as derivable from the latter. Thus, the evaluator

conception may be seen to be properly related to the grounds of a different independently attractive moral theory.

In the Conclusion, I offer some thoughts about the shape of the dialectical landscape going forward. My hope is that this dissertation will convince the reader that attending to the conceptions of human agency behind some of our best philosophical theories has real and significant payoffs.

Chapter One:

Three Theories of Self-Governance

We take ourselves to be capable of authoring our behavior in a strong sense. A theory of self-governance aims to capture this aspect of our understanding of our agency by articulating when a given action counts as self-governed and why. Contemporary theories of self-governance appeal to a distinction between motives that are "internal" to the agent and motives that are "external" to the agent. Internal motives issue in self-governed action because the agent is identified with them. External motives are alien, so behavior that issues from them is not self-governed.[1] My aim in this chapter is to establish that the way this distinction is drawn is only a reflection of deeper assumptions. The rival sets of internal and external motives specified by the leading contemporary theories of self-governance are specifications of inchoate conceptions of what we are like as agents.

Recognizing this has two payoffs. The first is that it allows us to appreciate the common structure of apparently very different views. Second, noticing that our leading theories of self-governance have a common structure

[1] Though it may still count as action. We should allow that one may perform actions that are not self-governed.

provides us with a better appreciation of the dialectical landscape. In particular, it suggests we may have to rethink the way we assess theories of human action to settle the question which one is best. In subsequent chapters, I detail an impasse in the contemporary literature and suggest a way forward. Roughly, the impasse is due to a lack of recognition by the parties to the debate that what really separates their respective views are different assumptions about what is fundamental to human agency. The way forward is to directly assess the merits of the background conceptions of what we are like as agents that account for the substantive differences between rival theories.

1. Frankfurt and Watson

In an influential series of papers,[2] Harry Frankfurt articulates a view about when a person's motives are his own. A motive is "internal" to an agent, on Frankfurt's view, when he endorses it as the motive he wants to move him to action. Frankfurt calls a second-order desire that a particular first-order desire move one to action a second-order "volition." And he says that the agent is identified with those desires that are the objects of his second-order volitions. When my desire to remain sober is the one I want to move me to act, it is internal to me. I am identified with it because I make it my will. When this desire moves me to order water instead of beer, my action issues from me.

[2] See the essays reprinted in Frankfurt (1988), esp., Frankfurt (1971) and Frankfurt (1977).

Frankfurt's original concern was with moral responsibility. The agent acts freely in the sense required for moral responsibility when his action is motivated by an internal desire. But the notion of internal desires has come to be used, by both Frankfurt and others, in analyses of self-governance.[3] And it is clear why. An internal desire is one with which I am identified. So when I am moved by an internal desire, my action issues from me. If I am identified with the desire that issues in an action, then the action is self-governed.[4]

Gary Watson has raised a well-known objection to Frankfurt's view.[5] The objection is that the view does not identify the right type of attitude to ground an adequate account of self-governance. The agent may be alienated from all of her desires. She may be, to use Frankfurt's term, a "wanton"—she may not care what motive moves her. And this suggests that Frankfurt's account of internal desires in terms of endorsement by higher-order desires is inadequate. That a given first-order desire is the object of an agent's second-order volition does not settle the question whether the desire is internal because the agent may be alienated from

[3] Fischer (2010) and Fischer (ms.) make these points well.

[4] Fischer (2010 and ms.) claims that there has been "mission creep" in both Frankfurt's development of his own view and the literature following from it. I agree with Fischer that Frankfurt's original concern was with moral responsibility. But I am inclined to disagree with Fischer that the later focus on autonomy was an instance of "mission creep." Rather, I think that Frankfurt, and others, failed to adequately distinguish moral responsibility from autonomy. So we can read them as using these terms in something like an interchangeable manner. On my view, there is no real difference between the focus of Frankfurt's earlier and later work. And those who take up the points about internality in his earlier writings in their own discussions of self-governance are not talking about something different than he was.

[5] See Watson (1975, at pp. 27-30 of his 2004).

all second-order volitions. A second-order volition is just a desire for another desire to be effective in action. But if one can be a wanton with respect to all of one's first-order desires, then why cannot one be a wanton with respect to all of one's second-order desires? The question applies at any order. So Frankfurt must either arbitrarily reject the question, or else the view does not really appeal to higher-order volitions in order to account for internal desires.

Watson's objection to Frankfurt's view has been influential.[6] But its full importance has not been adequately noticed. I want to bring out an overlooked aspect of Watson's criticism. His objection to Frankfurt's view and the positive theory Watson proposes in its stead tell us something about the structure of an adequate theory of self-governance.

2. The Structure of a Theory of Self-Governance

Watson considers his "primary point" regarding Frankfurt's view to be "two-fold: that desire can't deliver agential authority, no matter what level we ascend to, because it is not the right kind of concept, and, second, it is not the right kind of concept for these tasks because it lacks inherent normativity, or connection with reasons."[7] We can gloss Watson's understanding of his objection in terms of two claims. First, there is the claim that desire is not the right kind of

[6] Velleman (1992a, at p. 134 of his 2000) argues in a similar way against Watson's view, and Bratman (2003) defends Frankfurt's appeal to hierarchy against Watson's challenge.

[7] Watson (2005, 90).

attitude to ground an account of internal motives. Second, and in support of the first claim, there is the claim that an adequate ground for such an account must be inherently normative. I suggest we understand the inherent normativity Watson requires to be secured by the impossibility of disowning all attitudes of the relevant type. The upshot of Watson's objection, then, is that an adequate account of internal motives must appeal to a kind of attitude that the self-governing agent cannot be without.

This understanding of his objection to Frankfurt is supported by what Watson says in favor of his own view. Watson presents what he calls a "Platonic" account of free human agency. The account is Platonic, he tells us, "in the sense that it involves a distinction between valuing and desiring which depends on there being independent sources of motivation."[8] On this view, the agent's evaluative commitments, the ends and principles that (together with her beliefs) determine her evaluative judgments, are the grounds for specifying the agent's internal motivations. If I order the water because I take this to be the best course of action, then I am moved to act by an internal motivation. My action is self-governed, according to the Platonic theory, when it issues from the motivation to do something because one is committed to its worth.

[8] Watson (1975, at p. 18 of his 2004).

Evaluative commitments ground the Platonic theory's account of internal motivations, and it is clear that Watson thinks it is impossible for the self-governing agent to be alienated from all evaluative commitments.

> One's evaluational system may be said to constitute one's standpoint, the point of view from which one judges the world. The important feature of one's evaluational system is that one cannot coherently dissociate oneself from it *in its entirety*. For to dissociate oneself from the ends and principles that constitute one's evaluational system is to disclaim or repudiate them, and any ends and principles so disclaimed (self-deception aside) cease to be constitutive of one's valuational system. One can dissociate oneself from one set of ends and principles only from the standpoint of another such set that one does not disclaim. In short, one cannot dissociate oneself from all normative judgments without forfeiting all standpoints and therewith one's identity as an agent.[9]

Watson seems to be arguing here for the claim that the agent cannot be alienated from her evaluative commitments in their entirety by appeal to the claim that the agent can only dissociate from one set of values from the standpoint of another. But the better reading of this passage is that Watson is assuming the first claim and then bringing out that the second is an important implication of it. To see this, suppose that the first claim is false—suppose that the agent can be alienated from all of her evaluative commitments. But then the second claim is false. Given the supposition, one can dissociate from one's entire set of evaluative commitments from some other standpoint. For example, one might be able to dissociate from one's evaluative commitments in their entirety from the

[9] Watson (1975, at p. 26 of his 2004). Original italics.

standpoint of one's desires. So Watson's claim that the agent cannot be alienated from all evaluative commitments is assumed, and not supported, by the claim that the agent can only dissociate from one set of values from the standpoint of another.

On Watson's view, the self-governing agent can never be without values, and she assesses whether to reject or to endorse one set of evaluative commitments from the standpoint of some other set of evaluative commitments she already endorses.[10] Evaluative commitments have the required inherent normativity to ground an adequate account of internal motives because the agent cannot be alienated from all attitudes of this kind. This is in contrast to a view, such as Frankfurt's, on which the agent could be alienated from the grounds of her internal desires in their entirety. Watson's remarks suggest the following requirement: an attitude or commitment can adequately ground an account of internal desires only when the self-governing agent cannot be alienated from these attitudes or commitments in their entirety. Call this the *strong internality requirement*.[11] I think this requirement is at the heart of both Watson's objection to

[10] Perhaps, in another moment, this second set of evaluative commitments may come up for assessment and be rejected. The point is not that the agent must have a core set of substantive values that she never gives up, but rather that whenever she considers whether or not to endorse a set of values, she does so from the standpoint of some values she, at least in the moment, affirms.

[11] We should understand the strong internality requirement in the context of Watson's overall project. He characterizes his essays on human action as

> predominantly concerned with two intersecting questions: (1) What makes us agents—that is, individuals whose lives are attributable to them as something they (in part) conduct, not just as something that occurs? (2) What makes us responsible to one another

Frankfurt's view and the Platonic theory he proposes in its place. And I think the fact that Watson assumes, but does not argue for, the claim that evaluative commitments are strongly internal reveals some important features of the structure of his theory of self-governance.

On the Platonic theory, the agent's evaluative commitments are the grounds for specifying her internal desires. The agent is, on this view, identified with those desires she has because she values their objects. Watson's claim about the impossibility of alienation from all evaluative commitments while remaining a self-governing agent explains why he adopts this account of internal desires.[12] Values are *strongly internal* to the self-governing agent—they satisfy the strong internality requirement—and this is why they are the proper determining

---

> for how we carry out our lives? … My answer to both questions appeals to a single notion: the capacity for critical evaluation. We are agents because (and insofar as) we shape our lives by the exercise of normative intelligence; we are answerable to interpersonal norms of criticism because our lives are (in part) reflections of this capacity. That, at any rate, is the rough picture, which I first sketched in "Free Agency" … (Watson 2004, 1-2)

So when Watson claims that to dissociate from all evaluative commitments is to lose one's identity as an agent, we should understand him to be saying more than that one ceases to be the same agent. His point is not that one is not the same person without values. Rather, the point is that one ceases to be an agent (of the same type) without values. (The qualification allows that one might cease to be a self-governing agent while remaining an agent of some kind. That is, it leaves room for the thought that the form of agency we exhibit, the form that concerns Watson, is not the only form of agency and that one might cease to exhibit this form of agency and remain an agent.) The Platonic theory is a view about when the agent is to be identified with her motivations. It is a view about self-governance. So I take it that Watson's claim is that to dissociate from all evaluative commitments is to cease to be a self-governing agent. (I am indebted to Andrews Reath for helping me to see the need to clarify the points made here.)

[12] As mentioned above (footnotes 3 and 4), there is some dispute as to whether there has been a change in focus from the early discussions of free agency, including Watson's (1975) paper, from discussion of moral responsibility to discussion of self-governance. But I am not convinced that their focus has shifted over time. Rather, I think they have all along failed to adequately distinguish between self-governance and moral responsibility.

grounds for the motivations involved in exhibiting this form of agency. Since the self-governing agent cannot be without evaluative commitments, she may be identified with motivations to pursue what she takes to be of value.

The Platonic account of internal desires is determined by the assumption that self-governing agents cannot be alienated from all values. But what motivates this assumption? I propose that Watson makes this assumption because he has a particular *basic conception of what we are like as agents*. Call it the *evaluator conception*. On Watson's conception, justifying actions in terms of value is fundamental to human agency.[13] He conceives of human agency, in the first instance, in terms of pursuing worthwhile courses of action. The assumption of the evaluator conception can explain the claim that values are strongly internal because, on this view, the fundamental feature of human agency is evaluative. One could not justify one's actions in terms of value if one did not have any evaluative commitments. The evaluator conception can also explain the implication of the Platonic account of strong internality that Watson brings out—that we can only reject one set of values from the standpoint of another—because it entails that we occupy some evaluative standpoint or other.[14]

[13] I leave it open, as I think Watson does, what the nature of value is. The idea is that, whatever value is, we justify our actions to one another by appeal to it.

[14] Moreover, we can explain Watson's remark that an account of internal motivations must appeal to an inherently normative concept in terms of him adopting the evaluator conception as his basic conception of human agency. On a familiar sense of "normative," both justification and value are inherently normative. If these concepts are involved in Watson's basic conception of human agency, then an account of internal motivations that did not appeal to them would, from

My claim is that the evaluator conception is an inchoate conception of what we are like as agents that is basic to Watson's Platonic theory because, first, it is assumed as the grounds of the claim about values being essential to being a self-governing agent—the account of strongly internal attitudes—and then, second, it is specified in terms of a condition on motivations that issue in self-governed actions—the account of internal motivations. My view is that Watson's theory of self-governance has a three-part structure. At the surface is an account of how to specify internal motivations: the agent is identified with motivations she has because she judges their objects worthwhile. This account of internal motivations is determined by a claim about a certain kind of commitment that is strongly internal to the self-governing agent: a self-governing agent cannot be alienated from all values.[15] And this claim about strongly internal commitments is grounded in a basic conception of the human agent: acting in ways that can be justified in terms of value is fundamental to our form of agency.

In the next two sections, I will show that this structure is shared by two contemporary theories of self-governance that are rivals to the Platonic theory.

---

the perspective of this conception, appear inadequate. It would fail to take account of fundamental features of our agency.

[15] To claim that evaluative commitments are strongly internal and that they ground specification of internal motivations is to mark a distinction between a type of attitude from which the self-governing agent cannot be alienated and an attitude productive of action that bears a non-contingent connection to the agent in virtue of this essential feature of her agential identity.

3. Velleman's Theory of Self-Governance

In this section, I will argue on the basis of textual evidence that J. David Velleman's theory of self-governance[16] has the same structure as Watson's Platonic theory. To anticipate, Velleman's account of internal motivations—intelligible desires—is determined by the assumption that a particular attitude is strongly internal—the desire to act in accordance with reasons (where reasons are understood to be considerations that help one to make sense of what one is doing). And this claim about strong internality is assumed on the basis of a conception of what we are like as agents—that understanding the causes of one's behavior is fundamental to human agency.[17]

Velleman claims that the self-governing agent cannot be alienated from his desire to act in accordance with reasons. That this is an assumption, and not the conclusion of an argument, comes out in the following passage.

---

[16] For the purpose of exposition, I will focus on the view articulated in Velleman (1992a). As I discuss, in footnote 20, below, Velleman has changed the terminology in which he presents his view. It may be that the view itself has changed over time as well. I will show, however, that the claim I want to establish—that Velleman assumes that a particular attitude is strongly internal—is not affected by the development of Velleman's view. The assumption remains in place over time.

[17] It is worth noting an important difference between Velleman's claim about strong internality and Watson's. Whereas Watson claims that attitudes or commitments of a particular kind are strongly internal, Velleman claims that an attitude with a particular kind of content, not a particular kind of attitude, is strongly internal. The strongly internal attitude for Velleman must have a particular kind of content in order to play the functional role of the agent. But it need not be a particular attitude. "When I speak of a desire to act in accordance with reasons, I don't have a particular desire in mind; any one of several different desires would fit the bill. … In any of its forms, the desire to act in accordance with reasons can perform the functions that are attributed to its subject in his capacity as agent" (Velleman 1992a, at p. 141 of his 2000). This does not affect the similarity in structure between the views, but it does make Velleman's claim about strong internality seem even stronger than Watson's.

> Note that the desire to act in accordance with reasons cannot be disowned by an agent, although it can be disowned by the person in whom agency is embodied. A person can perhaps suppress his desire to act in accordance with reasons; but in doing so, he will have to execute a psychic manoeuvre quite different from suppressing his anger or his addiction to drugs or his other substantive motives for acting. In suppressing his anger, the person operates in his capacity as agent, rejecting anger as a reason for acting; whereas in suppressing his desire to act in accordance with reasons, he cannot reject it as a reason for acting, or he will in fact be manifesting his concern for reasons rather than suppressing it, after all. The only way for a person truly to suppress his concern for reasons is to stop making rational assessments of his motives, including this one, thus suspending the process of practical thought. And in suspending the process of practical thought, he will suspend the function in virtue of which he qualifies as an agent. Thus, the sense in which an agent cannot disown his desire to act in accordance with reasons is that he cannot disown it while remaining an agent.[18]

Here Velleman offers what looks like an argument in support of the claim that the agent cannot disown the desire to act in accordance with reasons, but what he says already depends on this being the case. Velleman supports the claim that to dissociate from one's desire to act in accordance with reasons is to cease to be an agent with the claim that to dissociate from this desire would be to cease the activity of rationally assessing one's motives, which is the distinctive function of an agent. But the second claim only supports the first if the activity of rationally assessing one's motives requires the desire to act in accordance with reasons. To see this, suppose it does not. Then even if rational assessment of one's motives is the distinctive function of an agent, there may be some other attitude or

[18] Velleman (1992a, at p. 142 of his 2000).

commitment from the standpoint of which one might rationally assess one's motives. For example, one might rationally assess one's motives from the standpoint of one's evaluative commitments. A motive might be rational, not (only) because it satisfies one's desire to act in accordance with reasons (understood as sense-making considerations), but (also) because one judges it worthwhile. Then one might disown one's desire to act in accordance with reasons through the exercise of one's capacity to rationally assess one's motives by judging the desire worthless. Perhaps one thinks it is better to "go with the flow" than to concern oneself with whether or not one is doing what makes sense. This would engage rational assessment from the standpoint of evaluation. Thus, if rational assessment does not require the desire to act in accordance with reasons, one could fulfill the distinctive function of an agent while disowning this desire. So one could remain an agent while disowning it.

Velleman is assuming that the distinctive function of an agent requires the desire to act in accordance with reasons. But since this activity is the function in virtue of which one is an agent, the claim that it requires this desire amounts to the claim that this desire is strongly internal—it is an attitude from which the self-governing agent cannot be alienated.[19] So Velleman's support for the claim

[19] Velleman's claim would seem to be that this is an attitude from which *an agent* (and not, specifically, a self-governing agent) cannot be alienated. But this would seem to deny the possibility of different forms of agency, some of which are weaker than the one characteristic of us. Why not allow that there are creatures who do not rationally assess their motives from the standpoint of the desire to act for reasons and that these creatures are agents (of a different kind

that the self-governing agent cannot disown the desire to act in accordance with reasons presupposes that this is the case. What looks like an argument for the claim is better thought of as bringing out an important implication of its truth. The best interpretation of this passage from Velleman is that he is doing something similar to what Watson was up to in the quotation from the previous section. There Watson is bringing out an important implication of his claim that evaluative commitments are strongly internal: a self-governing agent can only disown a set of values in their entirety from the standpoint of another set of values. Here Velleman is bringing out a similar implication of his claim that the desire to act in accordance with reasons is strongly internal: a self-governing agent can only disown a given desire from the standpoint of the desire to act in accordance with reasons, so he cannot disown, but only suppress, the desire to act in accordance with reasons.[20]

---

than us)? We can avoid the implausible implication if we take Velleman's focus to be human agency and suppose that humans are characteristically self-governing agents.

[20] Over time, the language in which Velleman characterizes his account of human agency has changed. But the difference in terminology does not affect his commitment to the claim that there is a particular motive that is strongly internal to the self-governing agent. Consider, for example, the following quotation from a more recent essay, in which he employs the same style of argument as in the above quotation. Here Velleman is concerned to defend his view of reasons for acting—that they are considerations in light of which acting in this way makes sense—from the charge that it cannot account for the normative force of such reasons. In defense of his view, he says that

> the intellectual drive that reasons for acting engage, in exerting their influence, carries a kind of authority by virtue of being inextricably identified with the agent himself. The agent cannot stand back from his drive toward self-understanding and regard it as an alien influence on him, because regarding it as an alien influence at all is an exercise of self-understanding, animated by the self-same drive, which consequently has not been banished to the realm of the alien, after all. (Velleman 2006b, at p. 282 of his 2006a)

Velleman's view is that an agent's behavior is an instance of self-governed action when it is motivated in part by the desire to act in accordance with reasons. This desire plays the functional role of the agent in the production of action by, first, assessing possible motivations and, second, throwing its motivational force behind only those that are intelligible grounds for action. The agent is identified with a first-order desire that satisfies his desire to act in accordance with reasons. This is how we specify internal desires. And internal desires may issue in action even when they are not stronger than the agent's other motives because their motivational force is combined with that of the desire to act in accordance with reasons.[21] My action is self-governed when I order water because it makes sense to me to act on my desire to remain sober. In this case, I act on the combined motivational force of my desire to remain sober and my desire to act for reasons. So even if my desire to have a beer is my strongest first-order motive, the combined motivational force of my (intelligible) first-order desire to remain sober and my (strongly internal) higher-order desire to act for reasons may move me to action.

Here Velleman argues that reasons as he conceives of them have normative force because the agent cannot come to regard the drive for self-understanding, which reasons engage, as alien. In other words, sense-making considerations have normative force as reasons because the attitude they engage is strongly internal.

[21] We might allow that the desire to act in accordance with reasons is an internal desire. It may take itself as object. Then self-governed action would be caused by a combination of internal desires, one of which is always the desire to act in accordance with reasons. If we allow that internal desires are only ever first-order, then self-governed action would be caused by a combination of an internal desire and something else—namely, the desire to act in accordance with reasons. I do not think that any of the claims I make in the text depend on our choosing one of these interpretive options over the other.

The structure of Velleman's view is the same as the structure of Watson's Platonic theory. The account of internal motivations is determined by the account of a strongly internal attitude.[22] And the latter is assumed, not argued for. We may now ask what grounds Velleman's assumption that the desire to act in accordance with reasons is strongly internal.

As with the Platonic theory, I propose that we explain the assumption about strong internality on the basis of a conception of what we are fundamentally like as agents. Call Velleman's view the Understanding theory because it gives pride of place to the agent's motive for understanding the causes of what he is doing.[23] For example, he conceives of reasons as considerations that allow one to make sense of what one is doing.[24] I propose that Velleman's claim that the self-governing agent cannot be alienated from his desire to act in accordance with reasons (understood as sense-making considerations) is

[22] Again, there is a difference between the two accounts in that, for Velleman, there is a single attitude of a particular kind of content that is strongly internal, but for Watson, a particular kind of attitude is strongly internal.

[23] Though this name for the view is my own, it is apt. Compare: "You can dissociate yourself from other springs of action within you, by reflecting on them from a critical or contemplative distance. But you cannot attain a similar distance from your understanding, because it is something that you must take along, so to speak, no matter how far you retreat in seeking a perspective on yourself. You must take your understanding along because you must continue to exercise it in adopting a perspective, where it remains identified with you as the subject of that perspective, no matter how far off it appears to you as object" (Velleman 2000, 30-31). And also: "The self to which autonomous actions are attributed must therefore be the agent's faculty of causal understanding. Insofar as a person's behavior is due to his causal understanding, its causes will appear to that understanding in reflexive guise, and the behavior will properly appear as due to the self" (Velleman 2006, 7).

[24] "This is what I mean when I say that reasons for doing something are considerations in light of which it would make sense. I mean that they are considerations that would provide the subject with an explanatory grasp of the behavior for which they are reasons" (Velleman 2000, 26).

assumed on the basis of the *explainer conception*. On this conception, causal understanding of what one is doing is fundamental to human agency. We understand what we are doing by grasping the causes of our doing it, by making sense of our behavior in terms of the desires that cause it.[25] This conception explains the claim that self-governing agents like us cannot be dissociated from our desire to act in accordance with reasons because to be dissociated from this desire would be to cease to occupy the standpoint from which one judges whether or not it makes sense to act on a given desire. It would be to cease to occupy the perspective from which causal-explanatory judgments about acting from a given desire are made.

Velleman's Understanding theory has a three-part structure. At the surface is an account of how to specify internal motivations: the agent is identified with motivations that make sense to him as grounds for action. This account of internal motivations is determined by a claim about an attitude of a certain kind of content that is strongly internal to the self-governing agent: a self-governing agent cannot be alienated from his desire to act in accordance with reasons. And this claim about strong internality is grounded in a basic conception of the human agent: understanding the causes of our own behavior is fundamental to human agency.

[25] Compare Velleman's remarks about his previous work on action in the "Introduction" to his (2006a, at p. 7): "Autonomous actions are actions performed for a reason, and reasons for performing an action, I argued, are considerations in light of which the action would be understandable in the causal terms of folk psychology."

4. Bratman's Theory of Self-Governance

In this section, I will show that Michael Bratman's theory of self-governance, too, shares this structure. But my argument that this is the case for Bratman will be different for two reasons. First, Bratman's commitment to the strong internality of a certain kind of attitude is implicit in his writings. There is no single quotation that reveals the commitment in the way that the above quotations from Watson and Velleman do. Second, Bratman does not claim that there are any attitudes that are strongly internal to the self-governing agent in the sense that both Watson and Velleman do. Bratman claims that there is a certain function from which the self-governing agent cannot be alienated, but he is a pluralist in that he allows that attitudes of multiple kinds (and not all with the same content) may serve this function. Exposition of Bratman's view is complicated for these reasons.

Bratman develops what he calls the Planning theory of human agency. On this view, intentions are planning attitudes that are not reducible to other psychological states and can be formed independently of evaluation. An agent forms intentions of different levels that fit into hierarchies that constitute "planning structures" in her psychological economy. Roughly, a planning structure is constituted by intentions to act that are part of (partial) plans of

action supported by general principles regarding what to count as reasons for action in deliberating about what to do. Intentions are central to human agency, according to the Planning theory, because the planning structures they constitute help to unify our agency over time. And if we look at why Bratman thinks this explanation works, we can see that the explanation depends on it being the case that attitudes that function to unify our agency over time are strongly internal.

Consider Bratman's answer to what he calls "the problem of agential authority," which "is, roughly, the problem of specifying psychological structures that are such that when they guide, the agent governs."

> Now, elsewhere I have argued that for an attitude to have agential authority for agents like us is in significant part for it to play central roles in the Lockean cross-temporal organization and integration of thought and action. And I have argued, further, that certain plan-type attitudes—in particular, policies concerning what to treat as justifying in practical reasoning—are central cases of attitudes with such authority. These authoritative policies of reasoning need to be embedded in structures of planning agency. So structures of planning agency are an essential element in this solution to the problem of agential authority.[26]

Bratman's conclusion is that planning attitudes are essential psychological elements in an adequate solution to the problem of determining when it is that the agent alone directs her behavior, "the problem of agential authority."[27] I take it that to perform an action with agential authority is to perform a self-governed

[26] Bratman (2009, 430) In footnotes to the quoted text, Bratman refers to his (2000) and (2004), respectively, for the arguments he mentions.

[27] On agential authority, see Bratman (2001, at pp. 91-2 of his 2007).

action. So I take it that his conclusion can be paraphrased as the claim that planning attitudes are essential to an adequate theory of self-governance.[28] Bratman supports this conclusion with three claims he has argued for elsewhere. First, a significant requirement on attitudes essential to a theory of self-governance is that they support "Lockean ties." Roughly, the claim is that they support the temporal extension of our agency. Second, general intentions to treat certain desires as reason-giving in deciding what to do, or "self-governing policies," are important attitudes that do support our temporally extended agency.[29] Third, self-governing policies are planning attitudes.

Putting it all together, Bratman supports his conclusion that planning attitudes are essential to an adequate theory of self-governance by appeal to their importance in supporting our temporally extended agency. But this claim only supports the conclusion on the assumption that the self-governing agent cannot be alienated from all of the attitudes (of whatever kind) that support the temporal extension of her agency. To see this, suppose that attitudes that support the temporal extension of one's agency are not strongly internal. But grant that planning attitudes are essential to accounting for temporally extended agency. By the supposition, the agent can be dissociated from her planning attitudes (and

[28] In Bratman (2010) he claims that we can understand central features of our shared agency as grounded in planning attitudes as well.

[29] See Bratman (2000) for discussion of self-governing policies and their role in supporting our temporally extended agency, and see Bratman (2004) for arguments against rival views of what is essential for self-governed agency.

any other attitudes that support her temporally extended agency) and yet remain a self-governing agent. So we would have no reason to think that planning attitudes are essential to an adequate theory of self-governance. The claim that planning attitudes are essential to an adequate theory of self-governance depends on the assumption that attitudes that support the temporal extension of our agency are strongly internal.

Bratman's assumption about strongly internal attitudes is similar to Watson's in that both of their claims allow that there may be (and likely are) many individual attitudes from which the self-governing agent cannot be alienated. For Watson, a kind of commitment—evaluative commitments—is strongly internal, and we likely have many evaluative commitments. For Bratman, attitudes that support our temporally extended agency are strongly internal, and we likely have many such attitudes (perhaps all of the same kind, perhaps not). In this respect, both the Platonic theory and the Planning theory are opposed to Velleman's Understanding theory. For Velleman, a single attitude with a particular kind of content—the desire to act in accordance with reasons—is strongly internal. But Bratman's claim about strong internality differs from both Watson and Velleman in this respect: Bratman does not claim that there is only one kind of attitude (or only an attitude with one kind of content) that is strongly internal. He allows that there may be multiple kinds of attitude (with different kinds of content) that support the temporal extension of our agency.

Bratman's pluralism comes out in the following.

> I do not say that such planning structures are the unique solution to the problem of agential authority. But they are one solution, and a solution that seems characteristic of us. And this role in our self-governance is part of a rationale for these planning structures—structures that involve guidance by norms of consistency and coherence of intention.[30]

Bratman is not claiming that his preferred kind of attitude—planning attitudes—is uniquely suited to serve the strongly internal function of supporting the temporal extension of our agency. He thinks that planning attitudes—specifically, self-governing policies with which the agent is satisfied—are attitudes that support our temporally extended agency and that they are "characteristic of us." But he does not claim that they are the only type of attitude that can support our temporally extended agency.[31] So he does not claim that planning attitudes are strongly internal in the sense that Watson claims that evaluative commitments are and Velleman claims that the desire to act for reasons is. To mark this difference, I will refer to Bratman's claim about strong internality in terms of a strongly internal function. Bratman claims that the self-governing agent cannot be alienated from the function of unifying her agency over time.

---

[30] Bratman (2009, 430).

[31] For example, he notices (Bratman 2000, at p. 44, note 60, of his 2007) that Frankfurtian caring attitudes "have policy-like roles in support of our temporally extended agency."

It is clear that Bratman thinks planning attitudes are especially well-suited to serve this function. He develops a theory of self-governance in terms of them and defends it against rival views. So while it is, strictly speaking, false to say that Bratman claims that planning attitudes are strongly internal—he admits that one might give a theory of self-governance that appeals to some other kind of attitude that supports our temporally extended agency—it would be to understate his view to treat it as though it places planning attitudes on a par with all other attitudes that might serve this function. Indeed, I think that once we recognize the structure of Bratman's theory of self-governance, we can see that his pluralism is not of central importance.

Though he allows that it may not be the only view capable of capturing what is essential to human agency, Bratman seems deeply committed to the truth of his Planning theory. His view has the same three-part structure as both Watson's and Velleman's. It contains an account of internal motivations that is determined by claims about strong internality. And these claims about strong internality are, in turn, grounded in a basic conception of what we are like as agents. This basic conception, I claim, is that we are fundamentally planning creatures. Bratman's pluralistic claims notwithstanding, the best way to understand his view is in terms of a fundamental commitment to our planning nature.

Let me lay out the three elements of the structure of Bratman's view in turn, beginning with the account of internal desires. Central to the Planning theory are self-governing policies with which the agent is satisfied.[32] A self-governing policy is a "higher-order policy about which desired ends to treat as reasons in one's motivationally effective deliberation."[33] It is a type of intention that has the form 'treat desire *d* as reason-giving in practical deliberation.' An internal desire, on this view, is one that is the object of a self-governing policy with which the agent is satisfied. That is, the agent is identified with those desires she intends to treat as reason-giving in deciding what to do. My action is self-governed when I order the water because I intend to treat my desire to remain sober as giving me reason to do so.

According to Bratman, self-governing policies help to unify our agency over time, so they are attitudes of a kind that can serve our strongly internal function. Thus, the Planning theory's account of internal desires is determined by its account of strong internality. Again, given Bratman's pluralism, he is not committed to the claim that planning attitudes are themselves strongly internal.

---

[32] An agent is satisfied with a given policy, roughly, just in case that policy is not undermined by other policies. It is possible for a policy to conflict with other policies such that it is unable to serve its distinctive functions of constituting the agent's identity and coordinating his activity over time. See Bratman (2000, at pp. 35, 44 of his 2007).

[33] Bratman (2001, at p. 101 of his 2007). It is important for Bratman that the policy not be simply to treat a desire as an effective motivation, but rather to treat it as both an effective motivation and a justifying end in practical deliberation. A self-governing policy endorses the desire that is its object as both a motive for action and a justifying end in certain forms of practical reasoning. Bratman glosses the content of self-governing policies to reflect this dual role of the desires that are their objects as follows: to treat a desire as providing a justifying reason in motivationally efficacious practical reasoning. See Bratman (2000, esp. at pp. 38-40 of his 2007).

But given that he does claim that attitudes that support our temporally extended agency are strongly internal and also that planning attitudes can do so, we can explain why the Planning theory specifies internal desires by reference to self-governing policies in terms of the connection between these policies and Bratman's claims about strong internality. The agent is appropriately identified with desires that are the objects of self-governing policies with which the agent is satisfied because these policies are a kind of attitude that can serve the function from which the self-governing agent cannot be alienated.

As with Watson and Velleman, I propose that we conceive of Bratman's account of strong internality as grounded in a conception of what we are fundamentally like as agents. Call the conception basic to the Planning theory the *planning conception*. On this conception, the fundamental feature of our agency is that we plan to reason in certain ways in the future. In the background of this conception is the assumption that we are fundamentally diachronic agents and our basic agential function is to extend our agency over time. According to the planning conception, we satisfy our basic agential function by making plans. We adopt intentions to treat certain desires as reason-giving in future deliberation. The planning conception explains why Bratman focuses on self-governing policies as a solution to the problem of identifying a type of attitude that can support the temporal extension of our agency over time. Given this conception, it makes sense to look to higher-order intentions to play the functional role from

which the self-governing agent cannot be alienated. Self-governing policies are essential to an adequate account of self-governance, as Bratman claims in the first quotation above, because they are the solution suggested by the planning conception to the problem of identifying an attitude that satisfies our strongly internal function.

To summarize, Bratman's Planning theory has the following, by now familiar, three-part structure. At the surface is an account of how to specify internal motivations: the agent is identified with motivations that she intends to treat as reason-giving in deciding what to do. This account of internal motivations is determined by a claim about a type of attitude that (possibly among others) can serve the function that is strongly internal to the self-governing agent: a self-governing agent cannot be alienated from those attitudes that unify her agency over time, and planning attitudes can unify one's agency over time. And this claim about strong internality is grounded in a basic conception of the human agent: planning to deliberate in certain ways is fundamental to human agency.

It is worth noting that, due to his pluralism, the elements of this structure relate differently on Bratman's view than they do on Watson's and Velleman's views. On Bratman's view, the planning conception explains why he focuses on the one kind of attitude—planning attitudes—and not others that can serve the strongly internal function of extending our agency over time. But the planning

conception does not explain the claim that attitudes that support the temporal extension of our agency are strongly internal. On Watson's and Velleman's views, the basic conceptions explain the claims about strong internality. Evaluative commitments are strongly internal because we are evaluators; the desire to act for reasons is strongly internal because we are explainers. But the fact that the basic conception does not serve exactly the same explanatory role for Bratman as it does for Watson and Velleman does not detract from the significance of the fact that all three views share these three elements in common. They all include an account of internal motives, determined by an account of strongly internal attitudes, commitments or functions, which is grounded in a basic conception of what we are like as agents.[34]

5. Conclusion

I have argued that three influential, contemporary theories of self-governance share the following structure. Each provides an account of internal motivations, which is determined by an account of strongly internal attitudes, commitments or functions—ones from which the self-governing agent cannot be

[34] It is worth noting that I think the basic conceptions just discussed are not narrowly of what we are like as self-governing agents. They do not only explain substantive claims about our self-governing agency. They are, more broadly, conceptions of what we are like as agents. Bratman, Velleman and Watson all take self-governance to be a central aspect of our agency with implications for other aspects, and I think these basic conceptions can explain claims they make about other aspects of our agency than self-governance (e.g., making up our minds what to do). I discuss the relationship between self-governance and moral responsibility in relation to basic conceptions of what we are like as agents in Chapter Three. (I would like to thank Agnieszka Jaworska for pressing me to clarify the points made in this footnote.)

alienated. And this account of strongly internal attitudes, commitments or functions is grounded in a basic conception of what we are like as agents. The resulting picture is of three theories of self-governance that differ on the question of what is fundamental to human agency. And it is this difference that explains the different accounts they give of internal motivations.

The rest of this dissertation articulates the importance of this insight into the common structure of these views for debates in the philosophy of action and moral philosophy. The main idea is that this insight provides us with a different appreciation of the dialectic between rival views of these two kinds.

Chapter Two:

The Power of an Argument from Intuition

In the previous chapter, we looked at Bratman's, Velleman's and Watson's theories of self-governance. I argued that all three share a common structure and that we can explain the substantive differences between them by appeal to basic conceptions of what we are like as agents. In this chapter and the next, I will bring out some implications of this observation for debates in the philosophy of action.

In this chapter, I will show how this observation helps us to see that a common form of argument found in the literature on the philosophy of action may be more forceful than it at first seems. One way to argue against a rival account of human agency is to appeal to intuitions about cases. One can argue from the fact that a rival theory of self-governance is inconsistent with an intuitive judgment about a given case to the conclusion that we have reason to reject that theory. I will show how a representative argument of this kind works by targeting the basic conception of what we are like as agents that grounds the rival theory. Then I will show how, though it explicitly targets just one rival theory of self-governance, this argument may provide reason to reject multiple theories at once.

1. Bratman's Argument

Recall the structure shared by the three theories of self-governance discussed in the previous chapter. On the basis of a conception of what is fundamental to human agency, a set of attitudes, commitments or functions is assumed to be strongly internal—if one were to lack this set of attitudes, commitments or functions one would not count as a self-governing agent. And this account of strongly internal attitudes, commitments or functions determines an account of internal motivations—motivations that issue in self-governed actions.

Call two theories of self-governance that specify different sets of internal motivations *rivals*. One way to determine which of two rivals is best is to determine which entails the correct set of internal motivations for a given agent. And one way to determine this is to appeal to intuitions about cases. One can argue that we have reason to reject a given theory of self-governance because it is inconsistent with an intuitive judgment about a given agent's set of internal motivations.[35] I will take one such argument of Bratman's against the Platonic theory as a working example and show that it may be more powerful than it at first seems. We can appreciate the potential force of this argument once we

[35] Alternatively, one might argue that we have reason to reject a rival theory because it is inconsistent with an intuitive judgment about whether a given action is self-governed. Given that self-governed actions are caused by internal desires, the two intuitions are equivalent.

recognize that its target is the basic conception of what we are like as agents that grounds the Platonic theory—the evaluator conception.

Here is Bratman's argument.

> An initial move [in favor of rejecting the Platonic theory] appeals to cases of weakness of will and the like—cases in which value judgments do not bring with them relevant commitments, and relevant commitments go against one's value judgments. Perhaps I think it strictly better to be a person who forgives and turns the other check [*sic*] but nevertheless, in a kind of self-indulgence, allow into my life a willingness to express reactive anger. Though this role of my desire to express my anger diverges from my relevant evaluative judgments, it is not a desire I reject or disown. … So, value judgment is one thing, and ownership another.[36]

The main idea is this. Even though it goes against my conception of the good,[37] my desire to express my anger is intuitively my own. It is an internal desire. But the Platonic theory is inconsistent with this intuition. According to the Platonic theory, evaluative commitments are strongly internal and determine both evaluative judgments and internal desires. So the Platonic theory entails that my set of internal desires does not include my desire to express my anger. This desire is alien to me. Thus, Bratman concludes, we have reason to reject the Platonic theory.

---

[36] Bratman (2003, at p. 144 of his 2007).

[37] Bratman says that "I think it strictly better" not to act on this desire and that allowing my behavior to issue from it "diverges from my relevant evaluative judgments." This is ambiguous between my judging the object of the desire worthless and my judging it less worthy than some other option. I will here take the disconnect between my evaluative judgments and my desires to be the stronger one, according to which the object of my desire is something I judge worthless.

This reason to reject the Platonic theory need not be decisive. All things considered, the fact that the Platonic theory is inconsistent with the intuition that my desire to express my anger is my own may be outweighed by other considerations in favor of the Platonic theory. The point is that this fact counts against the Platonic theory to some degree. It provides reason to reject it.[38]

Recall, from the previous chapter, that the Platonic theory is grounded in a conception of human agency, according to which justifying one's actions in terms of value is fundamental—the evaluator conception. This conception explains the Platonic claim that evaluative commitments are strongly internal. And the claim that evaluative commitments are strongly internal explains the claim that internal desires must share an object with the agent's evaluative judgments. It is because the Platonic theory is grounded in the evaluator conception that it identifies internal desires as it does. When Bratman argues that we have reason to reject the Platonic theory because it identifies the wrong set of internal desires in this case, the problematic feature of the view is its commitment to the evaluator conception. So we can understand Bratman's

[38] I will discuss a complication regarding Bratman's argument in the next chapter. To anticipate, it is not clear that our intuitive judgment about the case is the intuition required for the argument to work. The fact that the Platonic theory is inconsistent with the intuition that the desire is my own provides reason to reject it only if the sense in which the desire is my own is the sense relevant to self-governance. I will argue that the sense in which the desire is my own may be better understood as relevant to moral responsibility, as distinct from self-governance. So the fact that the Platonic theory is inconsistent with the intuition may not provide reason to reject it.

argument as impugning the evaluator conception.[39] We have reason to reject the Platonic theory because it is grounded in the evaluator conception.

2. Family Ties

I will now show that a theory of self-governance we have not looked at previously is also grounded in the evaluator conception. Then, in the next section, I will show that because of this Bratman's argument may provide reason to reject this theory of self-governance as well as the Platonic theory. This is particularly interesting because this fourth theory of self-governance is foundational to a Kantian moral theory. Thus, I will show how an argument against a theory of self-governance may provide a consideration in favor of rejecting a moral theory.

Christine Korsgaard argues that we can account for the sources of normativity from a conception of action. She develops a theory of practical reasons and a moral theory out of a theory of self-governance. Call her theory of self-governance the Practical Identity theory.

[39] According to Bratman's Planning theory, one's planning states might not be connected to one's conception of the good. So this argument does not impugn that view. By itself, however, the above argument does not to provide a consideration in favor of the Planning theory either. This would require arguing that the appealed to intuition—that my desire to express anger is internal—can be accounted for by the Planning theory, which requires that the desire be the object of one of the self-governing policies with which I am satisfied. The above argument does not establish this. But the possibility of the Planning theory's capturing this intuition is provided for by the Planning theory's conception of intentions as not necessarily formed on the basis of one's conception of the good. It may be that I have a self-governing policy with which I am satisfied to treat my desire to react in anger as reason-giving in practical deliberation, even though I do not judge this a worthwhile thing to do.

> When you deliberate, it is as if there were something over and above all of your desires, something which is *you*, and which *chooses* which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of *yourself*. … The conception of one's identity in question here is not a theoretical one, a view about what as a matter of inescapable scientific fact you are. It is better understood as a description under which you value yourself, a description under which you find your actions to be worth undertaking. So I will call this a conception of your practical identity. Practical identity is a complex matter and for the average person there will be a jumble of such conceptions. You are a human being, a woman or a man, an adherent of a certain religion, a member of an ethnic group, a member of a certain profession, someone's lover or friend, and so on. And all of these identities give rise to reasons and obligations. Your reasons express your identity, your nature; your obligations spring from what that identity forbids.[40]

Korsgaard claims that we can understand the experience of deliberation in terms of the agent, understood in terms of self-conceptions, standing at a "reflective distance" from her desires and making principled choices regarding which desires to act on. We might call those desires endorsed by the agent's principle of choice "internal desires" and those not endorsed by the principle "external desires." The Practical Identity theory is a theory of self-governance in the same sense as the three theories discussed in the previous chapter. It provides an account of when an action is self-governed—one that the agent is fully behind or that speaks for the agent in a particularly robust way—in terms of motivation by one source, as opposed to others.

[40] Korsgaard (1996a, 100-101).

What I will do in this section is show, first, that Korsgaard's Practical Identity theory shares a structure with the three theories of self-governance discussed in the previous chapter and, second, that it is grounded in the evaluator conception. In §4, I will discuss the significance of the connection between Korsgaard's theory of self-governance and her moral theory.

The standpoint from which the agent adjudicates between desires, on the Practical Identity theory, is a practical identity—a description under which she values herself. So the view includes a set of commitments to one's own worth under certain descriptions, or evaluative commitments, that internal desires bear a particular relationship to. Internal desires are endorsed from the standpoint of a practical identity in the sense that they are consistent with the principle of choice embodied in the identity. In the following passage, we can see both that this view is committed to the claim that there is a particular set of attitudes or commitments from which the self-governing agent cannot be alienated and that the particular set picked out by the view can be explained by the evaluator conception.

> It is necessary to have *some* conception of your practical identity, for without it you cannot have reasons to act. We endorse or reject impulses by determining whether they are consistent with the ways in which we identify ourselves. Yet most of the self-conceptions which govern us are contingent. … Because these conceptions are contingent, one or another of them may be shed. … What is not contingent is that you must be governed by *some* conception of your practical identity. For unless you are committed to some conception of your practical identity, you will lose

> your grip on yourself as having any reason to do one thing rather than another—and with it, your grip on yourself as having any reason to live and act at all. But *this* reason for conforming to your particular practical identities is not a reason that *springs from* one of those particular practical identities. It is a reason that springs from your humanity itself, from your identity simply as *a human being*, a reflective animal who needs reasons to act and to live. And so it is a reason you have only if you treat your humanity as a practical, normative, form of identity, that is, if you value yourself as a human being.[41]

Korsgaard claims that we have a necessary practical identity. As a self-governing agent, one cannot be alienated from one's human identity—the commitment to one's own worth under the description of a reflective creature who needs reasons to act. In other words, she claims that one's human identity is strongly internal.

The element from which the self-governing agent cannot be alienated, on Korsgaard's view, is a conception under which she *values* herself. It is an evaluative commitment—a commitment to one's worth under the description of a reflective creature who needs reasons to act. This follows from the human identity's being a practical identity and a practical identity being a description under which one values oneself. But Korsgaard does not offer an argument for the claim that a practical identity should be understood in this way. Rather, she assumes that this is so. We can explain this assumption by appeal to the evaluator conception. If we take the Practical Identity theory to be grounded in the evaluator conception, then we take it to be grounded in the conception that justifying one's actions in terms of value is fundamental to human agency.

[41] Korsgaard (1996a, 120-121).

To see more clearly that the Practical Identity theory is grounded in the evaluator conception in this way, suppose that it is not. Assume that we need practical identities to live and to act, but suppose that justifying one's actions in terms of value is not fundamental to human agency. Then we might identify ourselves as human beings, or anything else, on some other basis than descriptions under which we value ourselves. For example, we might adopt practical identities because they are descriptions under which we can make causal-explanatory sense of what we are doing. I might adopt the identity of a jazz-lover because I make sense to myself under this description. And, given this identity, certain actions might make sense for me to perform. Rather than being justifying considerations in terms of value, these reasons would be considerations that make sense of what I am doing.[42] It may even be the case that my particular practical identities depend on my human identity, where this is an intelligible description of myself as a creature who needs reasons to act.

What this shows is that the framework of Korsgaard's Practical Identity theory admits of different substantive versions.[43] Starting from the thought that we adjudicate between desires from the perspective of our conceptions of our identity, one of which is the necessary human identity, we might fill in the theory

[42] "Thus, for example, one's being interested in jazz would explain why one might frequent nightclubs, and so one can frequent nightclubs not only out of an interest in jazz but also on the grounds of that interest, regarded as explanatory of one's behavior" (Velleman, 2006, 8).

[43] I will return to this point in Chapter Four, where I argue that Velleman conflates his basic conception of what we are like as agents with the framework of Korsgaard's Practical Identity Theory in arguing that hers is a "concessive Kantianism."

on different grounds. The particular theory Korsgaard develops depends on a conception of the human agent, according to which justifying one's actions in terms of value is fundamental. It is grounded in the evaluator conception.

3. Collateral Damage

Both Watson's Platonic theory and Korsgaard's Practical Identity theory are grounded in the evaluator conception. I will now argue that because they are grounded in the same basic conception of what we are like as agents, Bratman's argument may provide reason to reject both of these theories.

Bratman's argument that we have reason to reject the Platonic theory appealed to an intuitive judgment about an agent's desire in a given case and the fact that this intuition is inconsistent with the set of internal desires entailed by the Platonic theory for the agent in this case. Since the inconsistency is due to the Platonic theory's being grounded in the evaluator conception, we should expect that a different theory of self-governance grounded (in the same way as the Platonic theory) in the evaluator conception would be equally inconsistent with the intuition appealed to in Bratman's argument.[44]

[44] The qualification allows that there may be theories of self-governance grounded in the evaluator conception but not in the same way as the Platonic theory. This allows, for example, that there may be a theory of self-governance that is committed to the evaluator conception but that does not contain an account of strongly internal commitments. I see no reason to eliminate this possibility but will not consider whether there are any such views in the literature.

We can see that this is, indeed, the case for the Practical Identity theory. On this view, the agent's set of internal desires is constituted by those desires endorsed from the standpoint of the agent's practical identities.[45] Since the agent's practical identities are descriptions under which she values herself, they are evaluative commitments—commitments to one's own worth under certain descriptions. On the natural assumption that the agent's practical identities are evaluative judgments,[46] the agent could not judge that a given action is not worthwhile and yet endorse the desire to perform it. The Practical Identity theory entails that my desire to express my anger that features in Bratman's case is external to me, and so the theory is inconsistent with the intuitive judgment appealed to in Bratman's argument. If this inconsistency provides reason to reject the Platonic theory, it also provides reason to reject the Practical Identity Theory.[47]

---

[45] One complication here is that the agent's human identity is a *necessary* practical identity. Thus, while it is a description under which one values oneself and while it constitutes part of the standpoint from which one evaluates candidate motives for acting, it is not just like the other elements of that standpoint. The other elements may be shed, but the human identity cannot. This marks a difference between the Practical Identity theory and the Platonic theory. For the latter, there is no particular necessary evaluative commitment. This difference is not relevant to the current point in the text, and I will come back to it in Chapter Four. (I thank Agnieszka Jaworska for urging me to clarify this difference here.)

[46] Not all evaluative judgments need be practical identities, but it seems clear that all practical identities are evaluative judgments. (I thank Agnieszka Jaworska for helping me to see the how to put this point.)

[47] I will argue in the next chapter, however, that Bratman's argument may not provide reason to reject the Platonic theory. The point here is about consistency. If the argument is successful against the one view, it is successful against the other as well. This remains the case, of course, even if the argument proves unsuccessful.

This means that Bratman's argument may be more forceful than it at first seems. It may tell against a further theory of self-governance in addition to the rival view it explicitly targets.

4. Moral Theory

The Practical Identity theory grounds both a theory of practical reasons and a moral theory. As Korsgaard puts it, "Your reasons express your identity, your nature; your obligations spring from what that identity forbids." I will discuss Korsgaard's moral theory in detail in Chapter Four, but let me say something brief here about how the Practical Identity theory is supposed to ground these other theories.

A practical identity gives you reasons to act because it is a conception of your actions as worthwhile and your life as worth living that serves as a principle for determining which desires to count as reasons. To act against a practical identity is to give up this conception and the reasons it gives you. But as a human being you need reasons to act and to live. So you cannot give up all practical identities. A given practical identity obligates you by forbidding you from treating certain desires as reasons. Some of these obligations are moral because one of your practical identities is your human identity. This identity is special because it is both necessary and universal. Your human identity is universal because every person is a human being like you, and this is why the

obligations it grounds are moral. It obligates you to respect the value of persons. Your human identity is necessary in the sense that it is strictly internal. You cannot be alienated from your human identity and remain the same kind of agent. It unconditionally forbids you from treating certain desires as reasons, and this applies both to reasons for acting and to reasons for adopting other identities. So your moral obligations are grounded in your human identity, and they apply both to what you ought to do and what sort of person you ought to be.

In light of the fact that the Practical Identity theory grounds both a moral theory and a theory of practical reasons, Bratman's argument may be much more forceful than it at first seems. If it may provide reason to reject this theory of self-governance, it may provide reason to reject a theory of practical reasons and a moral theory as well.

This indicates a point of connection between debates in agency theory and debates moral theory. Korsgaard's is not the only moral theory grounded in a theory of self-governance. Any Kantian moral theory will ground its account of obligation in an account of self-governance. And there are other ways for a moral theory to depend on a theory of self-governance. If we follow Bernard Williams in distinguishing moral theories from ethical theories on the grounds that blame is essential to the former, but not the latter, then we understand a moral theory

as owing us an account of moral responsibility.[48] I will argue in the next chapter that theories of moral responsibility should bear a particular relation to theories of self-governance. If this is right, then there is a universal point of connection between moral theories and theories of self-governance. Moral theories must say something about moral responsibility, and what they say should be properly related to a theory of self-governance. This means that arguments like Bratman's may have implications for which moral theories we should accept by providing reasons to reject the grounds of certain views. Such an argument may not provide decisive reason for anything. But it may provide some degree of reason to reject a moral theory, to modify its grounds or to modify its account of moral responsibility.

6. Conclusion

What I aim to have shown in this chapter is that arguments like Bratman's can extend beyond their explicit targets. These arguments appeal to intuitions about cases and challenge the basic conceptions of what we are like as agents that ground the particular theories of self-governance that are their targets. I have shown that one may challenge multiple theories of self-governance at once in this way. And I have suggested that it is possible for these arguments to challenge moral theories as well. If I am right about all this, then arguments like

[48] See Williams (1985).

Bratman's appear very attractive. But I think that any optimism this might engender should be tempered with caution. Though these arguments may be more forceful than they at first seem, it is deceptively difficult to make them work. In the next chapter, I spell out some of the difficulty.

Chapter Three:

The Difficulty of Persuasively Arguing from Intuition

In the previous chapter, I showed how the force of arguments from intuitions about cases is related to the observation that a theory of self-governance is grounded in a basic conception of what we are like as agents. I argued that an argument of Bratman's against the Platonic theory works by targeting its ground, the evaluator conception. I also argued that Bratman's argument may be more forceful than it at first seems. Because multiple theories may be (similarly) grounded in the evaluator conception, Bratman's argument may provide reason to reject more than just the Platonic theory. In particular, I showed how it may provide reason also to reject Korsgaard's Practical Identity theory and the moral theory it grounds.

In this chapter, I want to revisit Bratman's argument from a more critical perspective and show that there is room to doubt its success in providing reason to reject the Platonic theory in the first place. We have reason to doubt that the judgment on which Bratman's argument depends is really about self-governance. But the argument works only if the judgment is about self-governance. So we have reason to doubt that Bratman's argument is effective even against the theory it targets. Moreover, I will propose a way of expanding the Platonic

theory to provide conditions on moral responsibility for weak actions. Then I will show how the case that features in Bratman's argument may end up providing reason to accept this expanded version of the Platonic theory. So the conclusion of the previous chapter, that arguments like Bratman's may be more forceful than they at first seem, should be tempered with caution. This is the case only when the judgments they appeal to are clearly the right ones for the purpose. And, as I will suggest at the conclusion of this chapter, it is not at all clear how to decisively establish that the judgments appealed to by arguments of this type are up to the task.

1. The Right Intuition

For the purposes of this discussion, let us adopt a particular framework in which to discuss claims about self-governed actions. In a context where we are concerned with identifying self-governed actions, we let "the agent" refer to a set of desires, *D*. We evaluate the truth of claims of the form 'Action *X* is a self-governed action of agent *A*'s' by determining whether one (or more) member(s) of the set of desires with which (in this context) *A* is identified, D_A, (alone) causes *X*. The claim is true just in case the condition is satisfied.

Now we want some way of marking the difference between desires that when they cause actions these count as self-governed and desires that when they cause actions these do not count as self-governed. So call any particular desire, *d*,

that is a member of D_A *internal* to A, and call any particular desire, *d'*, that is not a member of D_A *external* to A. Thus, a self-governed action of *A*'s is caused (only) by one (or more) of *A*'s internal desires. Actions caused by external desires are not self-governed. The framework leaves it open how to distinguish between internal and external desires. Thus, we can characterize rival theories of self-governance—for example, the Platonic theory and the Planning theory—using the framework. One advantage of the framework is that it presents the differences between these rival theories in a particularly stark manner.[49]

With this framework in hand, let us return to Bratman's argument against the Platonic theory, presented in the previous chapter. Let us begin by looking again at the text from which Bratman's argument is taken.

> An initial move [in favor of rejecting the Platonic theory] appeals to cases of weakness of will and the like—cases in which value judgments do not bring with them relevant commitments, and relevant commitments go against one's value judgments. Perhaps I think it strictly better to be a person who forgives and turns the other check [*sic*] but nevertheless, in a kind of self-indulgence, allow into my life a willingness to express reactive anger. Though this role of my desire to express my anger diverges from my relevant evaluative judgments, it is not a desire I reject or disown. … So, value judgment is one thing, and ownership another.[50]

[49] Note that it is also compatible with the framework that there are other contexts, not concerned with self-governed actions, where the agent is identified with other psychological states than those in *D*, or not reduced to psychological states at all. So to adopt the framework is not to commit to a wholesale reduction of the agent. And one may adapt the framework for contexts concerned with other kinds of action than self-governed action.

[50] Bratman (2003, at p. 144 of his 2007).

Consider the claim that my desire to express my anger "is not a desire I reject or disown." For Bratman's argument to work, the sense of this claim must have to do with self-governance. Otherwise, the fact that I do not reject or disown the desire is irrelevant to whether or not we should reject the Platonic theory. As a theory of self-governance, the Platonic theory entails, for a given agent, a set of desires that can cause self-governed actions. To show that the Platonic theory entails the intuitively incorrect set of desires, "I" in this claim must refer to a self-governing agent and the senses of "to own" and "to reject" at issue must have to do with desires that are internal and external to this self-governing agent.

We might say that the referent of "I" here must be a set of attitudes or commitments from which I cannot be alienated and that grounds specification of a set of internal desires, *D*, that includes my particular desire to express my anger. This way of understanding the claim is consistent with Bratman's argument providing reason to reject the Platonic theory. Recall that the Platonic theory identifies my evaluative commitments as inalienable and as grounding a set of internal motivations that share objects with my evaluative judgments. In this case, my effective desire does not share an object with any of my evaluative judgments. So the Platonic theory entails that it is external.

But there is room to doubt whether this way of understanding the claim that the desire in Bratman's case is my own is the intuitively correct one. It may be that the intuitively correct judgment about this case is not that my desire is

internal to me in the sense relevant to self-governance, but that it is still one I do not reject or disown. It may be one with which I am identified for some other purpose than identifying self-governed actions. In order to see this more clearly, let me flesh out Bratman's case in more detail and explain why the intuitively correct judgment in this more detailed case is most plausibly about moral responsibility. Then I will explain why the intuitively correct judgment in the case as originally presented may also be about moral responsibility.

Consider the following more detailed version of Bratman's case. Suppose that my desire to express my anger causes me to berate a colleague for mispronouncing my name. Suppose also that everything is the same as in the original version of the case—for example, the desire that causes my behavior does not share an object with one of my evaluative judgments. Now suppose that my colleague resents me for the outburst. It was wildly out of line, and I should have been able to better control my temper. The Platonic theory does not tell us whether her resentment is appropriate. As a theory of self-governance, it tells us simply that my action is not self-governed, but it does not (by itself) tell us when I am morally responsible for my non-self-governed actions.[51] Given this more detailed version of the case, however, we want a theory that can tell us whether

[51] The qualification allows that we might pair the Platonic theory with claims, to which it is not committed, about the relationship between self-governed actions and actions for which the agent is morally responsible. See the discussion in §2.

my colleague's resentment is appropriate. We want to know whether I am morally responsible for my outburst.[52]

Given our concern about whether I am morally responsible for my outburst, the claim that I do not reject or disown my desire to express my anger may plausibly be taken to be about moral responsibility. My relationship to the desire is noteworthy because it bears on the object of concern. This, however, makes it the case that the claim may not bear on the question whether the Platonic theory entails the intuitively correct set of internal desires relevant to identifying self-governed actions.

Even if we analyze moral responsibility in the same way as is called for by our framework for analyzing self-governance—that is, by reducing the agent to a set of desires that can cause actions for which the agent is morally responsible—the content of a claim about moral responsibility may be distinct from the content of a claim about self-governance. If the claim that I do not reject or disown this desire is about moral responsibility, "I" may refer to something other than the set of attitudes or commitments identical to the self-governing agent and the senses of "to own" and "to reject" may have to do with desires that are distinct from

[52] It may be that we want to know both whether I am morally responsible for the outburst and whether it is self-governed. I think this is probably correct because I think self-governance and moral responsibility are deeply related. I discuss the relation in §2. The point here, however, remains: the Platonic theory, as a theory of self-governance, does not tell us everything we want to know about this case. I am assuming that a theory of moral responsibility (perhaps among other things) gives the conditions for the appropriateness of the reactive attitudes (guilt, resentment, indignation). This is a common way of understanding the purpose of a theory of moral responsibility. See, e.g., Fischer and Ravizza (1998) and Wallace (1994).

those that cause self-governed actions. So if we understand the relevant claim to be that this desire is internal to me for the purpose of identifying actions for which I am morally responsible, we are not licensed to infer that it has the same content as a claim about a desire internal to me for the purpose of identifying self-governed actions. But the Platonic theory has implications only for the truth or falsity of claims about whether desires are internal to an agent for the purpose of identifying self-governed actions. Thus, if the claim that I own my desire to express my anger is about moral responsibility, it may not provide reason to reject the Platonic theory on the grounds that it entails the incorrect set of internal desires.

If we return now to the case as originally presented by Bratman, we can ask whether the claim about my ownership of my desire to express my anger is most plausibly about self-governance or moral responsibility. If the argument is to provide reason to reject the Platonic theory, the claim must be about self-governance. So Bratman must takes our concern in this case to be about self-governance.[53] But the details of the case leave it an open question what exactly our concern is. Thus, I do not think that this way of understanding the claim is forced on us.

[53] It may be that he takes a concern about moral responsibility to be identical to a concern about self-governance, for instance, because he does not see a distinction between the two concepts. I am not sure what Bratman takes the relationship between self-governance and moral responsibility to be.

I will argue, in §5, that the claim is plausibly understood to be a claim about moral responsibility, but not self-governance. But I want to set up that argument with some considerations about the relationship between self-governance and moral responsibility, which I will develop in §§2-5. The point to notice now is that insofar as it is an open question whether the claim that my desire is my own is relevant to self-governance, there is room to doubt the force of Bratman's argument against the Platonic theory. This is not to say that Bratman cannot convince us that our concern in this case is about whether my action is self-governed. The point is that one might justifiably need more than Bratman's report of the intuition that I own my desire to be convinced that the case bears on the merits of the Platonic theory.

2. Expanding the Platonic theory

If the intuitive judgment to which Bratman's argument appeals is best understood as a judgment about moral responsibility, then Bratman's argument does not provide reason to reject the Platonic theory. But if this is the best way to understand the intuition in this case, then the Platonic theory leaves us wanting more. We want to know whether I am morally responsible for my angry outburst, and the Platonic theory (by itself) is silent regarding this question. In this section, I will propose a way of expanding the Platonic theory so that it can give us what we want. I will show how the Platonic account of self-governance

can be expanded into an account of when we are morally responsible for our weak actions. This does not yield a complete theory of moral responsibility, but it gives us a model for how we might develop one out of a theory of self-governance. In the next section, I will show how the Planning theory and Understanding theory can likewise be expanded.

In Chapter One, I summarized each of these three theories of self-governance in terms of (a) an assumption about a set of attitudes, commitments or functions from which the self-governing agent cannot be alienated—strongly internal attitudes, commitments or functions—that can be explained by (b) a basic conception of what we are like as agents and also determines (c) an account of motivations that issue in self-governed actions—internal motivations. Given the framework adopted in the previous section of this chapter, we can summarize these theories of self-governance in terms of a principle that specifies (i) the condition on internal desires and (ii) the causal role of internal desires in producing self-governed actions. The following principle summarizes the Platonic theory in this way.

> **Platonic SG**: For any self-governing agent, A, there is a set of internal desires, D_{SG}, such that for any $d \in D_{SG}$ (i) the object of d is both the object of this desire and of one of A's judgments of what is worthwhile because A has an evaluative commitment to it and (ii) self-governed action is caused by some $d \in D_{SG}$.

This principle captures the manner in which the Platonic theory tells us how to determine whether or not a given action is self-governed.

One way to develop a theory of moral responsibility would be to begin from a theory of self-governance captured in a principle like Platonic SG and articulate additional principled conditions on actions that are not self-governed but for which the agent is morally responsible. These conditions could then be added to the condition on self-governed action. The account of moral responsibility would develop out of the account of self-governance in stages by adding principled conditions on actions for which the agent is morally responsible to cover all necessary cases. Let me develop the proposal for the Platonic theory in some detail, focusing just on a condition that tells us which weak actions the agent is morally responsible for. With this slightly expanded account before us, I will provide reason to think that this is a promising way of developing a full theory of moral responsibility out of the Platonic theory.

The proposal is to begin from Platonic SG and articulate a principle that tells us which weak actions the agent is morally responsible for. I think we can identify a suitable principle for this purpose in Watson's account of the distinction between weak and compulsive desires. Consider, for example, the difference between Diego, who desires to eat a second piece of carrot cake, and Sam, a kleptomaniac who desires to steal a pair of shoes. Watson proposes that both Diego and Sam

> may be subject to desires of exactly the same strength. What makes the former weak is that they give in to desires which the possession of the normal degree of self-control would enable them to resist. In contrast, compulsive desires are such that the normal capacities of resistance are or would be insufficient to enable the agent to resist.[54]

Here is what Watson says about why Sam's desire is too strong, and so his action is compelled, but Diego's is not, so his action is weak.

> [T]here are capacities and skills of resistance which are generally acquired in the normal course of socialization and practice, and which we hold one another responsible for acquiring and maintaining. Weak agents fall short of standards of "reasonable and normal" self-control (for which we hold them responsible), whereas compulsive agents are motivated by desires which they could not resist even if they met those standards. … In the case of weakness, one acts contrary to one's better judgment *because one has failed* to meet standards of reasonable or normal self-control; whereas, this explanation does not hold of compulsive behavior.[55]

What does the work here of distinguishing between Diego's and Sam's desires is a *reasonable judgment about what is to be expected of a human agent*.[56] It is reasonable to expect a human agent to develop and maintain the capacity to resist the desire to eat a second piece of carrot cake that moves Diego, but it is not reasonable to

[54] Watson (1977, at pp. 48-9 of his 2004).

[55] Watson (1977, at p. 50 of his 2004).

[56] What makes the judgment 'reasonable' as opposed to simply the one we happen to accept? This raises issues that I do not need to enter into here. I leave it open that one might specify a 'reasonable judgment' in various ways compatible with the proposed amendment to the Platonic theory.

expect a human agent to develop and maintain the capacity to resist a desire to steal shoes of the sort that plagues Sam.

I think we can use reasonable judgments about what is to be expected of a human agent to distinguish between desires that cause weak actions for which the agent is morally responsible and those that cause weak actions for which the agent is not morally responsible.[57] They provide a principled way of expanding Platonic SG into an account of moral responsibility for weak actions.

Consider the following principle.

> **Platonic MR**: For any self-governing agent, A, there is a set of internal desires, D_{MR}, such that for any $d \in D_{MR}$ (i) the object of d is both the object of this desire and of one of A's judgments of what is worthwhile because A has an evaluative commitment to it or (ii) d is a desire that A should have developed and maintained the capacity to resist and (iii) A is morally responsible for an action caused by some $d \in D_{MR}$.

This principle tells us how to determine whether or not the agent is morally responsible for certain actions. In the spirit of the framework adopted in the previous section for discussing theories of self-governance, it does so by identifying the agent with a set of desires that may cause actions for which the

[57] I will not weigh in here on the question whether one is morally responsible for all weak actions. But the answer to this question has implications for the relationship between my view about moral responsible actions and Watson's (1977) view about weak actions. Watson's claim seems to be that all actions caused by desires that one should have developed and maintained the capacity to resist are weak. My claim will be that one is morally responsible for all actions caused by such desires. These claims jointly entail that one is morally responsible for all weak actions. If one wants to avoid the entailment, one must deny either Watson's or my claim. I will remain agnostic here regarding the truth of Watson's claim, and this allows me to remain agnostic regarding the entailment.

agent is morally responsible. By Platonic MR, the agent is morally responsible for an action caused by an internal desire or by a desire she should have developed and maintained the capacity to resist. The first clause entails that the agent is morally responsible for self-governed actions, and the second clause entails that the agent is morally responsible for certain weak actions.

Platonic MR entails that self-governed actions are a proper subset of those actions for which the agent is morally responsible. This is an intuitively correct result. Self-governance is an intuitively more demanding concept than moral responsibility in the sense that the conditions on self-governed actions are more stringent than the conditions on morally responsible actions. But these two sets of conditions are intuitively related. So if a given action satisfies the more stringent conditions, it seems plausible that it would satisfy the less stringent conditions as well.

One way to see this would be to imagine that we were constructing creatures capable of performing self-governed actions.[58] We could begin with a creature capable of purposive behavior and add capacities in stages until we arrive at a creature capable of performing self-governed actions. At each step in the sequence, we would add capacities to those possessed by the previous

[58] See Velleman (2000, 11-12, 22-23, 26) and Bratman (2000) where they use the method of "Gricean creature-construction" (following Grice (1975)) to build self-governing agents. The sequence of construction for both Velleman and Bratman begins from less demanding forms of agency, requiring less sophisticated capacities, and ends with a self-governing agent, with more sophisticated capacities on top of the less sophisticated ones. Neither Velleman nor Bratman consider morally responsible creatures in their sequences of construction, as I suggest we might in the text.

creature to arrive at a higher form of agency. The sequence of creatures would be like a set of Russian dolls. As we ascend the hierarchy of creatures, the set of capacities possessed by each creature contains the set possessed by the previous one and more. And it is plausible to suppose that the exercise of each higher form of agency requires the exercise of the lower form of agency plus some. So the capacities possessed by the creature in the previous stage of the construction would be exhibited in the exercise of the higher form of agency at the next level. If we think that morally responsible agency is a less demanding form of agency than self-governing agency, then the morally responsible creature will come earlier than the self-governing creature in the sequence of construction. The self-governing creature will have all of the capacities required for performing actions for which it is morally responsible plus some, and self-governed actions will manifest the exercise of the capacities required for morally responsible action plus some. Since one is morally responsible for actions that exhibit some, but not all, of the capacities exhibited in self-governed actions, one is morally responsible for all self-governed actions and some non-self-governed actions.

This relation Platonic MR entails between self-governance and moral responsibility points to an explanatory virtue of the principle. It helps to explain why it may seem as though Bratman's argument provides reason to reject the Platonic theory, even if the judgment in the case is really about moral responsibility. It is understandable that one might mistake a judgment about

moral responsibility for one about self-governance given that all self-governed actions are ones for which the agent is morally responsible. Difficulties arise because the converse does not hold. All actions for which the agent is morally responsible are not self-governed. So Platonic MR does not only show that we can expand the Platonic theory to account for the judgment that my desire to express my anger can cause actions for which I am morally responsible. It can also explain why it may seem as though the judgment about Bratman's case is that this desire is internal to me—that is, relevant to self-governance.

3. Expanding the Planning theory and the Understanding theory

We can come up with cases where the Planning theory or Understanding theory would entail that an action is not self-governed, but yet we want to know whether the agent is morally responsible for it. So we should be interested also in whether the Planning theory and Understanding theory can be expanded to account for moral responsibility for weak actions.

Consider the following case.

> **Mary**: Mary is a young woman who continues an unwanted pregnancy and believes it would not only be best to give up her child but bad for both her and the child to stay together as a family. But when the time comes to leave the child with the adoption agency, Mary signs the hospital release and takes her child home with her instead.[59]

[59] This case is modeled on one discussed by Watson (2002), which he borrows from Frankfurt.

Mary judges it best to give her baby up for adoption, but she takes it home with her instead. Suppose that Mary has a self-governing policy with which she is satisfied of treating the desire to take her baby home as reason-giving in motivationally efficacious practical deliberation. Given this detail, the Planning theory entails that Mary's action is self-governed. But suppose now that Mary has no such policy and the Planning theory entails that Mary's taking the baby home is not self-governed. We might still want to know whether this is something for which she is morally responsible. And, like the Platonic theory in the case of my angry reaction, the Planning theory (by itself) has nothing more to tell us. The same will be true of the Understanding theory. We could revise Mary's case such that the Understanding theory entails that her taking the baby home is not a self-governed action and we want to know whether she is morally responsible for it.

The upshot is that all three of these theories of self-governance only tell us how to discriminate between self-governed actions and non-self-governed actions. But given any one of them, there will be some non-self-governed actions for which we want to know whether the agent is morally responsible. In the case presented in Bratman's argument, the relevant action was weak-willed in the traditional sense of going against the agent's evaluative judgment. I express my anger even though I judge that it is not a good thing to do. For ease of exposition, I will refer to the relevant actions even in the cases involving the Planning theory

and Understanding theory as weak. This goes beyond the traditional sense of weakness of will, since, according to the Planning theory and Understanding theory, actions against one's evaluative judgment may be self-governed and non-self-governed actions may accord with one's evaluative judgment. But I do not think that this mars the point I wish to make. And it will be convenient to have a single term to pick out the type of action I am concerned to account for in expanding these theories of self-governance into theories of moral responsibility.[60] In what follows, then, I will suggest a way to expand these theories of self-governance to identify weak actions for which the agent is morally responsible. I recognize that the expansion suggested here does not clearly yield a theory of moral responsibility adequate to capture all actions for which the agent is morally responsible. The idea is to suggest how such an expansion might go.

I think we can use the same device—a reasonable judgment about what is to be expected of a human agent—as the basis for a principled expansion of all three theories of self-governance. We saw how this might go for the Platonic

[60] Though I will not argue for the point here, I think there is good reason to understand the meaning of the term "weakness of will" in relation to one's preferred theory of self-governance, as opposed to always picking out actions that go against one's evaluative judgment. Compare Richard Holton's remark: "One would expect the property opposed to weakness of will to be strength of will" (1999, 251-2). I am inclined to agree with Holton and also to take a theory of self-governance to be concerned with the property of strength of will. Interestingly, Holton argues that weakness of will is not *akrasia* and adopts Bratman's (1987) account as the correct account of strength of will.

theory in some detail. I will now sketch how it might go for the Planning theory and Understanding theory.

Consider the Planning theory first. Let the following principle summarize the Planning theory as a theory of self-governance.

> **Planning SG**: For any self-governing agent, A, there is a set of internal desires, D_{SG}, such that for any $d \in D_{SG}$ (i) d is the object of a self-governing policy with which A is satisfied and (ii) self-governed action is caused by some $d \in D_{SG}$.

This principle tells us how to determine whether or not a given action is self-governed.

We can use the device of a reasonable judgment about what is to be expected of a human agent to expand Planning SG into an account of moral responsibility for weak actions captured in the following principle.

> **Planning MR**: For any self-governing agent, A, there is a set of internal desires, D_{MR}, such that for any $d \in D_{MR}$ (i) d is the object of a self-governing policy with which A is satisfied or (ii) d is a desire that A should have developed and maintained the capacity to resist and (iii) A is morally responsible for an action caused by some $d \in D_{MR}$.

This principle tells us how to determine whether or not a given weak action is one for which the agent is morally responsible and captures the intuitively correct judgment that the agent is morally responsible for all self-governed actions and some non-self-governed actions. According to Planning MR, the

agent is morally responsible for an action caused by an internal desire or a desire she should have developed and maintained the capacity to resist. The former actions are self-governed, on this view, and the latter are weak.

Consider now the Understanding theory. We can expand it in the same way as the other two theories to yield the following pair of principles.

> **Understanding SG**: For any self-governing agent, A, there is a set of internal desires, D_{SG}, such that for any $d \in D_{SG}$ (i) d satisfies A's desire to act in accordance with reasons and (ii) self-governed action is caused by the combination of the desire to act in accordance with reasons and some $d \in D_{SG}$.
>
> **Understanding MR**: For any self-governing agent, A, there is a set of internal desires, D_{MR}, such that for any $d \in D_{MR}$ (i) d satisfies A's desire to act in accordance with reasons or (ii) d is a desire that A should have developed and maintained the capacity to resist and (iii) A is morally responsible for an action caused by some $d \in D_{MR}$.

This pair of principles tells us how to identify self-governed actions and weak actions for which the agent is morally responsible. And they entail the intuitively correct result that the agent is morally responsible for all self-governed actions and some non-self-governed actions.[61]

[61] It bears repeating that the expanded accounts just given of these theories of self-governance are not intended to be complete accounts of moral responsibility. They are intended to cover only cases of self-governed actions and weak actions for which the agent is morally responsible. I take myself to have shown how, in principle, these theories of self-governance may be expanded to give complete accounts of moral responsibility, but I do not take myself to have shown how this will go for all cases of morally responsible action. Nor have I engaged with the interesting question how the theories of moral responsibility that result from expanding these theories of self-governance in the suggested manner would relate to extant theories of moral responsibility

4. One Principle, Three Basic Conceptions

I have suggested that we use the same device—a reasonable judgment about what is to be expected of a human agent—to expand all three of these theories of self-governance. I think this device works especially well in this capacity because the term "reasonable" may be understood in very different ways. And one's understanding of the term may be determined by one's basic conception of what we are like as agents. So the expanded Platonic, Planning and Understanding accounts may be consistent with the very different grounds of the theories of self-governance that are their basis. I will explain why I think is a desirable result. But first, let me briefly discuss how one might interpret the claim that it is reasonable to expect something of a human agent consistently with different basic conceptions of what we are like as agents.

One interpretation of "reasonable" is evaluative. On this interpretation, the reasonableness of a reasonable judgment about what is to be expected of a human agent is grounded in considerations of value. One should, for example, develop and maintain certain capacities because it is bad to fail to do so. The values in question may form the basis for interpersonal convergence on such judgments among members of the moral community. Thus, these judgments may be a suitable basis for determining moral responsibility for actions. And

---

already familiar from the literature. These further questions are interesting but beyond the scope of this discussion.

grounding reasonable judgments about what is to be expected of a human agent in this way is consistent with a conception of the human agent according to which justification in terms of value is fundamental to human agency. If one fails to develop or maintain a capacity it is bad to fail to develop or maintain, then there is no justification in terms of value of your having failed to do so.[62] Thus, determining moral responsibility on the basis of these judgments is in the spirit of the evaluator conception that grounds the Platonic theory.

There are non-evaluative interpretations of "reasonable" as well. A judgment about what is to be expected of a human agent may be reasonable in the sense that it is grounded in desires or intentions widely shared among members of the moral community. Perhaps it is reasonable to expect that one develop and maintain the capacity to resist acting on a particular desire because that desire is inconsistent with widely shared desires or intentions. These common desires or intentions provide a basis for interpersonal convergence on such judgments, so these judgments may be suitable grounds for determining moral responsibility for actions. And grounding reasonable judgments about what is to be expected of a human agent in this way may reflect concern with understanding the causes of behavior or unifying agency over time. So

[62] There may be available justifications in terms of something else. For example, one might justify one's failing to develop a certain capacity in terms of its not cohering with one's planning commitments.

determining moral responsibility on the basis of these judgments may be in the spirit of the explainer or planner conceptions.

Admittedly, these remarks are sketchy. But they are suggestive of the main point I wish to make, which is that we can expand these theories of self-governance to determine moral responsibility for weak actions in ways that are consistent with the basic conceptions of what we are like as agents that ground them. The picture I have painted of the relationship between self-governance and moral responsibility seems to require this. I argued in Chapter One that, for each of the three theories of self-governance that are our focus, the conditions on self-governed action—strongly internal attitudes and internal motivations—are grounded in a basic conception of what we are like as agents. And I claimed, in §2 of this chapter, that it is plausible to conceive of the conditions on self-governed action as including but going beyond the conditions on morally responsible action. But if the conditions on self-governed action include the conditions on morally responsible action, then the conditions on morally responsible action one accepts should be consistent with the grounds of one's theory of self-governance. The way I have proposed expanding the theories of self-governance that are our focus allows for this.

5. Bratman's Argument Revisited

I want now to return to Bratman's argument and the dialectic between Bratman and Watson regarding which is the better theory of self-governance, the Planning theory or the Platonic theory. The preceding discussion has interesting implications for this debate.

There are two competing interpretations of the intuitive judgment that features in Bratman's argument. Bratman presents the judgment that my desire to express my anger is my own as a judgment about self-governance. It must be this way in order to support his argument against the Platonic theory. But there is an interpretation open to the defender of the Platonic theory, according to which the judgment is about moral responsibility. What I want to show now is that these different interpretations can be explained by the different basic conceptions of what we are like as agents that ground the Planning theory and the Platonic theory. Then I will provide some reason to think that the interpretation explained by the evaluator conception is better. The upshot will be that Bratman's argument is best construed, not as providing reason to reject the Platonic theory, but rather as providing reason to accept it.

Consider Bratman's interpretation of the intuitive judgment first. Bratman's argument appeals to the judgment that I do not disown or reject my desire to express my anger. For this to show that the Platonic theory entails the incorrect set of internal desires, it must be a judgment about self-governance. The

judgment must be that this desire is internal to me for the purpose of identifying self-governed actions. The referent of "I" must be a set of attitudes or commitments from which I cannot be alienated and that grounds specification of a set of internal desires that includes my particular desire to express my anger. This must be how Bratman understands the judgment because the Platonic theory is a theory of self-governance, entailing desires internal to the agent for the purpose of identifying self-governed actions, and he takes the judgment to support the conclusion that the Platonic theory entails the incorrect set of internal desires.

It is consistent with the planning conception that grounds Bratman's theory of self-governance to understand the judgment that I do not reject or disown my desire to express my anger as a judgment about self-governance. Recall that the planning conception takes unification of one's agency over time by means of plans to be fundamental to human agency. This conception explains why the Planning theory takes self-governing policies with which the agent is satisfied—a type of intention that can be formed independently of the agent's evaluative judgments—to be strongly internal. Recall also that, according to the Planning theory, the agent's internal desires are the objects of those self-governing policies with which the agent is satisfied. Given this account of internal desires, it is possible that my desire to express my anger is internal to me, even though it goes against my evaluative judgment.

But it is inconsistent with the evaluator conception that the judgment is about self-governance. The evaluator conception takes justification of one's actions in terms of value to be fundamental to human agency. This conception explains why the Platonic theory takes evaluative commitments to be strongly internal. And it is because these commitments are strongly internal, on this view, that the Platonic theory requires that the agent's internal desires share an object with the agent's evaluative judgments. Given this account of internal desires, it is not possible that my desire to express my anger is internal to me.

The fact that, if taken to be about self-governance, the judgment that my desire is my own is inconsistent with the Platonic theory but consistent with the Planning theory might explain why Bratman understands the judgment to be about self-governance. The judgment features in an argument that targets the Platonic theory's account of internal desires, in the context of arguing that the Planning theory's account is superior. As a proponent of the Planning theory, it is understandable why Bratman would construe the judgment that my desire is my own as having to do with self-governance. So construed, the judgment supports accepting his view and rejecting its rival.

The trouble with Bratman's argument, however, is that there is another way of understanding the judgment that my desire to express my anger is my own. The judgment may be that this desire is internal to me for the purpose of identifying actions for which I am morally responsible. The referent of "I" may

be something other than the set of attitudes or commitments identical to the self-governing agent and the senses of "to own" and "to reject" may have to do with desires that are internal and external to the morally responsible, but not necessarily to the self-governing, agent. This possibility is problematic for Bratman's argument because then the intuition does not straightforwardly engage with the question whether the Platonic theory, as a theory of self-governance, entails the correct set of internal desires.

This alternative understanding of the intuitive judgment about Bratman's case is consistent with the evaluator conception. The expanded Platonic account, consisting of the principles Platonic SG and Platonic MR, can be explained by this basic conception of what we are like as agents and entails that I am morally responsible for my angry outburst. The expanded account can be explained by the evaluator conception because Platonic SG is grounded in the evaluator conception, and the principle by which this account of self-governance is expanded admits of an evaluative interpretation. The expanded view entails that I am morally responsible for actions caused by desires I should have developed and maintained the capacity to resist. And supposing that I should have developed and maintained the capacity to resist my desire to express my anger—which, by my own lights, I should have, since I judge that this is not worthwhile—the view entails that I am morally responsible for the action caused by my desire to express my anger. So if the judgment about Bratman's case is

about moral responsibility, it is one that an account of human agency grounded in the evaluator conception can capture.

Just as we could explain Bratman's understanding of the judgment as about self-governance in terms of its supporting rejection of the Platonic theory, we might understand a proponent of the Platonic theory's alternative understanding of the judgment in terms of its supporting accepting the expanded Platonic account. If the judgment is about moral responsibility, it does not support rejecting the Platonic theory as an account of self-governance. This way of understanding the judgment undercuts Bratman's argument. And given that the expanded Platonic account can capture the judgment about moral responsibility, this way of understanding the judgment may support accepting it.

I have just argued that the way one understands the judgment that I own my desire to express my anger in Bratman's case may depend on which basic conception of what we are like as agents one adopts. And this makes it look as though the judgment can do no work in the context of deciding which of two rival theories of self-governance is best. If one's understanding of the judgment is determined by one's basic conception of what we are like as agents, and if one's basic conception determines which theory of self-governance one prefers, then the parties to the debate will understand the judgment in incompatible ways. A dialectical stalemate will ensue, with both parties to the debate talking past each other.

But there need be no stalemate if there are considerations that, independently of one's preferred basic conception, favor one way of understanding the judgment over the other. And I think there are such considerations. Recall that Bratman presents this as a case of *weak* action. So he is committed to the claim that my angry outburst is both weak-willed and self-governed. It is weak-willed because it goes against my evaluative judgment, and it is self-governed, on his view, because it accords with my planning commitments. There may be nothing inconsistent about these claims. But I think that the claim that an action is both weak and self-governed calls for explanation. If the action is self-governed, then what is weak about it? The reply may be that its weakness is due to its failure to accord with the agent's evaluative judgment. But if evaluative judgment is not essential to self-governance, as it is not on the Planning theory, then we want reason to think that it is relevant to the strength of one's will. The Platonic theory, on the other hand, entails that this weak action is not self-governed. On the grounds that there is something *prima facie* odd about claiming that a weak action is self-governed, then, we have reason to understand the judgment in Bratman's case as not having to do with self-governance. And this reason for understanding the judgment in a way that undercuts Bratman's argument is independent of a commitment to the evaluator conception. It is one that even a proponent of the Planning theory may accept.[63]

[63] For example, I think that Richard Holton, who accepts Bratman's account of intentional action,

Given that we have reason not to understand the judgment that my desire to express my anger is my own as a judgment about self-governance, we have reason to doubt the success of Bratman's argument. This may not decisively show that the argument fails, but it does show that more than just a report of the intuition is required to make it convincing. Moreover, if the choice is between understanding the judgment to be about self-governance or about moral responsibility, then we have reason to understand it to be about moral responsibility. Since the Platonic theory can be expanded to capture this judgment, it may end up being the case that the judgment supports our accepting that view. That is, we may find that Bratman's case not only fails to provide reason to reject the Platonic theory, but also provides reason to accept it as an element of an expanded account of human agency.

6. Conclusion

I have done two things in this chapter. First, I have shown that argument's against rival theories of self-governance, such as Bratman's, that appeal to intuitive judgments about cases may be more difficult to make work than it seems. The reason for this is that the judgments these arguments appeal to must be both intuitively plausible and about whether a given action is self-governed. But it is not always clear that they are either of these things. One's understanding

---

should favor the alternative interpretation of the intuition about Bratman's case. See footnote 60, above.

of a judgment about an action in a given case may be influenced by one's theoretical commitments. So it may not be intuitively plausible independent of these theoretical commitments. And it is not always clear that the judgment, even if intuitively plausible, is specifically about self-governance, as opposed to, say, moral responsibility. Self-governed actions are plausibly a proper subset of actions for which the agent is morally responsible, so one might understandably mistake a judgment about moral responsibility for one about self-governance. But the one should not be confused for the other. What follows from these considerations is that it may require more than just a report of one's intuition to make an argument like Bratman's convincing.

The second thing I have done in this chapter is to show how we can expand a theory of self-governance into a theory of moral responsibility. I proposed a way of expanding each of the theories of self-governance discussed in Chapter One to account for moral responsibility for weak actions. And I showed how this expansion might be consistent with the basic conception of what we are like as agents that grounds each theory of self-governance. The result is a more comprehensive account of human agency based in a conception of what is fundamental to human agency. The proposal is importantly incomplete as it stands, since I have not considered how to expand these theories of self-governance to account for all cases of morally responsible actions. But I take the discussion in this chapter to establish that this would be an interesting

avenue of inquiry. And I have already shown how these incomplete accounts of moral responsibility can factor into extant debates about which theory of self-governance is best. This is the first step in sketching a methodology for how to move the dialectic forward in the philosophy of action by mounting a holistic argument in favor of one's preferred basic conception of what we are like as agents.

This concludes my discussion of the implications for agency theory of the observation of Chapter One—that three prominent contemporary theories of self-governance are grounded in basic conceptions of what we are like as agents. In what follows, I turn to issues in moral theory.

Chapter Four:

A Defense of Traditional Kantianism

In this chapter, I turn to consideration of the importance of basic conceptions of what we are like as agents to debates in moral theory. My aim is to do two things. First, I will show how attending to the basic conceptions that ground rival theories can reveal why certain arguments are not apt to be fair or persuasive. As in the previous chapter, we can identify arguments against rival views that presuppose contentious conceptions of what we are like as agents and articulate why they are not fit to move the dialectic forward. Second, I will extend the project of suggesting how to mount a holistic defense of a basic conception, with particular attention to the evaluator conception. The evaluator conception can ground a traditional Kantian moral theory, and this may provide a consideration in favor of the evaluator conception for anyone who is attracted to a moral theory of this kind.

I will pursue these two goals by way of defending the Kantian moral theory developed by Christine Korsgaard in her Tanner Lectures[64] against two versions of the claim that it does not establish the irrationality of immorality. This is a criticism of the view insofar as one wishes, as I take it Korsgaard does,

[64] Korsgaard (1996a).

to defend a traditional Kantianism, on which all immoral action is irrational.[65] The trouble for Korsgaard's view is supposed to follow from her grounding morality in human, as opposed to rational, nature.[66] If this is right, then any traditional Kantianism that is grounded in a conception of what is fundamental to human agency will face the same challenge, no matter which basic conception it is grounded in. Thus, my argument that Korsgaard's Kantianism really does establish the irrationality of immorality shows how a traditional Kantian moral theory may be grounded in a basic conception of what we are like as agents. As I argued in Chapter Two, however, Korsgaard's view is grounded in the evaluator conception. So the argument of this chapter advances the project of suggesting a way of mounting a holistic argument in favor of the evaluator conception in particular.

1. Korsgaard's View

Let me begin with a summary of Korsgaard's moral theory. Some of this will repeat points made in Chapter Two. But the points bear repeating, especially because I will be arguing that her view has been misrepresented in respect to some of them.

---

[65] I borrow the label "traditional Kantianism" from Velleman (2004), which he contrasts with "concessive Kantianism." I discuss the contrast in §7, below.

[66] G. A. Cohen raises this charge forcefully in his comments on Korsgaard's lectures, at Korsgaard (1996a, 172-4). I discuss this point in §2.

According to Christine Korsgaard, our first-person experience of agency reveals a requirement. We need reasons to act and to live. You are a creature that stands at a "reflective distance" from your motives. You are not simply determined by your desires to act this way or that, but you can and do choose which desires to act on. "When you deliberate, it is as if there were something over and above all your desires, something which is *you*, and which *chooses* which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as expressive of *yourself*."[67] You regard your reasons as self-imposed.

In deliberation, on Korsgaard's view, you form conceptions of your "practical identity," and these are the source of your reasons and obligations. A conception of your practical identity is "a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking."[68] When you choose to act on a given desire, you do so because this is justified by some or other of your practical identities. Typically, one has many practical identities—a philosopher, a parent, a neighbor, a citizen, a student, a teacher—and each one gives rise to reasons and obligations. "Your reasons express your identity, your nature; your obligations spring from what that identity forbids."[69]

---

[67] Korsgaard (1996a, 100).

[68] Korsgaard (1996a, 101).

[69] Ibid.

There is one practical identity, however, that has pride of place on Korsgaard's view. Most of your practical identities are contingent. You could have valued yourself under different descriptions (and might do so in the future). But your identity as a human being is necessary. Korsgaard argues for this claim as follows:

> It is necessary to have *some* conception of your practical identity, for without it you cannot have reasons to act. We endorse or reject impulses by determining whether they are consistent with the ways in which we identify ourselves. Yet most of the self-conceptions which govern us are contingent. … Because these conceptions are contingent, one or another of them may be shed. … What is not contingent is that you must be governed by *some* conception of your practical identity. For unless you are committed to some conception of your practical identity, you will lose your grip on yourself as having any reason to do one thing rather than another—and with it, your grip on yourself as having any reason to live and act at all. But *this* reason for conforming to your particular practical identities is not a reason that *springs from* one of those particular practical identities. It is a reason that springs from your humanity itself, from your identity simply as *a human being*, a reflective animal who needs reasons to act and to live. And so it is a reason you have only if you treat your humanity as a practical, normative, form of identity, that is, if you value yourself as a human being.[70]

The claim that your human identity is a necessary part of your self-conception is central to Korsgaard's moral theory. Your human identity both "stands behind" your particular practical identities and is the source of your *moral* reasons and *moral* obligations. These points merit extrapolation.

[70] Korsgaard (1996a, 120-1).

The central idea of Korsgaard's view is that you impose reasons and obligations on yourself by valuing yourself under certain descriptions. Each particular description under which you value yourself is a particular practical identity of yours—teacher, parent, citizen—and imposes a consistency constraint on candidate motivations for actions. You have reason to act on those motivations *consistent* with one of your practical identities and are obligated not to act on those motivations *inconsistent* with any of your practical identities.

But the fact that you have any practical identities at all means that you value yourself under the description of one who needs reasons to act and to live. That you have *contingent* practical identities entails that you have the *necessary* practical identity of a human being. Your human identity is explained in the same way as all other practical identities: you value yourself under a certain description. But it is special in two ways. First, your human identity is implicitly affirmed in the adoption and maintenance of all of your particular practical identities. It "stands behind" them. Your particular identities are normative only given that it is normative. The normativity of any particular practical identities is parasitic on the normativity of your human identity. And your human identity requires that you have some particular practical identity or other. Without these particular identities you would not have reasons to do particular things. Your human identity is necessary, in other words, because you do act for reasons, and this presupposes that you value the need to do so. Second, your human identity

is the source of all of your specifically *moral* reasons and obligations. Your human identity gives you reason to value others' humanity, and it obligates you not to flout the value of others' humanity.[71]

This introduces a distinction between *moral* and *non-moral* reasons and obligations. One has a moral reason to act on those motivations consistent with one's human identity and a moral obligation not to act on those motivations inconsistent with one's human identity. One has a non-moral reason to act on those motivations consistent with one of one's particular practical identities and a non-moral obligation not to act on those motivations inconsistent with one of one's particular practical identities. This makes room for conflict between obligations. For example, a non-moral obligation may conflict with a moral obligation when it would be both inconsistent with one's human identity to act on a given desire and inconsistent with some particular practical identity not to.

The resolution of such conflicts is dictated by the structure of Korsgaard's view. From the two special characteristics of your human identity, it follows that morality is both *rationally inescapable* and *overriding*. Morality is rationally

---

[71] Compare what Korsgaard says about your moral identity (i.e., your human identity):

> But moral identity also stands in a special relationship to our other identities. First, moral identity is what makes it necessary to have other forms of practical identity, and they derive part of their importance, and so part of their normativity, from it. They are important, in part, because we need them. If we do not treat our humanity as a normative identity, none of our other identities can be normative, and then we can have no reasons to act at all. Moral identity is therefore inescapable. Second, and for that reason, moral identity exerts a kind of governing role over other kinds. Practical conceptions of your identity which are fundamentally inconsistent with the value of humanity must be given up. (Korsgaard 1996a, 129-130)

inescapable because your human identity, the practical identity that underwrites moral reasons and moral obligations, is necessary. It is not one that can be shed. Morality is overriding because, in a case of conflict, the conflict must be resolved by shedding the source of one or the other conflicting reasons or obligations. But since the source of your moral reasons cannot be shed, it will always be the case that a conflict between a moral reason or obligation and a non-moral reason or obligation will be resolved in favor of the moral reason or obligation. Morality always wins the day because the source of its normative force is a necessary feature of human agency.

2. Cohen's Mafioso and Korsgaard's Reply

Korsgaard's view is intended to answer what she calls "the normative question," a request for an account of "what *justifies* the claims that morality makes on us."[72] In his comments on Korsgaard's lectures, G. A. Cohen helpfully rephrases the normative question in familiar terms and distinguishes between two contexts in which it may be asked. The normative question is "Why should I be moral?" And this question may be asked either in "the context of self-justification" or in "the context of protest."[73] In the former context, the moralist offers a justification of her morally upright action to the skeptic who cannot see

[72] Korsgaard (1996a, 9-10).

[73] Korsgaard (1996a, 179).

why one should do what morality commands. In the latter context, the moralist offers a justification of why the skeptic should act in a morally upright manner.

Cohen thinks that Korsgaard's distinctive answer to the normative question, that the commands of morality are self-imposed in virtue of one's self-conceptions, provides an adequate and plausible answer to the normative question posed in the context of self-justification. It is perfectly natural to appeal to my conception of my identity in justifying my own actions. But he thinks that her account cannot provide an adequate answer to the normative question posed in the context of protest. The chief problem, according to Cohen, is that Korsgaard's answer appeals to the agent's self-conception, but the protestant version of the question comes out of the mouth of one who is alienated from morality. Given his alienation, it seems that the skeptic's self-conception cannot provide grounds for reason to be moral.

Cohen appeals to some arguments of Hobbes' to press the objection that normativity cannot be secured by self-imposed laws. The first Hobbesian argument is, roughly, that you are obligated to obey the laws you make because you made them. The second is, roughly, that you cannot be bound by a law you have made because in making it you have authority over it. These arguments share the premise that you make the law, but they arrive at opposite conclusions. The first of these arguments is friendly to Korsgaard's view, on which normativity is self-imposed. The second raises a problem for it.

I am not going to consider this problem or Korsgaard's response to it any further here. Rather, I am going to focus on a second problem for the view pressed by Cohen, one that seems to remain (or arise) given that the view can handle the first objection. The second problem Cohen raises for Korsgaard's view is that it entails that morality is rationally escapable.

To illustrate the problem Cohen presents the case of the "idealized Mafioso."

> I call him 'idealized' because an expert has told me that real Mafiosi don't have the heroic attitude that my Mafioso displays. This Mafioso does not believe in doing unto others as you would have them do unto you: in relieving suffering just because it is suffering, in keeping promises because they are promises, in telling the truth because it is the truth, and so on. Instead, he lives by a code of strength and honour that matters as much to him as some of the principles I said he disbelieves in matter to most of us. And when he has to do some hideous thing that goes against his inclinations, and he is tempted to fly, he steels himself and we can say of him as much as of us, with the same exaggeration or lack of it, that he steels himself on pain of risking a loss of identity.[74]

The trouble Cohen's Mafioso is supposed to present for Korsgaard's view is that the reasons and obligations he takes himself to have are grounded in his practical identity, just as the reasons and obligations, and especially the *moral* reasons and *moral* obligations, that we take ourselves to have are grounded in our practical identities. But the Mafioso takes himself to have reasons and obligations to do things that are immoral. Because the normativity of his obligations to do what is

[74] Korsgaard (1996a, 183).

immoral is secured in the same way as the normativity of our obligations to do what is moral Korsgaard's view cannot explain the difference. Her view seems to commit her to the claim that one might not have reason to be moral because one might have adopted practical identities that militate against moral action. It appears that she has not provided us with an answer to the normative question after all.

It is worth making explicit what the Mafioso's psychology must be like in order to pose the problem he is supposed to present for Korsgaard's view. We may suppose that the Mafioso can, but does not in fact, reflect as deeply as you do. If he did not share your reflective capacity, then it is hard to see why we should take him to be in the same game, as it were. We may also suppose that his human identity plays the role in his psychology that your human identity plays in your psychology—it stands behind his particular practical identities. But he simply does not reflect deeply enough to recognize that, say, the desire to murder his rival is inconsistent with his human identity.

The Mafioso example would seem to show that Korsgaard's view entails that morality is rationally escapable. Normativity is self-imposed, on Korsgaard's view. The fact that the Mafioso endorses the desire to murder his rival from the standpoint of his Mafioso identity makes it the case that he really is obligated to do so. Even though he would see that he ought to shed his Mafioso identity because its dictates conflict with his (necessary) human identity if he were to

reflect more deeply, the fact that the Mafioso does not reflect *that* deeply makes it the case that he is obligated to do the immoral thing.

In the beginning of his comments, Cohen focuses on the problem of securing the irrationality of immorality for Korsgaard's particular version of a Kantian moral theory. As Cohen succinctly puts his point, "Kant can say that you must be moral on pain of irrationality. Korsgaard cannot say that."[75] The reason she cannot, according to Cohen, is that she "humanizes" the source of morality. Whereas Kant derived the content of the moral law from reason as such, according to Cohen, Korsgaard derives it from human nature. The problem is that human nature is not up to the task. The feature of human nature central to Korsgaard's view is reflective self-consciousness, and, as Cohen puts it, the trouble is that "all manner of all-too-human peculiarities can gain strength in reflective consciousness."[76]

I take it that Cohen's Mafioso example is supposed to illustrate the moral frailty of human nature. The Mafioso deeply identifies with a particular conception of himself. The reasons and obligations that spring from this deeply held self-conception conflict with morality. But since the conception derives from the activity of his reflective consciousness, it has the same normative force as any other self-conception so derived. So there is no explanation, on Korsgaard's view,

[75] Korsgaard (1996a, 174).

[76] Ibid.

of why we should privilege the reasons and obligations that spring from the Mafioso's human identity over those that spring from his Mafioso identity. The Mafioso acts rationally when he acts immorally

Korsgaard begins her response to Cohen's Mafioso example by pointing out that she explicitly allows for, first, a distinction between *moral* and *non-moral* obligations and, second, conflicts between obligations. This suggests that she sees the Mafioso case as one involving a conflict between a moral obligation (not to murder the rival) and a non-moral obligation (to murder him). The next point she makes is that the explanations of these two kinds of obligation are identical: they both depend on the structure of the agent's psychology. The Mafioso has a moral obligation not to do that which is inconsistent with his human identity and a non-moral obligation not to do that which is inconsistent with his Mafioso identity.

Korsgaard then goes on to discuss "the special status of morality."[77] The justification for any obligation ultimately depends on the source of one's moral obligations. Reflection on the question why, as a Mafioso, one is obligated to murder a rival who wanders onto the block ultimately reveals that this obligation depends on valuing one's humanity. "But," Korsgaard admits, "this conclusion only emerges from a course of reflection, a course which may never be undertaken, or may only be partially carried out, and this does give rise to a

[77] Korsgaard (1996a, 256).

problem."[78] The problem is that, on Korsgaard's view, "it is the endorsement, not the explanations and arguments that provide the material for the endorsement, that does the normative work."[79] The Mafioso endorses his desire to murder his rival, and so he is obligated not to fail to do so. It remains true that "there is no coherent point of view *from which* [this desire] can be endorsed in the full light of reflection."[80] Yet "if one holds the view, as I do, that obligations exist in the first-person perspective, then in one sense the obligatory is like the visible: it depends on how much of the light of reflection is on."[81]

Though capable of deep reflection, the Mafioso fails to reflect deeply enough to uncover the moral obligation that conflicts with his non-moral obligation to murder his rival. In the absence of recognition of conflict, the Mafioso would recognize no reason to shed his Mafioso identity. But then the obligations that spring from it would remain normative for him. And the Mafioso really would be obligated to do what is immoral. Morality appears to be rationally escapable, so long as one does not reflect too deeply. It seems the Mafioso can act immorally in doing as he should.

Korsgaard concludes her reply to Cohen by citing her view that one of the rules constitutive of reflection is that "we should never stop reflecting until we

---

[78] Ibid.

[79] Korsgaard (1996a, 257).

[80] Korsgaard (1996a, 256).

[81] Korsgaard (1996a, 257).

have reached a satisfactory answer, one that admits of no further questioning."[82] The Mafioso ought to have reflected more fully than he did, and if he had, he would have recognized his moral obligation not to murder his rival. This is the sense in which his moral obligation is *deeper* than his non-moral obligation.

I must admit that it is unclear to me how these remarks about the constitutive rules of reflection are supposed to solve the problem, if they are. Yet I do think Korsgaard's view has the resources to show that morality is rationally inescapable.

3. Expanding on Korsgaard's Reply

Korsgaard's reply to Cohen seems unsatisfactory. She grants that the Mafioso's obligation to murder his rival is normative for him in the face of a failure to reflect deeply enough to uncover his moral obligation not to murder his rival. Thus, it appears that one can rationally escape the force of morality by failing to reflect as much as one ought.

But appearances can be misleading. Korsgaard says that the Mafioso has genuine obligations springing from his Mafioso identity. So he may have a genuine obligation to murder his rival. But nowhere does she say that the Mafioso does not *also* have a moral obligation not murder his rival. And nowhere does she say that this moral obligation does not override the conflicting non-

[82] Korsgaard (1996a, 258).

moral obligation. Thus, it would be consistent with what she says to claim that the Mafioso rationally ought to refrain from committing murder.[83]

Elsewhere, Korsgaard defines "true irrationality" as "a failure to respond appropriately to an available reason."[84] In the context of that paper, Korsgaard is interested in defending against skepticism about whether human action can be directed by reason and focuses, in particular, on issues about rational motivation. Whether reasons are rationally motivating is not the same issue as whether one has reason to be moral, but there are interesting connections between Korsgaard's discussion of skepticism about practical reason and Cohen's challenge to her view about the rationality of morality. And these connections lead us, I think, to see how her view can answer Cohen's challenge.

One of Cohen's criticisms is that Korsgaard's answer to the protestant version of the normative question is hopeless. An appeal to one's self-conception

---

[83] Velleman provides us with an example of one who is misled by appearances when he claims that the Mafioso has most reason to do what is immoral on the basis of the following: "Korsgaard does not say that the existence of this latent conflict between the mobster's commitment to humanity and his commitment to the role of a mobster already undermines the normative force of the latter commitment, even before the conflict is discovered and the latter commitment revoked. On the contrary, she says that the latter commitment gives rise to genuinely normative obligations." (Velleman 2004, at p. 305 of his 2006)

Velleman's mistake is to conclude that a given obligation determines what one has most reason to do because its normative force is not undermined by the normative force of a conflicting obligation. This does not follow. We can distinguish between undermining normative force—where this means that the force is no longer present at all—and outweighing or overriding normative force—where this means that the normative force is not decisive. A genuine obligation may be overridden and so not decisive regarding what one has most reason to do without losing all normative force.

[84] Korsgaard (1986, at p. 318 of her 1996b).

cannot reach one already alienated from morality. But consider what Korsgaard has to say about argumentation given the possibility of true irrationality.

> An interesting result of admitting the possibility of true irrationality is that it follows that it will not always be possible to argue someone into rational behavior. If people are acting irrationally only because they do not know about the relevant means/end connection, they may respond properly to argument: point the connection out to them, and their behavior will be modified accordingly. In such a person the motivational path, so to speak, from end to means is open. A person in whom this path is, from some cause, blocked or nonfunctioning may not respond to argument, even if this person understands the argument in a theoretical way.[85]

The result Korsgaard discusses here has to do with rational motivation and behavior. But we get a similar result when we apply the notion of true irrationality to reflection about what one has reason to do. A person who fails to recognize a conflict between a non-moral obligation and a moral obligation only because he has not reflected deeply enough may respond properly to argument: point the conflict out to him, and he will make up his mind what to do in a manner that reflects the overriding rational force of morality. This would be the case for a person whose reflective capacities are, so to speak, open. But one whose reflective capacities are, in some way or other, blocked or nonfunctioning may not respond to argument, even if he can appreciate in a theoretical manner the conflict between the normative dictates of morality and his particular

[85] Korsgaard (1986, at p. 323 of her 1996b).

practical identity. But this person would still be irrational because he would still fail to appropriately respond to an available reason.

Now Cohen has a choice. If, on the one hand, he insists that his Mafioso can reflect just as deeply as you and I, and means by this that his reflective capacities are open, then he poses no threat to Korsgaard's answer to the protestant form of the normative question. He will respond appropriately to the reasons he is presented with. In this case, the Mafioso example would not support Cohen's claim that Korsgaard's view cannot provide an adequate answer to the normative question in the protestant guise.

If, on the other hand, Cohen insists that his Mafioso can reflect as deeply as you and I, and means by this that he has the capacity to do so but that this capacity is presently blocked or nonfunctioning, then the Mafioso may indeed present a hard case for a view that intends to convince *him* that he has reason to be moral. But he does not provide a counterexample to the claim that he, in fact, has reason to be moral. It is the possibility of *irrationality*, after all, that makes room for this characterization of the case. It is because the Mafioso fails to respond to an available reason, even when made aware of it, that he is not a candidate for being argued into the moral way of life. But if he is irrational, then there must be reason for him to be moral. So the Mafioso does not show that morality is, on Korsgaard's view, rationally escapable.

4. Velleman's Take and Where He Gets Korsgaard's View Wrong

Velleman disagrees with my assessment of Korsgaard's response to Cohen. In a recent paper, he argues that Korsgaard's view is a "concessive Kantianism."[86] It entails that one always has reason to *be* moral, but it does not entail that one always has reason to *act* morally. Velleman argues for this claim by looking at Korsgaard's response to Cohen. But Velleman's understanding of her response is very different from the one I have just offered. I will, first, present Velleman's interpretation of Korsgaard's view in the form of his argument for the claim that it is concessive and, then, explain where the interpretation goes wrong and why. My conclusion will be that Velleman has conflated some of his own view of human agency with Korsgaard's practical identity theory. In particular, my claim will be that Velleman's understanding of Korsgaard's practical identity theory presupposes the explainer conception, the conception of what we are like as agents basic to his own account of human agency, but not Korsgaard's moral theory.

According to Velleman, Cohen's basic objection "is that being adopted at will would drain rules or laws of any significant normative force."[87] This is reflected in the second argument he considers out of Hobbes. But Korsgaard meets this objection, according to Velleman, by allowing that the practical

[86] Velleman (2004).

[87] Velleman (2004, at p. 298 of his 2006).

identities one has in place at the moment of choice guide and constrain the present decision. Not just anything goes when one chooses whether to act on a given motivation because one's choice is made from the standpoint of the practical identities one inhabits from the past—including those due to one's previous choices, upbringing, etc.

Korsgaard allows, however, that one may alter one's practical identities by adopting new ones and shedding old ones. So, as Velleman points out, it appears that Cohen's objection re-emerges at the level of decisions about how to make up one's mind. Korsgaard insists that "if I am to be an agent, I cannot change my law without changing my mind, and I cannot change my mind without a reason."[88] But, as Velleman puts it, "the ready availability of reasons will take the bite out of any restriction on changes of mind."[89] According to Velleman, all it takes to have a reason, on Korsgaard's view, is to adopt a principle that constitutes some consideration as a reason. But these are easy to conjure up. So it appears that one's practical identities at a time do not really constrain choice at that time because they can be revoked at will.

But, according to Velleman, on Korsgaard's view, an agent can "alter the range of available reasons only by adopting, shedding, or somehow modifying his practical identities, *and this process takes time*."[90] So the agent cannot, at the

[88] Korsgaard (1996a, 234).

[89] Velleman (2004, at p. 300 of his 2006).

[90] Velleman (2004, at p. 304 of his 2006). Emphasis added.

moment of choice, effect a change in her practical identities at will. Thus, they can guide and constrain her choice.

On Velleman's interpretation, Korsgaard's view is able to meet Cohen's objection that self-imposed laws cannot be normative for the agent, but it entails that one may not have most reason to do what is moral. It is, so to speak, out of Hobbes' frying pan and into the Mafioso's fire.

The reason, according to Velleman, that Korsgaard's view entails that immorality may be rational is that it may be that one has, at the moment of choice, a practical identity from the standpoint of which a motivation to do something immoral is a reason. Because, as Korsgaard herself admits, it is the fact of endorsement that does the normative work, the fact that one identifies with the standpoint from which this consideration is a reason makes it a reason for you. Cohen's Mafioso illustrates this possibility. At the moment when he is choosing whether to act on his desire to murder the rival who wanders onto his block, his deliberative standpoint includes his Mafioso identity. And from the standpoint of this identity, he has reason to murder the rival. On Korsgaard's view, the Mafioso really does have reason to murder the rival and is obligated not to refrain from murdering him. Korsgaard must concede that immoral action can be rational.

Velleman goes on to argue that Korsgaard's view does not entail that the Mafioso's murdering his rival is altogether rational.

> To be sure, the mobster has countervailing reasons, based in his fundamental identity as a human being, as expressed in the Categorical Imperative. But these reasons weigh against acts of murder only indirectly, by committing him to "giving up his role as a Mafioso." They are reasons for him to revoke his commitment to that more particular identity, which turns out to conflict with his underlying identity as a human being, and so they are reasons for him to become someone who no longer has reasons for committing murder. The mobster is irrational to commit murder, not because he doesn't have reasons for committing such an act, but rather because he has reasons against being the sort of person who has those reasons.[91]

According to Velleman, Korsgaard's view shares with Kant's the commitment that acting immorally displays that one is an irrational agent. One should have a different set of reasons. But it does not share the commitment that acting immorally displays irrationality in so doing. On Korsgaard's view, but not on Kant's, the balance of reasons may favor acting immorally. The difference, according to Velleman, is that Korsgaard's view "has eliminated the mechanism by which [the Categorical Imperative] militates against those acts."[92] Kant's view requires that the agent will an act in conjunction with a universalized principle. Since the principles behind immoral actions cannot be universalized, there is a contradiction in willing them. Hence, they are irrational. "In Korsgaard's version of the theory, however, the agent may already be committed to the principle by virtue of having adopted it earlier and not repealed it since. In that case, there

[91] Velleman (2004, at p. 306 of his 2006).

[92] Ibid.

would seem to be no need for him to will the principle afresh in acting on it again."[93] But then there would be no contradiction in his willing the immoral action, and so it would not be irrational.

Here is a reconstruction of Velleman's argument for the claim that Korsgaard's Kantianism is concessive put in terms of an argument for the claim that Cohen's Mafioso is obligated to murder his rival and also to become the sort of person who does not have reason to do so.

> 1. The Mafioso's practical identity at a given time guides and constrains his choice at that time whether to act on the desire to murder his rival.
>
> 2. From the standpoint of his Mafioso identity, the Mafioso is obligated to murder his rival.
>
> 3. From the standpoint of his human identity, the Mafioso is obligated to give up his Mafioso identity.
>
> 4. But it takes time to shed a particular practical identity.
>
> So: 5. At the moment of choice, the Mafioso is obligated to murder his rival.

I think that this argument contains two mistakes.

First, premise (3) is not quite right: from the standpoint of his human identity, the Mafioso is obligated *both* to give up his Mafioso identity *and* not to murder his rival. *Pace* Velleman, your human identity militates against adopting particular identities *as well as* performing particular actions.

[93] Velleman (2004, at pp. 306-7 of his 2006).

Korsgaard's view is that you decide which motivations to act on from the standpoint of your practical identity. This standpoint is constituted by those descriptions under which you value yourself. Most of these descriptions are contingent and particular—parent, citizen, teacher—but one is necessary and shared by all persons: your human identity. As an element of your practical identity, your human identity is part of the standpoint from which you assess candidate motives for acting.[94] And it provides consistency constraints on actions just as any practical identity does. You have reason to act on motivations consistent with your human identity, and you are obligated not to act on motivations inconsistent with it.

You sometimes decide whether or not to adopt a particular practical identity. In these cases, you make the decision also from the standpoint of your practical identity, which is constituted by your human identity and all of your particular practical identities. As with decisions about what to do, decisions about what to be are constrained by your practical identities, including your human identity. You have reason to be someone consistent with your current self and are obligated not to be someone inconsistent with your current self.

Thus, Korsgaard's view entails that your human identity militates *both* against becoming an immoral person *and* performing immoral actions. It does so

[94] "In one sense, your moral identity is just like any other form of practical identity. To act morally is to act a certain way because you are human, to act as one who values her humanity should" (Korsgaard 1996a, 129).

in the same way in both cases, by placing consistency constraints on your decisions. Velleman's premise (3) is incomplete.

The second mistake in Velleman's argument is that premise (4) is ambiguous between a claim about shedding the *motivational force* of a particular practical identity and shedding the *normative force* of a particular practical identity. And the premise is false on the sense required for the argument to engage Korsgaard's view.

If the claim of premise (4) is that it takes time to shed the motivational force of a particular practical identity, then (4) is true. But there is a distinction to be made between *action* and *mere behavior*. Actions issue only from motivations with which the agent is identified.[95] But both actions and mere behavior issue from motivational forces within the agent. So it may be the case that some particular practical identity of mine continues to influence my behavior without it being the case that it gives me reasons or obligations. Korsgaard's view is about the sources of normativity, not the causes of behavior. So this sense of premise (4) does not engage her view.

Alternatively, the claim of premise (4) may be that it takes time to shed the normative force of a particular practical identity. This premise would engage Korsgaard's view. But it would also be false. I see no compelling reason to think

[95] This is consistent with distinguishing between different types of action and different sets of motivations with which the agent is identified for the purposes of identifying actions of these different types. Recall the discussion in Chapter Three, where I show how this goes for self-governed actions and actions for which the agent is morally responsible.

that a given description of myself is something I continue to value—and not something that merely continues to grip me motivationally—after I come to see that I should not value it. But this is what it would have to be like, on Korsgaard's view, for a particular practical identity to continue to have normative force for me even after I recognize that I ought to shed it.

Thus, even if we grant Velleman the claim that, on Korsgaard's view, one's human identity does not militate directly against actions, his argument is problematic. If premise (4) is about normative force, then the argument is unsound. If premise (4) is about motivational force, the conclusion—a claim about normativity—does not follow.

What this examination shows is that Velleman's argument does not establish that Korsgaard's Kantianism is concessive. At the moment of choice, the Mafioso is obligated not to murder his rival. This is because his human identity militates directly against the action—premise (3) is incomplete. Of course, the claim that the Mafioso is obligated not to murder his rival does not by itself show that the Mafioso does not have most reason to do so. It may be that he is also obligated to murder his rival—an obligation springing from his Mafioso identity—and that this obligation overrides the obligation grounded in his human identity. But this is not the case. On Korsgaard's view, conflicts between reasons or obligations springing from different identities are to be settled by shedding or otherwise modifying those identities. But one's human identity is

necessary, so it cannot be shed. Thus, any particular identity that issues in a normative conflict with morality must be shed in order to resolve the conflict.

There are now two cases to consider, analogous to the two characterizations of the Mafioso example considered in §3. Either the Mafioso recognizes that there is a conflict between his human identity and his Mafioso identity or he does not. If he recognizes the conflict, then his Mafioso identity immediately ceases to be normative for him—premise (4) in the normative sense is false. There is no time at which the Mafioso both recognizes reason to *be* moral and yet still has most reason to *act* immorally. And this is the case even if his Mafioso identity continues for some time to exert motivational influence on his behavior. If the Mafioso does not recognize the conflict, he is irrational. His failure to recognize that he should shed his Mafioso identity does not alter the fact that he should. It merely leaves the conflict between the two identities in place and reveals that he fails to appropriately respond to available reasons.[96]

5. Why Velleman Gets Korsgaard's View Wrong

I have just identified two places where Velleman's argument for the claim that Korsgaard's Kantianism is concessive goes wrong. These mistakes involve a

[96] There is a third possibility, which I do not consider here. In §2, I discussed the possibility that the Mafioso may not be able to reflect as deeply as you and I. So he may not be able to recognize the conflict between his human identity and his Mafioso identity. But, as I mentioned above, then he could not present the problem for Korsgaard's view he is supposed to. For this reason, I leave this possibility out of account in discussing Velleman's claim that Korsgaard's view is concessive.

misunderstanding of her view. Now I want to offer an explanation for why he misunderstands her view in these ways. The explanation is, I believe, rather straightforward: he conflates his own view of human agency with hers.

Velleman begins his paper with some remarks about the structure of human agency that he will be working with. On this view, "an agent's motivational set is supposed to represent the contingent, individually variable input to his practical reasoning."[97] And, on top of this motivational set, "each agent must have something else—a project, it might be called—that isn't an end in this sense."[98] The required project is "the project of coping with the reasons that issue from his motivational set, a project that requires a motivational set that issues in reasons with which he can cope."[99]

This conception of human agency does not sound like the one Korsgaard is working with. To begin with, as we have seen, on her view, not all inputs to one's practical reasoning are contingent and variable. One's human identity is a necessary practical identity that all human agents share. And it is an input into practical reasoning because it is an element of the standpoint from which the human agent decides what motivations to act on. Moreover, even if we grant Velleman his talk of motivational sets and a required project of human agency, on Korsgaard's view, it is not *only* the case that one's human identity—the

[97] Velleman (2004, at p. 288 of his 2006).

[98] Ibid.

[99] Velleman (2004, at p. 289 of his 2006).

necessary element of human agency—requires a "motivational set." It is *also* required by the motivational set. On Korsgaard's view, the human identity is (implicitly) affirmed in every instance of acting for a reason. So every time one takes a motivation to issue in a reason one affirms the value of one's humanity. There is nothing like this direction of influence in the conception of human agency Velleman is working with in this paper.

But the conception he is working with does sound a lot like Velleman's own view, the Understanding theory. On the Understanding theory, recall, human agency requires the motivation to make sense. This higher-order motivation is a necessary feature of human agency. And this higher-order motivation requires first-order motivations in order to issue in action. One's set of first-order motivations is constituted by contingent motivations acquired through choice, circumstance, upbringing, etc. One acts when one's motivation to make sense adds its motivational force to a first-order motivation that satisfies it and, thus, the intelligible first-order motivation becomes the strongest motivation in one's psychic economy. Human action, on Velleman's view, issues from an intelligible, contingent motivation in combination with the necessary, higher-order motivation to make sense.

I think that the problem in premise (3) of Velleman's argument can be explained by the differences between Velleman's background conception of human agency and Korsgaard's. Recall that Velleman's premise (3) claims that

the Mafioso's human identity militates only against his Mafioso identity, and not any particular actions. This makes sense given Velleman's conception of human agency. On that conception, the inputs to practical reasoning are the contingent and variable first-order motivations assessed from the perspective of the necessary, higher-order motivation to make sense. Given that the human identity is not contingent and variable, it would fit into this conception, not as an input to practical reasoning, but rather as a higher-order perspective from which the various inputs are assessed. Thus, it would not bear directly on decisions about what to do. Rather, it would bear directly on decisions about the status of first-order motivations, as the motivation to make sense does.

I suspect that Velleman's reading of Korsgaard is influenced by his own views about human agency and that he conceives of the agent's human identity as playing the functional role, in Korsgaard's theory, of the necessary, higher-order "project of coping with the reasons that issue from [one's] motivational set." His idea seems to be that, given Korsgaard's practical identity theory, the way to cope with the reasons that issue from one's motivational set is to shed, adopt or otherwise alter particular practical identities. And so the Mafioso's human identity obligates him to shed his Mafioso identity, but it does not bear directly on the question which action to perform.

However, this is only half the story. Velleman's conception of human agency fails to allow for the possibility that the Mafioso's human identity is an

element of the standpoint from which he makes all practical decisions, both about what to be and what to do. But this is exactly how it functions in Korsgaard's view. The Mafioso's human identity militates directly against the action of murdering his rival because it places consistency constraints on his decisions, *both* about what identities to adopt *and* what motivations to take as giving him reason to act.

My suggestion is that we can explain Velleman's failure to take account of the full role of the agent's human identity on Korsgaard's view by appeal to the different conceptions of human agency they are working with. Velleman does not allow that any of the inputs to practical reasoning are necessary elements of the agent's motivational set. For Korsgaard, however, the normativity of morality is explained precisely by a necessary element of the agent's motivational set, namely, his human identity. Let me turn now to the second mistake in Velleman's argument I pointed out in the previous section and suggest that we can explain this mistake in a similar manner as the first, by appeal to features of Velleman's conception of human agency that Korsgaard does not share.

Velleman's premise (4) is ambiguous between a claim about it taking time to shed the motivational force of a practical identity one recognizes that one should not inhabit and a claim about it taking time to shed the normative force of such an identity. The claim about normative force is the only one that bears on

Korsgaard's view, and this claim is false. It does not take time to shed the normative force of a particular practical identity one comes to see that one should not inhabit. I think we can explain why Velleman does not distinguish between the two senses of this claim and also why he might think that the continued motivational influence on one's behavior of a feature of one's psychology is normatively relevant by appeal to the explainer conception.

Recall from Chapter One that self-governed action, on Velleman's view, issues from a motive that satisfies the agent's motivation to make sense. The motivation to make sense is strongly internal, and the assumption that this is the case can be explained in terms of Velleman affirming the explainer conception—that understanding the causes of one's own behavior is fundamental to human agency. Since the Mafioso's motive to protect the family's turf could explain his murdering the rival who wanders onto his block, the act of murder would, on this view, count as self-governed.

Given the Understanding theory, a feature of one's psychology continues to be normatively relevant so long as it has a sustained motivational grip on one's behavior. Velleman's conception of self-governed action can explain why he conflates a claim about the sustained motivational force of an element in the Mafioso's psychology with a claim about its sustained normative force. And since it is plausible that one cannot instantly get rid of the motivational force of an element of one's psychology simply by recognizing that one should no longer

take it to provide one with reasons, we can see why Velleman would think that it takes time to shed a particular practical identity one recognizes one should not inhabit.

But, again, these claims follow from Velleman's conception of human agency, not Korsgaard's. On the Practical Identity theory, autonomous action issues from motivations that satisfy descriptions you value yourself under. Together with the fact that we can be (and all too often are) motivated by considerations we do not take to be valuable, this view makes room for a robust distinction between elements of your psychology in light of which you can understand the causes of your behavior and elements of one's psychology that justify it as valuable. And human agency, the Practical Identity theory assumes, is basically about the latter. Thus, we can see that Velleman's premise (4) is ambiguous and that neither way of resolving the ambiguity is a happy one for his argument. Either the premise does not support his conclusion or else the premise is false.

6. The Light of Reflection

I would like now to tie together my discussion of Cohen's and Velleman's treatments of Korsgaard's view. I will do so by clarifying what I take to be the role of the agent's human identity in Korsgaard's moral psychology and

reflecting a little further on how both Cohen and Velleman misunderstand her view on precisely this point.

In my discussion of her reply to Cohen, I quoted Korsgaard as saying that "if one holds the view, as I do, that obligations exist in the first-person perspective, then in one sense the obligatory is like the visible: it depends on how much of the light of reflection is on."[100] This comment bears on the example of Cohen's Mafioso in the following way. The Mafioso's obligation to murder his rival is genuine because he does not shed the Mafioso identity from which it springs. He does not even see that he has reason to shed this identity. And he does not see this because he does not reflect deeply enough to recognize that his Mafioso identity conflicts with his identity as a human being. The light of his reflection does not reach all the way down, so to speak.

But, as I argued in my extension of Korsgaard's reply to Cohen, it does not follow from the Mafioso's failure to recognize this conflict that the conflict is not there. On Korsgaard's view, as I understand it, *the Mafioso affirms his human identity each and every time he acts for a reason, including when he decides to adopt or maintain his Mafioso identity or when he decides to act on a consideration that constitutes a reason from the standpoint of his Mafioso identity*. What does follow from the fact of the Mafioso's dim reflection is either that he can be argued into morality by helping him to cast the glow a bit deeper or that he irrationally fails

[100] Korsgaard (1996a, 257).

to respond to an available reason. In neither case is morality rationally escapable for him.

Velleman displays a similar misunderstanding of what follows from the fact of the Mafioso's dim reflection when he says

> In order to maneuver the Mafioso out from under the force of reasons for committing murder, then, we would have to "get him to a place" from which he could see something that he can't currently see from the place he's in, at the moment of pulling the trigger. Indeed, we'd have to get him to a place where he could turn around and see that he couldn't find his way back, a place that would therefore have to be far removed, in the space of reasons, from the place he currently occupies.[101]

Velleman seems to understand the fix to the Mafioso's problem as involving a change in his psychology. We have to help him change what he is. But this is not Korsgaard's understanding of the situation. On her view, the Mafioso already is what he needs to be, a human being. We need only get him to rationally recognize this present, necessary and normative feature of himself.

7. Conclusion

In this chapter, I have defended Korsgaard's traditional Kantianism against the charge that it entails that one might have most reason to do what is immoral. This is significant, in the context of this dissertation, for two reasons. First, the way in which I defended Korsgaard's view against Velleman's

[101] Velleman (2004, at p. 308 of his 2006).

argument was to show that the critique depended on assuming a basic conception of what we are like as agents that is not shared. Velleman's argument depends in key places on his Understanding theory, grounded in the explainer conception, but Korsgaard's traditional Kantianism depends in key places on her Practical Identity theory, which is grounded in the evaluator conception. Noticing this allowed me to diagnose the specific difficulties with Velleman's argument and to show how Korsgaard's view had the resources to establish the irrationality of immorality. In essence, this was the same problem I found, in the previous chapter, in debates in the philosophy of action. Attending to the conceptions of human agency in the background of rival views can allow us to avoid these problems.

The second reason the discussion of this chapter is significant in the context of this dissertation is that it shows one way to begin mounting a holistic defense of the evaluator conception—and, by extension, a basic conception of what we are like as agents in general. Korsgaard's traditional Kantianism is grounded in the evaluator conception, the same conception that grounds the Platonic theory—and the expanded version of it sketched in the previous chapter. Thus, one who holds that justification in terms of value is fundamental to human agency has open to her both a comprehensive account of human action and a traditional Kantian moral theory. Insofar as these are attractive theories in

their own right, the fact that they may be grounded in the evaluator conception provides a consideration in its favor.

Chapter Five:

Scanlon's Trouble with Psychopaths

In the previous chapter, I showed how attending to conceptions of what is fundamental to human agency can illuminate debates in moral theory in a similar manner to the way in which I showed, in Chapter Three, that it can illuminate debates in agency theory—we can recognize when an argument is not apt to be fair or persuasive because it assumes a contentious basic conception of what we are like as agents. In this chapter, I will show that attending to background conceptions of agency can be illuminating in another way—we can recognize when particular claims are inconsistent with more fundamental commitments of a theory. And I will demonstrate how we can argue, on this basis, that a given theory should be revised in certain ways. This is a second way in which attending to basic conceptions of what we are like as agents can inform philosophical debates.

My focus, in this chapter, will be on T. M. Scanlon's contractualist moral theory. This allows for pursuit of a second aim as well. In the previous chapter, I showed that a proponent of the evaluator conception has open to her a traditional Kantian moral theory. In this chapter, I will show that she may,

instead,[102] adopt Scanlonian contractualism. This is interesting, in the context of this dissertation, because it shows another way of mounting a holistic argument in favor of the evaluator conception. It also shows something interesting about the sort of holistic argument I am suggesting. The relation between the evaluator conception and the conception that grounds Scanlonian contractualism is not identity, as it was in the case of the evaluator conception and the conception that grounds Korsgaard's traditional Kantianism. Rather, the two conceptions are related in an appropriate way, which I spell out below. This suggests that one's holistic argument in favor of a given basic conception of what we are like as agents need not show that the same conception can ground various kinds of plausible philosophical theories. It may, instead, show that the basic conception being argued for can stand in the right relation to conceptions that ground various kinds of plausible theories. This would be a more pluralistic way of mounting the holistic argument than that suggested in the previous chapter.

It will be worthwhile to lay out what is to come in this chapter. Scanlon has, over the past thirty or so years, developed a very influential contractualist moral theory, one element of which is a novel account of blame. In his discussions of moral responsibility, Scanlon has consistently claimed that psychopaths, understood as incapable of grasping moral reasons, may still be properly judged morally blameworthy and morally blamed for what they do. In

[102] I don't see that contractualism and traditional Kantianism are compatible, such that one may consistently affirm both at once.

this chapter, I argue that Scanlon's claims about psychopaths are inconsistent with central elements of his contractualism.

The structure of the chapter is as follows. After summarizing Scanlon's account of blame and differentiating my criticism from other well-known critiques of Scanlon's views on moral responsibility, I offer three arguments to show that Scanlon's claims about psychopaths are inconsistent with central aspects of his view. I offer two arguments for the claim that psychopaths do not satisfy a conception of agency that Scanlon's view presupposes applies to those who stand in the moral relationship, the relationship that, on his view, grounds moral blame. Then I offer a third argument for the claim that psychopaths' actions cannot have the meaning they must in order to impair the moral relationship, so psychopaths are, on Scanlon's view, properly held to be exempt from this relationship. If these arguments are successful, there is a tension between Scanlon's claims about the moral responsibility of psychopaths and central aspects of his moral theory. I conclude that he ought to resolve the tension by dropping the claims about psychopaths, rather than the commitments they conflict with, because the latter are more central to his view than the former. I conclude by offering an explanation of why Scanlon might have made the mistake I argue he has.

1. Scanlon's Account of Moral Blame

Scanlon's contractualist moral theory contains a novel account of blame. He summarizes the views as follows:

> to claim that a person is *blameworthy* for an action is to claim that the action shows something about the agent's attitudes toward others that impairs the relations that others can have with him or her. To *blame* a person is to judge him or her to be blameworthy and to take your relationship with him or her to be modified in a way that this judgment of impaired relations holds to be appropriate.[103]

It is important to highlight that this account of blame is relationship-based. All instances of blame are grounded in a particular relationship that provides standards relevant both to determining the status of the agent's conduct—whether he or she is blameworthy for it—and to determining the appropriate response—whether and how it would be appropriate to modify one's relations with him or her on the basis of this conduct.

My focus here will be on Scanlon's account of moral blame, a special case of his general account of blame. On Scanlon's account, we distinguish between kinds of blame by reference to the distinct kinds relationships that ground them. Moral blame is grounded in "the moral relationship: the kind of concern that,

[103] Scanlon (2008, 128-9).

ideally, we all have toward other rational beings."[104] And, according to Scanlon, the moral relationship is a "default relationship" assumed to hold "between us and the strangers we pass on the road or interact with in the market."[105] But, for example, the friendship relation grounds a kind of blame, call it friendship blame, that is distinct from moral blame. Friendship blame is an appropriate (by the standards of the friendship relation) reaction to the impairment of the friendship relation. Moral blame is an appropriate (by the standards of the moral relationship) reaction to impairment of the moral relationship. These two kinds of blame may overlap, but they are distinct because the friendship relation is not identical to the moral relationship. Thus, the standards for impairment and appropriate reaction in the one case are not identical to those in the other.

Scanlon's account of moral blame is embedded in his contractualist theory of the morality of what we owe to each other, and we can see that the relationship that grounds moral blame—the moral relationship—provides a point of connection with other central elements of his view. Scanlon's contractualism is intended to provide a unified account of the distinctive subject matter, epistemology and, primarily, reason-giving force of judgments of right and wrong.[106] It centers on a distinctive account of wrongness: "an act is wrong if

[104] Scanlon (2008, 140).

[105] Scanlon (2008, 141).

[106] See Scanlon (1998, 3) for his summary of these questions and his claim that the question of motivation is primary.

its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement."[107] This account of wrongness is central to Scanlon's account of the subject matter of morality. He tells us that moral judgments are "about reasons and justification."[108] And it is connected to his account of moral motivation via a particular interpersonal relation.

> The contractualist ideal of acting in accord with principles that others (similarly motivated) could not reasonably reject is meant to characterize the relation with others the value and appeal of which underlies our reasons to do what morality requires. This relation, much less personal than friendship, might be called a relation of mutual recognition.[109]

I think it is fair to say that Scanlon is referring to this same relation when he speaks of "the moral relationship." Thus, it is fair to say that Scanlon's accounts of wrongness, moral motivation and moral blame are essentially connected via a single, interpersonal relationship. I will use "the moral relationship" to refer to this relationship in what follows.

The connection between the moral relationship and each of these elements of Scanlon's moral theory may be put as follows. According to Scanlon's account

---

[107] Scanlon (1998, 153).

[108] Scanlon (1998, 4).

[109] Scanlon (1998, 162).

of wrongness, we judge that actions are wrong if they violate general standards for conduct that could not be reasonably rejected by those who are suitably motivated and with whom we stand in this relationship. According to his account of moral motivation, we are motivated by the value we take this relationship to have not to perform wrong actions. According to his account of moral blame, we judge those who act in ways that violate this relationship[110] to be morally blameworthy for what they have done, and we morally blame them by modifying our relations with them in ways deemed appropriate by the standards of this relationship. It is worth bearing in mind the essential connection, via the moral relationship, between Scanlon's account of moral blame and these other central elements of his contractualist moral theory.

2. Criticisms of the Account

Scanlon's account of moral blame has significant appeal, much of which is inherited from the attractiveness of his contractualist moral theory. For example, Scanlon's account of moral blame provides a compelling account of the condemnatory force of moral blaming responses because his contractualism provides such a compelling account of moral motivation. If we value standing in the moral relationship with others, then we will feel the sting of judgments that

[110] Note that, on Scanlon's view, one might violate the moral relationship without performing a wrong action.

we have failed to live up to this ideal and the force of others' withdrawal of some of the intentions and dispositions characteristic of this relationship.

For all its attractiveness, however, Scanlon's account of moral blame has its detractors. It is worth noting how my criticism will differ from theirs. R. Jay Wallace charges Scanlon with leaving "the blame out of blame."[111] Because Scanlon's account of moral blame does not make the reactive attitudes of resentment, indignation and guilt essential, Wallace claims, it omits an important element in any adequate account of moral blame, an element Wallace characterizes as "opprobrium." This is an interesting criticism. However, I will not here engage with the issues Wallace raises and will assume that Scanlon's account of moral blame is adequate with respect to them.

A second criticism of Scanlon's account of moral blame is closer to the issues that concern me here. Watson also charges Scanlon with leaving an essential aspect of moral blame out of his account, albeit a different aspect than the one cited by Wallace. According to Watson, moral responsibility has two "faces." The attributability face involves assessment of the agent in ethical terms for the attitudes expressed in her behavior—"she was callous," "she was dishonest." This first face requires only the "general capacity to recognize and respond to some practical reasons."[112] The accountability face, by contrast,

[111] Wallace (2011, 349).

[112] Watson (2011, 310).

involves interpersonal accountability to basic moral norms. This second face requires "the capacity for moral reciprocity or mutual recognition that is necessary for intelligibly holding someone accountable to basic moral demands and expectations."[113] Scanlon's account of moral blame leaves out the accountability face of moral responsibility because it allows that agents who do not have the stated capacity can nevertheless perform morally blameworthy actions that may license the modifications of intentions and dispositions Scanlon claims are characteristic of moral blaming responses.

Watson focuses in particular on Scanlon's treatment of the case of the psychopath. Scanlon understands psychopaths to "be rational in a general sense, and capable of means-ends reasoning, but nonetheless unable to understand why they have any reason to take moral requirements seriously as limits in the pursuit of their aims."[114] And yet he holds that psychopaths' behavior may be attributable to them for the purposes of moral assessment, including moral blame. "If they see no reason not to kill, injure, or manipulate us when it promotes their ends, then this judgment about reasons is attributable to them. It is their considered judgment about the reasons they have."[115] The judgment that there is no reason not to kill another person shows a lack of concern with the justifiability of one's actions to others. If we can attribute this judgment to an

[113] Watson (2011, 308).

[114] Scanlon (ms., 11).

[115] Scanlon (ms., 14-15).

individual on the basis of his behavior, this shows that our moral relations with that individual are impaired. He is morally blameworthy and it may be appropriate to morally blame him for what he has done.

Watson complains that Scanlon's treatment of the case of the psychopath shows that his account of moral blame is conceptually flawed. Scanlon's account requires only that the morally responsible agent have the general capacity to recognize and respond to practical reasons, and not the more robust capacity to recognize others as co-members in the Kingdom of Equals.[116] But the latter capacity is required for moral accountability. Thus, Scanlon's account of blame fails to "do justice to moral responsibility because it does not do justice to moral accountability."[117]

Like Watson, I will appeal to Scanlon's treatment of the case of the psychopath in order to argue that his account of moral blame is problematic. But, unlike Watson, my aim is not to show that Scanlon's view cannot capture a true claim about the concept of moral responsibility. My aim is to show that Scanlon's

---

[116] I borrow this nice phrase from Pamela Hieronymi. She articulates Scanlon's conception of the contractualist situation as follows:

> we imagine ourselves both as legislators and citizens, creating the principles by which we will then govern ourselves. We imagine that we are symmetrically situated, that each of us has a veto, and that we must come to some kind of reasonable agreement. The principles of morality, as Scanlon understands it, are the principles that we would agree to in this contractualist situation. They are thus the terms of self-governance adopted by those who recognize each other as having a symmetric standing to determine the terms of their mutual self-governance. They are, we might say, the principles that would be agreed to in a Kingdom of Equals, each of whom is committed to living in a kind of harmony with the rest and so accords to each one a symmetric standing in determining the terms of his or her own self-governance. (Hieronymi 2011, 106)

[117] Watson (2011, 316).

treatment of the case of the psychopath is in tension with his overall contractualist moral theory. I will argue that the view suffers from an internal inconsistency. Given Scanlon's contractualism and given his account of moral blame, he should not hold that psychopaths may be morally blameworthy or the proper objects of moral blame.

My criticism is importantly different from both Watson's and Wallace's. Wallace and Watson each take issue with a claim of Scanlon's (that reactive attitudes are not essential to blame, that psychopaths may be morally responsible) on the basis of a conceptual claim that he does not hold to be true (that blame essentially involves the reactive attitudes, that moral responsibility essentially involves interpersonal accountability). It is open to Scanlon to rebut their criticisms by denying the conceptual claims on which they rest. But it is not open to Scanlon to respond to my criticism in like manner. My challenge appeals only to the commitments of Scanlon's own view. I will argue that his claims about psychopaths are inconsistent with his other commitments. In order to meet my challenge, Scanlon would have to either show that the relevant commitments are really not inconsistent or else give up some of them in favor of others. Accordingly, I aim to establish that this first avenue of response is closed and also to show that, of the claims Scanlon might give up in order to dissolve the inconsistency, his claims about the moral responsibility of psychopaths are the

most peripheral. Hence, I will recommend that Scanlon revise his view by giving them up.

3. The Contractualist Conception

The moral relationship grounds Scanlon's account of moral blame in the sense that (i) it "provides the standards relative to which the attitudes that an agent's action reveals constitute an impairment" and (ii) "[t]hese standards also determine the appropriateness of various responses to this impairment."[118] As noted in §1, it is not only Scanlon's account of moral responsibility that is tightly connected with the moral relationship. His accounts of wrongness and moral motivation are connected to each other via this relationship as well. In this section, I will argue that the moral relationship is grounded in a particular conception of moral agency and that this conception is foundational to Scanlon's thinking about morality in general.

Scanlon is explicit that he takes justifiability to others[119] to be "basic" and reasons to be "primitive." These two notions are foundational to the contractualist framework he develops for thinking about morality.

---

[118] Scanlon (2008, 138).

[119] It is significant that the notion of justifiability Scanlon takes to be basic is justifiability *to others*. The notion of justifiability is the notion of offering reasons in support of something. It may be characterized in terms of a three-place relation, where two places take persons and one place takes reasons. We can distinguish between *intra*personal and *inter*personal justifiability as follows. In the case of intrapersonal justifiability one person provides the reasons to herself—the same person takes both person-places. In the case of interpersonal justifiability one person provides reasons to another—the person-places are taken by distinct individuals. The claim that

A reason, according to Scanlon, is a consideration that counts in favor of something. But he distinguishes between two senses of "reason." The primary notion is of a reason in the "standard normative sense." A reason in this sense just is "a *good* reason—a consideration that really counts in favor of the thing in question."[120] An "operative reason," and this is the second sense of "reason," is one that a particular agent takes to be a good reason.[121] So we can sensibly say that *A*'s reason for *p* was no reason at all or that a given reason could not have been *A*'s reason for *p*. In talking this way, we are trading between the two senses of "reason." In the first claim, the word "reason" is first used in the operative sense and then in the standard normative sense; in the second claim the word "reason" is first used in the standard normative sense and then in the operative sense. I will return to this distinction, and especially the claim that a normative reason could not have been an operative reason for a given agent, in §7. For now, it is enough to note that the notion of a reason is foundational to Scanlon's moral

---

justifiability to others is basic can be recast as the claim that we understand intrapersonal justifiability on the model of interpersonal justifiability—intrapersonal justification is a special case in which the same person takes both person-places. This is a substantive claim with interesting implications. Perhaps one implication can be seen in Scanlon's (1982) criticism of Rawls' version of contractualism. There Scanlon registers suspicion about the reduction of choice behind the veil of ignorance to the choice of a single individual. In contrast to Scanlon's view, one might claim that justifiability to oneself is basic, understanding interpersonal justification as a special case where the person-places are occupied by distinct individuals. Perhaps this latter claim is implicit in Korsgaard's view, with the caveat that an intrapersonal justification must be in principle interpersonally intelligible.

[120] Scanlon (1998, 19).

[121] Compare Dancy's (2000) distinction between "motivating reasons" and "normative reasons;" also Schroeder's (2008) distinction between "subjective reasons" and "objective reasons."

theory because he assumes it in articulating his distinctive account of what it is for an action to be wrong.

On Scanlon's account, what it is for an action to be wrong is for it to be justified only by principles that could be reasonably rejected by people with the right kind of motivation. A principle is reasonably rejectable if there is sufficient reason[122] for a generic individual, situated as an equal with others who are also motivated to agree on principles for the general regulation of behavior, to reject it. The picture here is of a group of similarly motivated individuals with equal standing offering each other reasons in favor of and against candidate principles for the regulation of behavior between them.[123] The notion of a reason is basic relative to Scanlon's notion of wrongness because reasons function as a currency for justification (and, conversely, rejection) and wrongness is a property of actions that could not be justified in the relevant way to the relevant individuals.

This brings us to the other notion foundational to Scanlon's contractualist moral framework: justifiability to others. Scanlon tells us that his view

> holds that thinking about right and wrong is, at the most basic level, thinking about what could be justified to others on grounds that they, if appropriately motivated, could not reasonably reject. On this view the

[122] More precisely, these reasons must be "generic" and "personal". See (Scanlon 1998, 204 and 219) for these claims, respectively. (Hieronymi (2011, 108) notes this qualification.) This does not mar the point in the text. Certainly the notion of a reason is more fundamental than the notions of a generic reason or a personal reason.

[123] See the quotation from Hieronymi in footnote 116, above, for a nice description of the contractualist situation.

> idea of justifiability to others is taken to be basic in two ways. First, it is by thinking about what could be justified to others on grounds that they could not reasonably reject that we determine the shape of the more specific moral notions such as murder or betrayal. Second, the idea that we have reason to avoid actions that could not be justified in this way accounts for the distinctive normative force of moral wrongness.[124]

In other words, the notion of justifiability to others grounds both the content and characteristic force of particular moral claims. For example, it is a commonplace that murder is wrong but not that not all acts of killing are murder. On Scanlon's view, we determine which acts of killing are murder, and so wrong, by appeal to the notion of justifiability to others. Those acts of killing that are allowed only by principles that could not reasonably be justified to suitably motivated persons are wrong, and we label them acts of "murder." We appeal to the same notion to explain the characteristic force of the judgment that a given act was an act of murder. It is because the act of murder is unjustifiable to others that we experience our reasons not to engage in it as so strong.

---

[124] Scanlon (1998, 5). And compare: "What is distinctive about my version of contractualism is that it takes the idea of justifiability to be basic in two ways: this idea provides both the normative basis of the morality of right and wrong and the most general characterization of its content" (Scanlon 1998, 189). And also:

> … Scanlon simply notes, in effect, that it is plausible that we owe it to each other (in some pretheoretical sense) to grant to one another standing to partially determine the terms of our mutual self-governance, so long as such standing is exercised consistently with each the standing of each to do the same [*sic*]. Further, and crucially, Scanlon thinks we owe *only* this to one another. That is to say, we do not, in constructing these moral principles, appeal to any other, prior or independent, moral standard (though we may appeal to moral principles established in some other iteration of the holistic contractualist method). Rather, Scanlon identifies, as the moral standard, *whichever* principles no one could reasonably reject, if we were all committed to finding such principles. Again, as he puts it, justifiability is basic. His is, so to speak, a minimalist account. (Hieronymi 2011, 119)

We can articulate a conception of moral agency that combines the two notions that are Scanlon's starting points for thinking about morality: *justification of one's actions to others in terms of reasons is fundamental to moral agency*. Call this *the contractualist conception*. I take it that the contractualist conception captures the background picture of moral agency assumed by Scanlon's moral theory. When we think about agency in the context of morality, on Scanlon's view, we are thinking in terms of justifiability to others in terms of reasons.[125]

The contractualist conception is closely related, but not identical, to the evaluator conception—that justification in terms of values is fundamental to human agency. I think it is apt to consider the contractualist conception as derivable from the evaluator conception. Let me explain why. One difference between the two conceptions has to do with the form of justification they invoke. The contractualist conception is stated explicitly in terms of interpersonal justification, whereas the evaluator conception is not. A form of agency exhibited in intrapersonal justification of action would satisfy the evaluator conception, but perhaps not the contractualist conception.[126] Thus, the notion of justification

[125] Notice that this formulation leaves it open that a different background conception may be applicable in contexts other than thinking about morality. For example, it may be that when we think about issues of criminal punishment we conceive of human beings somewhat differently than when we think about issues of morality, or it may be that when we think about self-governance or weakness of will we conceive of human agents somewhat differently.

[126] Perhaps Korsgaard's Practical Identity theory can provide an example of a view that satisfies the evaluator conception, but perhaps not the contractualist conception. On that view, reasons and obligations issue from the agent's practical identities, which are descriptions under which the agent values herself. Justified actions issue from motivations endorsed by the agent's practical identities because she has reasons, and may be obligated, to perform these actions. But if a

invoked in the contractualist conception is stronger than that invoked in the evaluator conception. It is apt to consider the former as a moralized version of the latter. This is why I have referred to it as a conception of moral agency, as opposed to human agency. There are aspects of human agency that are not essentially moral. The evaluator conception may aim to capture these, but the contractualist conception does not. This is fine insofar as the contractualist conception is taken to be in the background of a moral theory, as it is here.

There is a possible second difference between the evaluator conception and the contractualist conception, depending on one's metaethical view about the relation between reasons and values. Scanlon's is a buck-passing account of value: "to call something valuable is to say that it has other properties that provide reasons for behaving in certain ways with regard to it."[127] On a buck-passing account, claims about values are transparent to claims about reasons. This is one reason why the justification that is the focus of the contractualist conception is put in terms of reasons. The justification that is the focus of the evaluator conception is put in terms of values. This difference in terminology may mark a difference in conceptions, depending on whether or not one accepts a buck-passing account of value.

---

practical identity is a description under which the agent values herself, then there is no appeal to interpersonal justification as such. Intrapersonal justification will suffice. (Although, on Korsgaard's view, intrapersonal justification is, at least in principle, interpersonally valid because the normativity of practical reasons is essentially relational. See her (1996a, 136-145).)

[127] Scanlon (1998, 96).

If one is a buck-passer about value, then one will be inclined to read the evaluator conception—put in terms of values—as transparent to a conception of agency put in terms of reasons. That is, one will have read the evaluator conception in a buck-passing way all along. The reasons terminology in the contractualist conception will just make explicit the way in which one has all along understood the evaluator conception. The only difference between the two conceptions will be that the former is moralized in the way described in the previous paragraph.

But if one is not a buck-passer about value, then one may take the reasons terminology in the contractualist conception to mark a genuine difference between it and the evaluator conception. A non-buck-passer about value may hold that a claim about reasons depends on a claim about values—for example, that a consideration is a reason to *a* only if it shows that *a*-ing is good in some way. Thus, one may hold that the contractualist conception is derivable from the evaluator conception, not only in the sense that it requires a stronger sense of justification, but also in the sense that it calls for the currency of justification to be reasons, and the notion of a reason is dependent on the notion of value.[128]

---

[128] Scanlon accepts the buck-passing account of value, in part, in response to Moore's open-question argument. If we take claims about value to be transparent to claims about reasons to behave in certain ways, we can explain why questions of the form "'X is P, but is it good?' where 'P' is a term for some natural or metaphysical property" have an "open feel." These questions ask whether the specified property is a reason for behaving in certain ways towards the object that has it. They call for one to "draw a practical conclusion." And even if one thinks that one should draw the relevant conclusion, "just saying that something has these properties does not involve drawing it." (Scanlon 1998, 96-7). But one might think that Scanlon has only served to explain the

My claim is that the contractualist conception is a background picture of moral agency that animates Scanlon's thinking about agency in the context of morality. In the next section, I will argue that understanding Scanlon's thought in this way can help us to make sense of his claims about the abilities required to be the proper object of moral assessment, including judgments of blameworthiness and moral blaming responses. Before turning to that argument, however, I would like to consider a more direct way in which the contractualist conception might bear on these judgments and responses.

As we have seen, Scanlon's thinking about morality essentially involves the moral relationship. This relationship is central to his accounts of wrongness, moral motivation and moral blame. My claim that the contractualist conception is in the background of Scanlon's thinking about morality may be put in terms of

---

open feel of questions about value in terms of a notion, reasons, that, on his understanding, admit of questions that feel open in their own way. Scanlon begins his book with the following assertion:

> I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. "Counts in favor how?" one might ask. "By providing a reason for it" seems to be the only answer. (Scanlon 1998, 17).

Perhaps his buck-passing account of value has blinded Scanlon to what seems to be an obvious answer to the question "Counts in favor how?"—namely, "By showing it to be good in a way." This is the answer suggested by my characterization of the view of one who is a non-buck-passer and accepts the evaluator conception. But there are other ways. One who accepts the explainer conception might reply that a consideration counts in favor of an action "By showing it to make causal-explanatory sense." My aim is not to argue against the buck-passing account of value, but rather to suggest that an appeal to the open-question argument does not decisively favor this account. To be fair, Scanlon argues that the buck-passing account of value is also supported by intuitions about reasons to respond to the valuable. But these arguments do not strike me as decisive either. In the least, Scanlon's arguments do not engage with the claim that there are ways of conceiving of what is fundamental to human agency that can provide substantive answers to the question how a reason counts in favor of an action.

its articulating an assumed condition on who can stand in the moral relationship. Scanlon's thinking about morality assumes that only those agents who satisfy the contractualist conception—only those capable of justifying their actions to others in terms of reasons—can stand in the moral relationship. This amounts to articulating a condition on who may be the proper object of judgments of moral blameworthiness and moral blaming responses. Such judgments and responses are conditioned by the moral relationship. If satisfying the contractualist conception is a condition on standing in the moral relationship, and if standing in the moral relationship is a condition on being the proper object of judgments of moral blameworthiness and moral blaming responses, then if one does not satisfy the contractualist conception, one is not a proper object of judgments of moral blameworthiness or moral blaming responses.

4. Abilities and Moral Assessment

In this section, I want to consider another way of understanding the importance for Scanlon's account of moral responsibility of my claim that the contractualist conception is behind Scanlon's thinking about morality. I will argue that my claim about the contractualist conception is helpful in getting clear on what Scanlon says about the abilities required to be the proper object of moral assessment. In the next section, §5, I will argue from the results of this and the previous section that Scanlon's own view commits him to the claim that

psychopaths are not morally responsible for what they do. This is interesting because it contradicts his explicit claims to the contrary, which I consider in §6.

Moral judgments do not apply to just anyone. For example, we do not think that young children or human beings with certain mental disorders are morally responsible for what they do. A plausible account of moral responsibility must make a principled distinction between those who are and those who are not proper objects of moral assessment. Here is how Scanlon's contractualism makes the distinction.

> On this view, moral judgments apply to people considered as possible participants in a system of co-deliberation. Moral praise and blame can thus be rendered inapplicable by abnormalities which make this kind of participation impossible.[129]

Young children and those with certain mental disorders are not properly subject to moral assessment for their behavior, on Scanlon's view, because they are not (yet or anymore) members of the Kingdom of Equals. They cannot (yet or any longer) participate in the determination of reasonable principles for the general regulation of behavior.

If membership in the Kingdom of Equals is Scanlon's criterion for being a proper object of moral assessment, we should want to know what such membership requires.

---

[129] Scanlon (1986, 167).

> According to contractualism, thought about right and wrong is a search for principles "for the regulation of behavior" which others, similarly motivated, have reason to accept. What kind of "regulation" is intended here? Not regulation "from without" through a system of social sanctions but regulation "from within" through critical reflection on one's own conduct under the pressure provided by the desire to be able to justify one's actions to others on grounds they could not reasonably reject. This idea of regulation has two components, one specifically moral, the other not. The specifically moral component is the ability to reason about what could be justified to others. The nonmoral component is the more general capacity through which the results of such reasoning make a difference to what one does. Let me call this the capacity for critically reflective, rational self-governance—"critically reflective" because it involves the ability to reflect and pass judgment upon one's actions and the thought processes leading up to them; "rational" in the broad sense of involving sensitivity to reasons and the ability to weigh them; "self-governance" because it is a process which makes a difference to how one acts.[130]

To be a member of the Kingdom of Equals, on Scanlon's view, one must have both the specifically moral ability to reason about what could be justified to others and the general capacity for critically reflective, rational self-governance.

What is the relation between this moral ability and this general capacity? Since Scanlon distinguishes between them, it should be safe to assume they are distinct. But then what more is required than the general capacity in order to possess the moral ability? It is not easy to glean Scanlon's answer to this question from what he says. He tells us that the moral ability is the ability to reason about what could be justified to others, but he does not tell us how this differs from the capacity for critically reflective, rational self-governance. However, Scanlon does

[130] Scanlon (1986, 173-4).

make some suggestive remarks about the general capacity and its relation to morality and moral motivation. Beginning from these remarks, we can tease out Scanlon's commitments regarding the difference between these jointly necessary conditions on membership in the Kingdom of Equals.

To begin with, Scanlon is clear that to think of one as subject to the demands of morality is to presuppose that one has the general capacity.

> This general capacity for critically reflective, rational self-governance is not specifically moral, and someone could have it who was entirely unconcerned with morality. Morality does not tell one to have this capacity, and failing to have it in general or on a particular occasion is not a moral fault. Rather, morality is addressed to people who are assumed to have this general capacity, and it tells them how the capacity should be exercised. The most general moral demand is that we exercise our capacity for self-governance in ways that others could reasonably be expected to authorize. More specific moral requirements follow from this.[131]

When we consider the actions of agents in moral terms, we assume that those whose actions we are considering have the general capacity for critically reflective, rational self-governance. And one might have this capacity without the characteristic contractualist motivation to find principles that no one can reasonably reject. The capacity is, we might say, prior to and independent of morality.

[131] Scanlon (1986, 174).

But not everyone has this general capacity. One may be exempt from moral assessment precisely because one lacks it.[132] For example, we do not normally consider young children, those subject to posthypnotic suggestion and the mentally ill to be morally responsible.

> It is important to our reactions in such cases, however, that what is impaired or suspended is a *general* capacity for critically reflective, rational self-governance. If what is "lost" is more specifically moral—if, for example, a person lacks any concern for the welfare of others—then the result begins to look more like a species of moral fault.[133]

Thus, on Scanlon's view, while morality does not require one to have the general capacity for critically reflective, rational self-governance (moral thought assumes that one has this), morality does require concern for others. A lack of such concern would suggest a moral fault.

This last remark entails, on Scanlon's view, that concern for others is not what separates one who merely has the general capacity from one who has both the general capacity and the moral ability. Scanlon claims that, on the assumption that one has the general capacity, a lack of concern for others suggests a moral fault. But to attribute a moral fault is to make a moral judgment. And moral judgments, on his view, apply only to those who have both the

[132] See Watson (1987) for discussion of a distinction between excuses and exemptions.

[133] Scanlon (1986, 175).

general capacity and the moral ability. It follows that one can have both the moral ability and the general capacity while lacking concern for others.

It seems that possession of the moral ability must require something over and above possession of the general capacity, or else the moral ability would reduce to the general capacity and there would be no need to distinguish between them as Scanlon does. But the added element cannot be a concern for others, or else to lack this concern could not be a moral fault. So what is it that one must possess over and above the general capacity for critically reflective, rational self-governance in order to possess the moral ability to reason about what could be justified to others? Here is a proposal: the ability to recognize and respond to the force of moral reasons.

Admittedly, I do not find this proposal in Scanlon's own words. But it gains support from what he says. First, it preserves Scanlon's distinction between the general capacity and the moral ability. Second, it preserves the suggested relation between them—namely, that the moral ability presupposes the general capacity. Third, it is suggested by the contractualist conception, which, I argued in the previous section, is behind Scanlon's thinking about moral agency. And, fourth, it allows that lacking concern for others might be a moral fault, because it allows that one might possess the moral ability without actually being concerned for others. Let me take these points in turn.

The ability to recognize and respond to the force of moral reasons presupposes the general capacity for critically reflective, rational self-governance. For one to be able to recognize and respond to the force of moral reasons in particular, one must be capable of recognizing and responding to the force of reasons in general. But it goes beyond the general capacity because it requires not just sensitivity to reasons in general, but sensitivity to a particular class of reasons—namely, moral reasons.

Scanlon does not make it clear what exactly a moral reason is on his view. He claims that there are agents who cannot see the force of moral reasons.[134] So he is committed to the existence of moral reasons and to a distinction between the force of moral reasons and the force other kinds of reasons. But he does not provide an explicit characterization of moral reasons.

Nevertheless, I think Scanlon is committed to the following. The force of a moral reason, on his view, is at least partially generated from within the Kingdom of Equals. For example, the force of a moral reason not to stomp on my foot is generated by the procedure of determining that this action would be allowed only by principles that could be reasonably rejected by suitably motivated individuals in pursuit of principles for the general regulation of behavior. This is what distinguishes the force of a moral reason from the force of a reason of another kind. The force of a moral reason is related to the

[134] See, e.g., Scanlon (1998, 288). I quote the relevant passage in §9, below.

determination in the Kingdom of Equals that some action would be wrong in the sense that this determination provides the reason with greater force than it would have if the action were not wrong. The force of a non-moral reason is not so related to wrongness.

There are at least two kinds of consideration the force of which could be related to the determination that some action is wrong in the way characteristic of moral reasons. The first is just the consideration that some action is wrong. For example, the consideration that stomping on my foot is wrong is a reason not to stomp on my foot, and it is clear how the force of this reason is related to the determination that stomping on my foot is wrong. So a moral reason not to *a* may just be the consideration that *a*-ing is wrong. The second is a consideration that factors in the determination that some action is wrong. For example, the consideration that stomping on my foot would cause me pain is a reason not to stomp on my foot, and, given that stomping on my foot is wrong, this reason has greater force than it otherwise would. To see this, contrast a case in which stomping on my foot would be wrong with a case in which it would not. For example, suppose that there is a poisonous spider on my shoe and stomping on my foot would kill the spider and save me from its bite.[135] In this case, it is plausible that stomping on my foot would be permissible. The pain it would cause me is still a reason not to stomp on my foot, but this reason is not so

[135] I borrow this case from Scanlon (1998, 279) and discuss it again in §9, below.

forceful that you should not do it. The force of other considerations that speak in favor of stomping on my foot is greater. This is in contrast to a case in which you simply want to see me grimace. In this case, stomping on my foot would be wrong and the consideration that it would cause me pain is more forceful because of this.[136]

I think Scanlon is committed to claiming that, in the case where stomping on my foot is wrong, both the consideration that stomping on my foot is wrong and the consideration that it would cause me pain are moral reasons not to stomp on my foot. What makes them moral reasons is that their force is related to the determination that this action is wrong. One important difference between these two moral reasons is that the first, but not the second, is always a moral reason. The consideration that an action is wrong is essentially related to wrongness. But the consideration that stomping on my foot would cause me pain may not be related to wrongness, as evidenced by the case where there is a spider on my shoe.

---

[136] Compare Hieronymi's discussion of Scanlon's reply to an objection by Judith Jarvis Thomson. Thomson claims that what makes it wrong to torture babies is not that this action would be allowed only by reasonably rejectable principles. Scanlon replies that there is, on his view, a distinction between what *makes* an action wrong and what it is for an action to *be* wrong.

> Scanlon here, in effect, draws attention to the fact that his is a "two-level" view, in which wrongness provides a "higher-order" reason. For an action to *be* wrong, according to Scanlon, is for it to be in violation of principles that no one could reasonably reject, etc. But, for an action to be wrong, there must be other, strong, "lower-order" reasons that count against it—other reasons that provide winning grounds for rejecting any principle that would allow the action. (Hieronymi 2011, 113)

I am here indebted to Hieronymi's illuminating discussion, though what I say goes beyond what she says.

My proposal is that what separates the agent who possesses only the general capacity for critically reflective, rational self-governance from the agent with the moral ability to reason about what could be justified to others is that the former lacks the ability to recognize and respond to the force of moral reasons. He cannot recognize and respond to the force of the consideration against stomping on my foot that it is wrong, nor the full force of the consideration that this would cause me pain in contexts where stomping on my foot would be wrong. This agent can, however, recognize and respond to the full force of the consideration against stomping on my foot that it would cause me pain, where this force is not related to the determination that this action would be wrong, and to at least some of the force of this consideration in contexts in which it would be wrong to stomp on my foot. What separates him from an agent with the moral ability is that his ability to recognize and respond to this consideration would not change between the two cases considered above. He would recognize and respond to this consideration as if it had the same force whether there was a spider on me shoe or not. The agent with the moral ability, by contrast, would recognize and respond to this reason not to stomp on my foot as if it had different force in these different contexts.

The proposal that what separates the moral ability from the general capacity is the ability to recognize and respond to the force of moral reasons makes sense given the claim that the contractualist conception is in the

background of Scanlon's thinking about morality. The force of moral reasons is related to considerations about what could be reasonably justified to others. If justifying one's actions to others in terms of reasons is fundamental to moral agency, then it would make sense to distinguish between a general capacity sufficient for rational agency and a specifically moral ability required for moral agency on the basis of the ability to recognize and respond to the force of moral reasons. Those who possess the moral ability, but not those who possess only the general capacity, would satisfy the contractualist conception. Thus, the contractualist conception may be seen to determine Scanlon's understanding of who (in the context of thinking about morality)[137] counts as a rational agent. Justifying one's actions to others in terms of reasons is fundamental to rational (moral) agency. And this presupposes that the rational (moral) agent be able to recognize and respond to the force of moral reasons.

Notice that the ability to recognize and respond to the force of moral reasons need not presuppose that one is actually concerned for others. Any analysis of the notion of ability that does not make the ability to recognize and respond to considerations related to wrongness (in Scanlon's sense) depend on actual concern for others would be consistent with an agent having the ability to recognize and respond to moral reasons and lacking concern for others. For example, if we understand the ability to recognize and respond to moral reasons

[137] This leaves it open that, in other contexts, Scanlon might conceive of rational agency differently. My claim is about how Scanlon thinks about morality.

along the lines of[138] Fischer and Ravizza's account of "moderate reasons-responsiveness," then the ability would require "regular reasons-receptivity"—that the agent "exhibit an understandable pattern of reasons-recognition"—and "weak reasons-reactivity"—that the agent react to "*some* incentive to (say) do other than he actually does."[139] It would be enough for regular receptivity to moral reasons that the agent exhibits an intelligible grasp of the notion of reasonable rejectability, say, by identifying an understandable pattern of wrong actions when asked.[140] She might do so without actually being concerned for me because, say, she has a purely "intellectual" grasp of moral requirements. It would be enough for weak reactivity to moral reasons that there be some possible world in which the agent would react to a moral reason to do otherwise than she actually does—for example, there is some possible world in which the agent who actually steps on my foot refrains from doing so because it would be wrong. It may be the case that she would be concerned for me in this possible world, and that may even be the explanation of why she would react to this

---

[138] One important difference between the way I put things in the text and Fischer and Ravizza's account is that theirs is "mechanism-based," whereas the account in the suggestion in the text is "agent-based." This difference, however, does not mar the point that their account shows that we can give an analysis of ability to recognize and respond to moral reasons that does not presuppose concern for others.

[139] Fischer and Ravizza (1998, 75).

[140] See Fischer and Ravizza (1998, 71-2) for the "imaginary interview" test for determination of whether an agent's pattern of reasons-recognition is "understandable." I should note that I am not claiming that Scanlon endorses Fischer and Ravizza's account. Rather, I am claiming that this is one account of the ability to recognize and respond to reasons on which the ability to recognize and respond to moral reasons would not presuppose actual concern for others.

moral reason in that world. But this would not change the fact that she is not actually—in the actual world—concerned for me. So she may be weakly reactive to moral reasons without actually being concerned for me.

I have been proposing that what separates the moral ability to reason about what could be justified to others from the general capacity for critically reflective, rational self-governance is the ability to recognize and respond to the force of moral reasons. This proposal preserves the implied distinction between the moral ability and the general capacity and the implied relation between them, is consistent with Scanlon's claim that lack of concern for others can be a moral fault and makes sense given the claim that the contractualist conception is behind Scanlon's thinking about morality. I take these to be considerations in its favor.

We were led to consider what separates the moral ability to reason about what could be justified to others from the general capacity for critically reflective, rational self-governance in the light of two claims of Scanlon's. The first claim was that moral assessment applies only to possible members of the Kingdom of Equals. The second claim was that membership in the Kingdom of Equals requires both the moral ability and the general capacity. If we accept my proposal that what must be added to the general capacity to possess the moral ability is the ability to recognize and respond to the force of moral reasons, it follows that moral assessment applies only to those who can recognize and

respond to the force of moral reasons. Since my focus in this paper is on Scanlon's account of moral blame, the relevant moral assessments are judgments of moral blameworthiness and moral blaming responses. So the more particular conclusion that interests me is that judgments of moral blameworthiness and moral blaming responses apply only to those who can recognize and respond to the force of moral reasons.

5. Psychopaths and Moral Responsibility I

I have been arguing that Scanlon's thinking about morality is determined by the contractualist conception of moral agency, according to which justification of one's actions to others in terms of reasons is fundamental. And I have articulated two ways in which this claim may present a condition on who may be the proper object of judgments of moral blameworthiness and moral blaming responses. The first condition is that one must satisfy the contractualist conception. The second condition is that one must be able to recognize and respond to the force of moral reasons. In this section, I will turn to the question whether psychopaths may be properly judged morally blameworthy or properly morally blamed for what they do. I will argue that, given Scanlon's own understanding of psychopathy, it follows rather straightforwardly that psychopaths do not satisfy either of the two conditions just articulated. Thus, they are not proper objects of judgments of moral blameworthiness or of moral

blaming responses. In the next section, I will consider Scanlon's explicit claim to the contrary.

According to Scanlon, recall, psychopaths "may be rational in a general sense, and capable of means-ends reasoning, but nonetheless unable to understand why they have any reason to take moral requirements seriously as limits on the pursuit of their aims."[141] The psychopath so described fits the profile of an agent who has the general capacity for critically reflective, rational self-governance. He can critically reflect on his judgments and actions, and he can recognize, weigh and react to reasons. But the psychopath, so understood, lacks the concern for others morality requires. He also lacks the ability to recognize and respond to the force of moral reasons. One who cannot understand why he has any reason to let moral requirements constrain his aims must not be able to recognize the force of moral reasons. Thus, Scanlon's understanding of psychopaths entails that they do not satisfy the condition on who can be a proper object of moral assessment considered in the previous section.

Scanlon's understanding of psychopaths also entails that they do not satisfy the contractualist conception. Scanlon understands psychopaths to be rational agents, but not rational agents who can understand why moral requirements—considerations that result from a collection of individuals offering

[141] Scanlon (ms., 11).

reasons to each other in order to show that particular actions are not justifiable—should constrain their aims. But if psychopaths are unable to grasp considerations that result from a process of interpersonal justification in terms of reasons as reasons, then they do not exhibit a form of rational agency to which justifying one's actions to others in terms of reasons is fundamental.[142] Thus, psychopaths do not satisfy the contractualist conception or the condition on standing in the moral relationship that follows from it.

Because they do not satisfy either of these conditions, I conclude that, on Scanlon's view, psychopaths are not properly judged morally blameworthy for their actions and are not appropriate objects of moral blaming responses.

6. Psychopaths and Moral Responsibility II

Scanlon, however, claims that, to the contrary, psychopaths may be properly judged morally blameworthy for their actions and may be appropriate objects of moral blaming responses. After giving the above quoted characterization of psychopaths, he continues: "If they see no reason not to kill, injure, or manipulate us when this promotes their ends, then this judgment about reasons is attributable to them. It is their considered judgment about the reasons

---

[142] Compare:

> To elaborate, and to preview the argument, psychopathy (I am assuming) involves the incapacity to engage with others as individuals with the standing to object when their interests and concerns are disregarded as unimportant. The mutual recognition of this standing is, in my view (and I take it in Scanlon's), what morality is fundamentally about. (Watson 2011, 308-9)

they have."[143] The idea here is that if an agent has the general capacity for critically reflective, rational self-governance, then we can attribute judgments about reasons to that agent on the basis of her actions. If, for example, she stomps on my foot, we can attribute to her the judgment that the fact that this will cause me pain is not a reason not to do it.

Scanlon claims that it follows from this idea that, when we consider the significance of their actions, we should group together agents with the moral ability to reason about what could be justified to others with those who possess only the general capacity for critically reflective, rational self-governance and oppose them to agents who lack the general capacity.

> If a creature cannot make judgments about whether anything matters, it cannot judge that harm to us does not matter, and its actions cannot reflect such judgments. By contrast, a rational creature who fails to see the force of moral reasons—who fails, for example, to see any reason for being concerned with moral requirements at all or with the justifiability of its actions to others—can nonetheless understand that a given action will injure others and can judge that this constitutes no reason against so acting. So the actions of such a creature would have implications for its relations with others that are at least very similar to (if not identical with) those of an agent who understood the relevant moral reasons but simply rejected them.[144]

---

[143] Scanlon (ms., 11). And compare: "A person who is unable to see why the fact that his action would injure me should count against it still holds that this *doesn't* count against it" (Scanlon 1998, 288). Compare also: "I do not think that blame is undermined by the fact that a person had no control over the factors that made him the kind of person that he is, or by the fact that, given the kind of person he is, he is incapable of understanding the reasons against acting the way he does" (Scanlon 2008, 178).

[144] Scanlon (1998, 288).

Scanlon seems to be saying here that, on his view, all that is required for an agent to be morally blameworthy for his actions is that he have the general capacity for critically reflective, rational self-governance. Given this capacity, even if one cannot recognize and respond to moral reasons, one's actions have something like the significance (or even the same significance) as the actions of one who can recognize and respond to moral reasons. Given Scanlon's understanding of psychopaths, this entails that psychopaths may be morally blameworthy for their actions because they may show improper concern toward other rational beings.

My argument for the claim that, on Scanlon's view, psychopaths are not proper objects of judgments of moral blameworthiness focused on the conditions for such judgments. Scanlon's claim that psychopaths may be morally blameworthy, by contrast, focuses on the significance of their actions. Perhaps this difference in focus has led me to miss the relevant commitments of Scanlon's view.

I do not, however, think this is the case, and, in the following two sections, I will argue that, on Scanlon's view, psychopaths' actions cannot have the significance he claims they do. Thus, his focus on the significance of their actions does not establish that psychopaths are properly judged morally blameworthy or properly subject to moral blaming responses. It is worth noting that, though it aims at the same conclusion, this argument will be independent of the claims of

the preceding sections. So one who has found my arguments to this point lacking may yet find reason to agree with me.

7. Impairing the Moral Relationship

In this section, I will appeal to two independent distinctions, introduced already and to which Scanlon is committed, between kinds of reasons: the distinction between operative reasons and normative reasons and the distinction between moral reasons and non-moral reasons. I will argue that, given these distinctions and given Scanlon's account of what impairment of the moral relationship consists in, psychopaths cannot act in ways that impair the moral relationship. Given that judgments of moral blameworthiness are judgments that one has impaired the moral relationship, it follows that psychopaths' actions do not make them proper objects of such judgments.

Scanlon defines the "meaning" of an action as "the significance, for the agent and others, of the agent's willingness to perform that action for the reasons he or she does."[145] And he is explicit about the connection between the meaning of actions and his account of blame.

> To say that an action is blameworthy is to make a claim about its meaning: to claim that the action indicates something about the agent's attitudes that impairs his or her relations with others. To blame someone, in my

[145] Scanlon (2008, 4).

> view, is to understand one's relations with that person as modified in the way that such a judgment holds to be appropriate.[146]

To claim that an agent is morally blameworthy for an action is to claim that the agent performed the action for reasons that impair his moral relations with others.

> It is relatively easy to say what this type of impairment consists in. It occurs when a person governs him- or herself in a way that shows a lack of concern with the justifiability of his or her actions, or an indifference to considerations that justifiable standards of conduct require one to attend to.[147]

This is a disjunctive account of the conditions under which impairment of the moral relationship occurs. It is worth considering each condition in isolation. We will see that it is clear that psychopaths cannot meet the first condition, that the second condition is crucially ambiguous and that, even given the ambiguity, Scanlon's view does not support the claim that psychopaths can meet it.

Scanlon's first condition on impairment of the moral relationship—that impairment occurs when one "shows a lack of concern with the justifiability of his or her actions"—is naturally interpreted as requiring that the agent judge that it does not matter whether her action is justifiable to others. The invocation of a judgment about justifiability is important here. It would be too weak a condition

[146] Scanlon (2008, 6).

[147] Scanlon (2008, 141).

if interpreted straightforwardly as involving mere lack of concern. Animals, such as tigers, exhibit lack of concern with the justifiability of their actions. But, since we do not expect this concern of them to begin with, their lack of it is not robust enough to support any claim that they have impaired their relationship with us.[148] It is plausible that we do not expect this concern of tigers (and other non-human animals) because we suppose they are incapable of exhibiting it.[149] Our understanding of the significance of their behavior does not warrant the attribution to them of judgments about the justifiability of their actions. In order to distinguish between the lack of concern with the justifiability of her actions shown by a moral agent who does something blameworthy and the lack of concern shown by a tiger, it is natural to interpret this first condition as involving appropriate attribution of a judgment about the importance of the justifiability of one's actions.

Now consider the second condition—that impairment of the moral relationship occurs when one exhibits "indifference to considerations that

---

[148] On Scanlon's view, we can blame non-human animals so long as we have a grounding relationship with them. See Scanlon (2008, 166). I assume this blame would not be moral blame because we do not stand in the moral relation with non-human animals. Compare:

> Blame as I interpret it has this aspect of condemnation because it involves withholding trust, cooperation and so on from a person *because of* attitudes that person holds that are faulty by the standards of some relationship to which he or she is a party. This explains why it is not a form of blame to withhold trust from a tiger. (Scanlon ms., 14)

This comment presupposes that we do not stand in *any* relationship with tigers. The presupposition may be false of some of us—for example, Roy Horn (of Siegfried and Roy).

[149] This is also a plausible explanation for why we do not take ourselves to stand in any relationship with them.

justifiable standards of conduct require one to attend to." We can see that this condition is ambiguous. Recall the distinction, discussed in §4, between the force of a moral reason not to perform some action—for example, the force of the consideration that stomping on my foot would cause me pain when this action would be wrong—and the force of this very same consideration against a permissible action—say, when you justifiably think there is a spider on my shoe. The claim is that the consideration—stomping on my foot would cause me pain—has greater force when stomping on my foot would be wrong. This second condition is ambiguous because we might interpret "considerations" to include the consideration that some action is wrong. Indifference to the force of a moral reason may be thought to include indifference to this consideration. But if the condition includes this consideration, then one who, like a psychopath, is unable to reason about what could be justified to others, could not grasp it. Alternatively, we could interpret "considerations" to include only the particular considerations that count against wrong actions, and not also the consideration that these actions are wrong. Psychopaths can grasp these considerations.

Thus, I take it that Scanlon's second condition on impairing the moral relationship is ambiguous between a claim about the specifically moral reason that some action is wrong and a claim about reasons that may or may not be moral reasons—for example, the consideration that some action will cause me pain. One way to bring out the significance of this ambiguity for the question

whether psychopaths can, on Scanlon's view, impair the moral relationship in this way is to consider it in the light of the natural claim that if one cannot understand reasons of a certain kind, these reasons cannot figure in the meaning of one's actions. Call this *the reasons claim*.

The reasons claim is a natural fit with Scanlon's definition of the meaning of an agent's action as dependent on the reason for which the agent acted. If an agent *A* cannot grasp reasons of kind *K*, then a particular reason of that kind, *r*, cannot be the reason for which *A* performed action *a*. To use Scanlon's terminology, *r* cannot be *A*'s operative reason for *a*-ing. This is all consistent with (i) *r* being a normative reason against *a*-ing, (ii) *r* being a consideration that factors, in the Kingdom of Equals, in the decision that it would be reasonable to reject any principles that allow *a* and (iii) *r* being a moral reason not to *a*.

The reasons claim is both of a piece with the claim, central to Scanlon's view, that the meaning of an action is about the reason for which the agent performed it and consistent with other claims important to his overall view. For example, the reasons claim is consistent with Scanlon's claim that the permissibility of a given action can come apart from its meaning.[150] Even if *A* cannot *a* for reason *r*, it may be the case that *r* figures in the explanation for why *a*-ing is wrong or that *r* just is the consideration that *a*-ing is wrong. Thus, it seems that Scanlon ought to accept the reasons claim.

[150] See the Introduction to Scanlon (2008) for a summary of his view on this matter.

At times, however, Scanlon seems to deny the reasons claim. Consider what he says in a recent response to Watson's criticism of his view, discussed in §2. Scanlon attributes to Watson the view that "the ability to see that one has a reason of a certain kind [is] a necessary condition for having such a reason." Scanlon opposes this to his own view: "I believe, on the contrary, that in the sense relevant to questions of blame a person can be blind to reasons that he really does have. This disagreement may be important to our conflicting views about psychopaths."[151] Though it certainly seems that there is a disagreement here, I am not sure that there is. It is not clear that Scanlon is using the notion of "having a reason" in the same way when he attributes a view to Watson and when he states his own view.[152]

The view Scanlon attributes to Watson is plausibly just a version of the reasons claim: a necessary condition on agent *A* performing action *a* for reason *r* is that *A* is able to grasp *r*.[153] But it is consistent with *A* being unable to grasp *r* that *r* is a reason against *a*-ing. In Scanlon's terminology, *r* may be a normative reason for *A* not to *a*, even if *r* cannot be *A*'s operative reason for *a*-ing.

---

[151] Scanlon (ms., 12).

[152] See Schroeder (2008) for discussion of a distinction between two sense of "reason," the "subjective" and the "objective" senses, as it bears on claims about having reasons. What I say in the text is indebted to this article. It is also interesting to note that Schroeder takes the subjective sense of reason to be tied to the concept of blame.

[153] This is actually stronger than the reasons claim, as stated above. There may be cases in which one cannot grasp a particular reason, yet one can grasp reasons of that kind. In such cases, one would satisfy the reasons claim but not the view Scanlon attributes to Watson. I should note, as well, that I remain agnostic as to whether Scanlon has got Watson's view right.

Normative reasons are relevant to questions of blame. We are often concerned, in the context of blame, with wrong actions.[154] And the agent's quality of will is often a function of his operative reasons given his normative reasons. It is consistent with accepting the reasons claim, and so consistent with the view attributed to Watson, to hold that "in the sense relevant to questions of blame a person can be blind to reasons that he really does have" if the sense of "having a reason" here involves normative reasons.

However, if the sense of "having a reason" here involves operative reasons, Scanlon's claim, in the above quotation, is inconsistent with the reasons claim. This fits his characterization of the opposition between his view and the one he attributes to Watson. But then it is inconsistent with his claims that meaning is about the reason for which the agent acted and that blame is about the meaning of actions. An agent cannot act for a reason that cannot be his operative reason. But then that reason is not relevant to questions of blame. Since the claims about the meaning of actions and the connection between meaning and blame are more central to Scanlon's view, a charitable reading of what he says will preserve them, even at the cost of committing him to denying the claim that he is in disagreement with Watson. Thus, we should interpret the above claim of Scanlon's in a way that dissolves the disagreement with Watson, and we should not take it to suggest that Scanlon denies the reasons claim. When

[154] Often, but not always. Scanlon is clear that, on his view, blame can come apart from wrongness. See Scanlon (2008, 124).

Scanlon says that "in the sense relevant to questions of blame a person can be blind to reasons that he really does have," we should take him to be talking about normative reasons.

Let me return now to the two conditions Scanlon gives for impairment of the moral relationship. In the light of the reasons claim, we are in a position to see that Scanlon should deny the claim that psychopaths can impair the moral relationship in either way.

Recall the first condition on impairment of the moral relationship—that impairment occurs when one "shows a lack of concern with the justifiability of his or her actions." This condition is naturally interpreted as involving an agent judging that the consideration that some action is wrong is not a reason not to perform the action. That is, it involves attributing to the agent a judgment about a moral reason. But psychopaths, as Scanlon understands them, cannot understand moral reasons, so they cannot make judgments about them. So, given the reasons claim, psychopaths cannot impair the moral relationship in this first way.

Consider now the second condition—that impairment of the moral relationship occurs when one exhibits "indifference to considerations that justifiable standards of conduct require one to attend to." This condition is ambiguous between a claim about specifically moral reasons and a claim about reasons that could figure in contractualist determination of whether an action is

wrong. Given the reasons claim, if we interpret this condition in terms of moral reasons, psychopaths cannot satisfy it. Psychopaths cannot grasp moral reasons, so these reasons cannot figure in the meaning of their actions. Psychopaths' indifference to moral reasons is like tigers' for the same reason that their lack of concern for moral reasons is. Thus, they cannot exhibit the relevant indifference to moral reasons.

Alternatively, we can interpret this condition in terms of reasons against actions that one is required to attend to by reasonably justified principles for the general regulation of conduct. It may be the case that psychopaths can be indifferent to such considerations—for example, the consideration that an action would cause someone pain. But it is not the case that psychopaths can be indifferent to these considerations in relation to justifiable standards of conduct. Psychopaths are incapable of grasping these standards, and this incapacity changes the significance of their indifference. Even if we take the relevant considerations to be non-moral reasons against actions, the significance of psychopaths' actions in the light of them is more like that of tigers than that of full-blown moral agents. The way psychopaths govern themselves does not signify indifference to the justifiability of their actions, but rather indifference to

considerations that those concerned with justifiability take to have special significance.[155]

This last claim merits further development. At one point, Scanlon considers the importance of consciousness to the form of rational agency required of moral agents.

> [R]eal governance, in the sense presupposed by moral interaction, requires not only the right kind of regular connection between action "outputs" and the reason-giving force of the considerations presented as "inputs" but something more, namely that these "outputs" depend at crucial junctures on the force that these considerations *seem to the agent* to have.[156]

Scanlon's point here is not the same as the one I am trying to make. Scanlon is here concerned with the question whether universal causal determination would rule out moral agency, entailing that we are just like sophisticated computers. His answer is that it would not. It is consistent with universal causal determination that an agent's actions are affected by his conscious judgment. But computers are not conscious. So this does not entail that human beings are just sophisticated computers.

---

[155] Compare: "A plausible test for deciding whether a given condition should be taken to rule out moral criticism is to ask whether the behavior of a creature which has that condition would, for that reason, lack the distinctive significance that moral failings generally have for relations with others" (Scanlon 1998, 287-8).

[156] Scanlon (1998, 282).

What Scanlon says here is relevant to the point I want to make because it shows that, on his view, moral agency crucially involves the agent's perception of the reasons for which she acts. Moral agency involves conscious appreciation of the *force* of reasons. This claim makes room for a distinction between, on the one hand, an agent who is indifferent to the consideration that stomping on my foot will cause me pain and who appreciates the force of this consideration in the light of its being a reason that factors in the determination that this action is wrong and, on the other hand, an agent who is indifferent to this reason and does not appreciate this force because he is incapable of reasoning about what could be justified to others. Tigers and psychopaths would be agents of the second type. As Scanlon holds that tigers' actions do not exhibit meaning in relation to moral standards, he should hold that psychopaths' actions do not either.

I conclude that Scanlon's view commits him to the claim that psychopaths cannot impair the moral relationship. They cannot satisfy the first condition on impairing the moral relationship because they cannot grasp moral reasons. They cannot satisfy the second condition, either for the same reason, or else because they cannot appreciate the force of reasons against actions in relation to the wrongness of those actions. But a judgment of moral blameworthiness is a judgment that one has acted so as to impair the moral relationship. Thus,

psychopaths' actions do not make them appropriate objects of judgments of moral blameworthiness.

8. Psychological Accuracy and Psychopathic Agency

The argument of the previous section differs from the arguments of §3 and §4 in two ways. First, it does not rely on the contractualist conception. Thus, one could accept the conclusion that Scanlon's view commits him to denying that psychopaths' actions can make them appropriate objects of judgments of blameworthiness without agreeing that his thinking about morality is determined by that particular conception of moral agency. Second, the arguments of §3 and §4 established that Scanlon's view commits him to the claim that psychopaths do not stand in the moral relationship. The argument of the previous section has not shown that. Rather, it has shown that, even if they do stand in the moral relationship, psychopaths' actions cannot impair this relationship.

In this section, I will argue that this second difference between the sets of arguments is only apparent. In the light of a requirement Scanlon explicitly adopts as part of his account of blame, the argument of the previous section establishes that psychopaths do not stand in the moral relationship. Thus, we have an argument that is independent of the contractualist conception and yet leads to the same conclusion as the arguments that depend on it.

To continue to hold that one's judgment about the significance of someone's action is correct, even in the light of new information that shows that their reasons for acting were not what the judgment assumes they were, violates what Scanlon calls the *requirement of psychological accuracy*.

> The requirement of psychological accuracy is straightforward. Insofar as blame depends on the reasons for which an agent acted, a judgment that blame is called for can be modified or undermined by factors that change our view of what those reasons were.[157]

We can get a sense of how the requirement of psychological accuracy is supposed to work by considering what Scanlon says about the following case of Susan Wolf's.

> [A] woman fails to give her friend a book that she very much wants because, as a result of her "personality and social development," she is either "too self-centered for the thought, 'My friend would like this book' to occur to her" or "so unfamiliar with the examples of sincere, non-instrumental friendships that the thought 'I should buy this book, just to make my friend happy' cannot help appearing irrational to her."[158]

---

[157] Scanlon (2008, 180).

[158] Scanlon (1998, 283); he is quoting from Wolf (1990, 85). It may seem anachronistic to consider Scanlon's (1998) response to this case in the light of his (2008) requirement. But the anachronism is only apparent. The discussion in Scanlon (1998) is clearly a precursor to the explicit requirement in Scanlon (2008). Scanlon considers this case as a motivation for objecting to his view on grounds that it is unfair. The objection centers of "the question of accuracy. It is unfair to condemn a person for a certain action if that condemnation is based on inaccurate or incomplete information, when a fuller or more accurate account would reveal that the person is not as bad as he is being portrayed" (Scanlon 1998, 283). In other words, the objection calls for something like the requirement of psychological accuracy. And Scanlon's argument that Wolf's case does not pose a problem for his view can be seen to be an argument for why one might maintain the judgment that the woman is blameworthy for not buying her friend the book, even in the light of this requirement.

This is a case in which someone apparently violates the standards of friendship, and so it seems appropriate to judge that she is (friendship) blameworthy for not buying her friend the book. But when we learn the explanation for why she fails to buy the book, this judgment no longer seems appropriate.

Scanlon considers what difference the further information we learn about this woman makes for our judgment regarding her action. He says that the woman in Wolf's case

> can be seen as someone who is trying just as hard as any of us to do the best thing, but who because of her character or lack of experience cannot see correctly what this is. … Our moral assessment of a person can certainly be affected by additional information about his or her background and circumstances. If we imagine that the woman Wolf describes is sincerely trying to be a good friend but just cannot figure out how to do it, then we might judge her less harshly than we would if she "just didn't care."[159]

Given more information about her background and circumstances, we may have reason to revise our judgment about the meaning of the woman's action. We may have reason to give up our initial judgment that her action impaired her relation with her friend and adopt, instead, the judgment that she was sincerely, but misguidedly, trying to be a good friend. This seems to follow from the requirement of psychological accuracy.

[159] Scanlon (1998, 283).

Scanlon concludes, however, that the specific further information we learn about this woman in Wolf's case tells against revising our judgment about her action.

> But this interpretation of the case is undermined by the suggestion that the woman cannot see a reason to buy a book for her friend because she is too self-centered. If that is the explanation, then the woman is not struggling unsuccessfully to figure out the best thing to do. Rather, she fails to think of what would please her friend because pleasing her friend does not occur to her as important. So moral criticism still seems to be warranted.[160]

Even in the light of full information, the appropriate judgment is that the woman in this case acted for reasons that show that she failed to live up to the standards of friendship. In particular, the self-centered reasons for which she did not buy the book for her friend reveal that she does not recognize that, as her friend, the person who would have appreciated the book has a certain standing. The standards of friendship require that, as her friend, she give the other person's interests and preferences weight (even if not equal weight) in proportion to her own. Genuine friendship is incompatible with self-centeredness. So the requirement of psychological accuracy does not entail that we revise our judgment about the reasons for which the woman acted in the light of the further information we are given in Wolf's case.

[160] Scanlon (1998, 283-4).

On Scanlon's view, there are at least two appropriate responses that the friend might have to the woman who, in Wolf's case, did not buy her the book for self-centered reasons. The first is to blame her. The woman's not buying her friend the book warrants the judgment that she impaired their friendship and revision of the friend's intentions and expectations with respect to her. This blame may come in varying degrees. At the extreme, she might come to see that her friend was "not really a friend after all."[161] This would still count as blame, according to Scanlon, because it involves a modification of intentions and expectations relative to a relationship.[162] And this extreme response may be warranted by the further information regarding the woman's self-centeredness.

The second option would be for the friend to conclude that not only is the woman not really a friend after all, but that she never had any reason to think that she was her friend—"the idea was a mistake or a fantasy."[163] This would warrant revision of her intentions and expectations, but it would, on Scanlon's view, involve "nothing analogous to blame."[164] The reason that this would not involve blame, I take it, is that the grounding relationship never actually obtained. Perhaps the further information about the woman in Wolf's case

[161] Scanlon (2008, 136).

[162] To be clear, this would be friendship blame, not moral blame.

[163] Scanlon (2008, 225, n. 12).

[164] Ibid.

warrants this, even more extreme, response. Perhaps it was a mistake to think that a self-centered person such as she could ever be a friend.

Scanlon claims that this second response—judging that the relationship never obtained in the first place—is possible with respect to the moral relationship. Speaking of a character named Joe, he says:

> The corresponding possibility in regard to the moral relationship that is my main concern would be that I was mistaken in thinking that Joe was a rational agent at all, capable of standing in moral relations with others. In this case as well, the change in attitude that would be called for would not be analogous to blame, because it would not be occasioned by Joe's failure to live up to the standards involved in a relationship he was a party to.[165]

This is exactly the response I think is called for, on Scanlon's view, by the behavior of psychopaths.

To see this, consider the case of a psychopath, call him Joe, who stomps on my foot. Insofar as he resembles everyone else, it is natural to assume that Joe and I stand in the moral relationship.[166] And his stomping on my foot is naturally taken to impair this relationship, signifying either that Joe is unconcerned with the justifiability to others of his action, or that Joe is indifferent to

---

[165] Scanlon (2008, 225-6, n. 12).

[166] Compare what Scanlon says about the moral relationship:

> We assume that this default relationship of mutual regard and forbearance holds between us and the strangers we pass on the road or interact with in the market. When someone does not manifest this concern, it is this relationship that is the standard relative to which our actual relation with them is seen as impaired. (Scanlon 2008, 141)

considerations—such as my pain—that reasonably justified standards of conduct require him to attend to, or both.

But now suppose that I come to learn that Joe is a psychopath. He is incapable of understanding moral reasons and incapable of understanding the force of reasons against actions in relation to justifiable standards of conduct. Doesn't this new information give me reason to revise my assessment of the significance of Joe's action?

I think it does. As I argued in the previous section, psychopaths' actions cannot have the significance required to satisfy either of Scanlon's two conditions on impairing the moral relationship. Thus, the new information that Joe is a psychopath gives me reason to deny that Joe can impair the moral relationship. This calls for revising my initial judgment that he is morally blameworthy for stomping on my foot.[167]

We should distinguish between the revision called for when I learn that Joe is a psychopath from a different revision that may be called for by different information. Suppose I learn that Joe thought that there was a spider on my shoe and that he was saving me from a poisonous bite.[168] This would also call for a

---

[167] It may be that I never come to learn that Joe is a psychopath—and so incapable of reasoning about what could be justified to others. In that case, it may be that I never come to have any further information that would call for revising my judgment about the moral significance of Joe's action in the way I am arguing I should revise it. But the issue at hand is not whether Scanlon's view entails that I should or should not revise my judgment in the light of faulty or incomplete information. The issue is whether the view entails that I should revise my judgment given the further information that Joe is a psychopath.

[168] See Scanlon (1998, 279).

revision of my judgment that Joe impaired our moral relationship when he stomped on my foot. But it would call for this revision in the light a revision of the attitude I take his action indicate. As Scanlon puts it, Joe's action "may have been hasty, but it was not ill-intended."[169]

When I learn that Joe thought there was a spider on my shoe, I come to judge that Joe did not impair the moral relationship because he *did not* act for morally objectionable reasons. His action did not have a morally objectionable meaning. When I learn that Joe is a psychopath, I come to judge that Joe did not impair the moral relationship because he *could not* act for morally objectionable reasons. His action could not have a morally objectionable meaning. The difference between these two revisions mirrors the difference between excuses and exemptions.[170] To claim that Joe is excused from wrongdoing is to claim that he only apparently acted for objectionable reasons, but in fact he did not. This is like the revision that Scanlon considers, but ultimately rejects, with respect to the judgment that the woman in Wolf's case is blameworthy for not buying her friend the book. To claim that Joe is exempted from wrongdoing is to claim that he is not "a potential term in moral relationships."[171] Since he is a psychopath, it is mistaken to hold that Joe and I ever stood in the moral relationship. His

[169] Scanlon (1998, 279).

[170] The discussion here owes a great deal to Watson's discussion of Strawson's view in Watson (1987, esp. at pp. 223-5 of his 2004).

[171] Watson (1987, at p. 225 of his 2004).

stomping on my foot could not have impaired his moral relationship with me (or anyone else) because there was nothing to impair in the first place.

9. Explaining Scanlon's Mistaken Claims about Psychopaths

I have offered two independent arguments for the conclusion that Scanlon's explicit claims that psychopaths may be the proper objects of judgments of moral blameworthiness or that they may be the proper objects of moral blaming responses is inconsistent with the commitments of his own view. In particular, I have argued that Scanlon's view commits him to the claim that psychopaths do not stand in the moral relationship.

It is a bold charge to claim, as I have about Scanlon, that someone has misunderstood the commitments of his own view. In this section, I will back up this claim by offering an explanation for why he might have made this mistake. My explanation will appeal to the contractualist conception of moral agency, according to which justification of one's actions to others in terms of reasons is fundamental to moral agency. In short, it will be that Scanlon continues to consider the actions of psychopaths through the lens of this conception, even though he is aware that it does not apply. This is understandable, however, because this conception is our default lens through which to consider the significance of human actions and because it is such an attractive conception of our agency.

Scanlon claims that the moral relationship is the "default relationship" assumed to hold between oneself and others. This will be true even when one comes into contact with psychopaths. Because psychopaths resemble other people one interacts with on a regular basis, one assumes that one stands in the moral relationship with them. And when they do things that are morally objectionable—for example, stomping on one's foot—it is natural to judge, on the basis of this assumption, that they are morally blameworthy for doing so.

I have been arguing that the further information that someone is a psychopath—and so incapable of reasoning about what could be justified to others—exempts him from the moral relationship. His actions do not have the same significance as the actions of a moral agent because it is inaccurate to attribute judgments about reasons to him that, directly or indirectly, involve the notion of justifiability to others. Since this notion is foundational to Scanlon's view, and since Scanlon adopts the requirement of psychological accuracy, I take it that Scanlon's view commits him to agreement with me on this point.

Scanlon, however, does not agree with my assessment. In discussing the very objection I have been pressing, he considers what difference it makes to moral assessment of an agent's action that the agent is incapable of seeing the force of moral reasons.

> When we see that a person is unable to avoid a certain action, or unable to see that that action will cause harm, this inability makes a difference

> because it intervenes between the agent's action and his or he assessment of the relevant reasons: because of this inability, that action need not reflect a judgment on the agent's part that the harm caused by the action did not count against performing it. But an inability to see the force of a certain reason, or of moral considerations in general, does not have this same effect. A person who is unable to see why the fact that his action would injure me should count against it still holds this *doesn't* count against it.[172]

Scanlon here contrasts two kinds of inability. Applying his comments to the case of someone stomping on my foot, we might put Scanlon's point as follows: the significance for moral assessment of (a) the inability to see that stomping on my foot would cause me harm is crucially different from that of (b) the inability to see why the consideration that stomping on my foot would cause me harm counts against stomping on my foot. In the case of (a), we cannot accurately attribute a judgment to the agent about the reason-giving force of considerations pertaining to the harm caused by his action. In the case of (b), however, we can attribute such a judgment to the agent. We can attribute to him the judgment that this consideration does not have sufficient reason-giving force to count against stomping on my foot.

I agree with Scanlon about the difference between these two inabilities. But his discussion is not exhaustive. He does not consider the difference between (b) the inability to see why the consideration that stomping on my foot would cause me harm counts against stomping on my foot and (c) the inability to see

[172] Scanlon (1998, 288).

why the consideration that stomping on my foot would cause me harm counts against stomping on my foot *in relation to the justifiability to others of this action*. The former involves attributing to an agent the judgment that some consideration does not have sufficient reason-giving force to count against performing some action. The latter goes beyond this and involves attributing to an agent the judgment that some consideration does not have sufficient reason-giving force to count against performing some action *in relation to the justifiability to others of this action*. We can accurately attribute a judgment of the former kind to an agent incapable of seeing the force of moral reasons, but we cannot accurately attribute to him a judgment of the latter kind. In the case of an agent capable for seeing the force of moral reasons, we can attribute both judgments.

When an agent who satisfies the contractualist conception stomps on my foot we can (barring further relevant facts) accurately attribute to her (i) the judgment that the consideration that this will cause me harm is no reason against stomping on my foot and (ii) the judgment that the consideration that this reason would factor in a decision that this action would be wrong does not count against performing it. The fact that we cannot accurately attribute judgments of type (ii) to agents who are incapable of reasoning about what could be justified to others—for example, our psychopath, Joe—shows that the significance of their actions is importantly different than those of agents capable of such reasoning.

The difference between a judgment of type (i) and a judgment of type (ii) is subtle. And the claim that a type (ii) judgment is not attributable to a given agent depends on conceiving of his agency as not satisfying the contractualist conception. But since it is our default stance to consider the actions of others through the lens of the contractualist conception, our default conception of others' agency will lead to our attributing type (ii) judgments to them on the basis of their behavior. It may be that only when we explicitly attend to information that speaks against his satisfying our normal conception of human agency that we come to judge that someone's action does not warrant attributing a type (ii) judgment to him. But my attention may be easily drawn away from this information—say, by the pain in my foot—and I may easily slip back into my default conception of human agency, even in the light of this information. So it is understandable why I might judge that a psychopath's action has a significance it does not. When I consider his action in the normal way, I attribute a greater significance to it than is warranted by the facts. And, even if I try to attend more carefully to the facts, my habitual way of thinking about such things may creep back in mind and cloud my assessment of the situation.

I think this is a plausible explanation for why Scanlon might mistakenly claim that psychopaths may be morally responsible for what they do. When Scanlon thinks about morality, he conceives of human agency under the guise of a conception, according to which justification of one's actions to others in terms

of reasons is fundamental. Because this is his normal way of thinking about the significance of others' actions, it is hard to attend to what would normally be morally significant behavior other than through the guise of this conception. Thus, even granting that psychopaths do not have the capacities that seem required by his default conception of moral agency, Scanlon may have come to think about the significance of their behavior through the lens of this conception because that is what he is used to doing.[173] This explanation seems especially well-suited to Scanlon's own judgments, because it seems that one would be more likely to assume the contractualist conception when thinking about the moral significance of human actions the more one is enmeshed in Scanlon's way of thinking about morality.

The explanation I am offering of why Scanlon mistakenly claims that psychopaths may be proper objects of judgments of moral blameworthiness or moral blaming responses is that he is thinking about the significance of their actions through the lens of the contractualist conception, even when this is not warranted by the facts. This explanation is strengthened, I think, once we note the attractiveness of the contractualist conception as a background picture of moral agency. It is difficult to consider potentially morally salient actions other than from the perspective of this conception because it is such an attractive conception in the first place.

[173] Perhaps a similar mistake is behind the thinking embodied in what Watson calls "the Affirmative Argument." See Watson (2011, 309).

Pamela Hieronymi claims that much of the appeal of Scanlon's contractualism derives from the central notion of respect associated with the notion of justifiability to others.

> This narrow notion of justifiability to others is associated with a specific form of respect, and this form of respect gives contractualism its appeal. Again, according to contractualism, the significance of moral wrongdoing lies in the fact that one has violated the principles that recognize the standing of each to partially, symmetrically, determine how one shall act—one has violated the terms that would be agreed to in the Kingdom of Equals.[174]

I would make the similar claim that much of the appeal of Scanlon's contractualism derives from the basic conception of moral agency that grounds it, of which the narrow notion of justifiability Hieronymi discusses is a significant part. It is a compelling conception of moral agency that justifying one's actions to others in terms of reasons is fundamental to it. And the moral theory that Scanlon develops out of this conception inherits the appeal of this way of conceiving of our agency.

Perhaps the appeal of this conception explains why it may be difficult to set it aside when considering the actions of agents like Joe. These are cases in which a putatively normal human agent performs what would, under normal circumstances, be a morally objectionable action, but they are special cases because the agent in question does not, in fact, satisfy the attractive conception of

[174] Hieronymi (2011, 117).

moral agency. And if one keeps the appealing conception in mind when considering such cases, it seems that one would arrive at the assessment that there is no relevant difference between the moral significance of this agent's action and that of a normal moral agent, as Scanlon does.

If I conceive of Joe as I do most everyone else—that is, in the light of the attractive contractualist conception of moral agency—it seems appropriate to endow the judgment, appropriately attributed to him, that the consideration that this action will cause me pain is no reason not to perform it with the same significance as this judgment would have for me. It would be natural to claim, with Scanlon, that Joe's inability to see the moral reasons against stomping on my foot makes no difference to the moral significance of his action. But this is natural only if we consider Joe in the light of the contractualist conception. I have argued, however, that Joe does not satisfy this conception. Thus, it is not appropriate to consider him in the light of it, and the seemingly appropriate claim is mistaken. There is a difference between the significance of Joe's actions and mine. And we can see this when we attend to the fact that Joe and I are (in the context of morality) different kinds of agents.

I think it adds to the appeal of Scanlon's contractualism that it can account for the difference between my action and Joe's in this way. To claim that Joe is exempt from moral assessment because he is (in the context of morality) a different kind of agent than I am is to be upfront about the "gulf"—to borrow a

term of Scanlon's—between Joe and myself.[175] There is a distance that (in this context) separates me from Joe. I am able to reason about the justifiability to others of my actions, but he is not. This explanation of the gulf between us, however, tells against attributing the same significance to Joe's actions as to mine. And this is why we should not hold that Joe stands in the moral relationship with me or anyone else.

My proposed explanation for why Scanlon mistakenly claims that psychopaths may be proper objects of moral assessment is that he may have been too faithful to the attractive conception of moral agency behind his moral theory. Even in considering cases of actions performed by agents who do not satisfy it, Scanlon may have kept the contractualist conception in mind. And this may have led him to attribute judgments about reasons to psychopaths that are not warranted by his own account of moral agency and that violate his requirement of psychological accuracy. But when we attend to the inappropriateness, on Scanlon's own view, of attributing these judgments to psychopaths, we recognize

[175] "People with a consuming interest in one activity often feel that a large gulf separates them from those who cannot see the point or value of that pursuit. … What I am suggesting is that almost all of us have reason to see the gulf separating us from an 'amoralist' as having this character, and that this accounts for the special importance we attach to seeing the force of moral considerations" (Scanlon 1998, 159-60). I would claim that the psychopath's amoralism is a special case of amoralism, one that obtains because the individual is incapable of understanding moral reasons. The run-of-the-mill amoralist simply does not care about morality. It is an interesting question, but beyond the scope of this discussion, what difference the psychopath's incapacity makes to the gulf between us and him, in relation to the gulf between us and the run-of-the-mill-amoralist.

that they are exempt from the moral relationship. And, thus, they are not morally blameworthy or properly morally blamed for what they do.

9. Conclusion

I have offered several arguments for the conclusion that Scanlon's claim that psychopaths may be morally blameworthy or properly morally blamed for what they do is inconsistent with commitments of his contractualist moral theory. The third of these arguments is independent of the first two in the sense that one could accept it without accepting the others as well. But it is possible to see these arguments as having a common ground. If one accepts the contractualist conception, then one accepts that justification of one's actions to others in terms of reasons is fundamental to moral agency. It would be natural, on the basis of this conception of moral agency, to accept both that the ability to recognize and respond to moral reasons is required in order to be the proper object of moral assessment and that the ability to reason about what could be justified to others is required for one's actions to have moral significance. In other words, the contractualist conception can, but need not, be behind all three of the arguments I have offered. As we saw in the previous section, the contractualist conception is also behind a possible explanation for why one might mistakenly claim that psychopaths are morally responsible, even (and especially) if one thinks about morality as Scanlon does. So we can take the contractualist

conception to be in the background of Scanlon's thinking about morality, both correct and mistaken.[176]

If I am right about the way in which the contractualist conception is behind Scanlon's moral theory, then this conception is especially central to his view. If I am right that Scanlon's claims about the appropriateness of moral assessment of psychopaths' for their actions are inconsistent with the contractualist conception, then these claims are inconsistent with a central aspect of his view. The inconsistency may be resolved by giving up either the contractualist conception or the claims about psychopaths. Since the former is more central to the view than the latter, I conclude that Scanlon should revise his view by dropping these claims about the moral responsibility of psychopaths.

In the context of this dissertation, this argument from a conception of what is fundamental to a certain form of agency to a conclusion about how to revise a particular philosophical theory is interesting because it shows a further way in which attending to basic conceptions of what we are like as agents is important. Not only can we, as we saw in Chapter Three and Chapter Four, identify arguments against rival views that are not apt to be fair or persuasive because they presuppose contentious conceptions of what we are like as agents, but we can also identify claims that are in tension with the background

[176] Moreover, since my arguments have appealed to writings of Scanlon's over a long period of time, we can take the contractualist conception to have been in the background of his thinking about morality all along.

conception animating a particular view and that should, on that basis, be expunged from the theory.

A further element of this chapter that is especially interesting in the context of this dissertation is that we have seen that the sort of holistic argument in favor of a particular basic conception of what we are like as agents that I am advocating should replace familiar arguments from intuitions about cases need not proceed by identifying theories of various kinds grounded in the same basic conception. Rather, one might identify a theory grounded in one basic conception and a different kind of theory grounded in a second conception that may be derived from the first. In this chapter, we saw that the contractualist conception, behind Scanlon's moral theory, may be seen as derived from the evaluator conception, behind Watson's theory of self-governance (and an expanded account of human agency stemming from it). A holistic argument may appeal to this fact in support of the evaluator conception. Not only can this conception ground an attractive account of human agency, but it can also serve as the basis for a conception that grounds an attractive moral theory.

Conclusion

This dissertation has focused on two questions: What is right/wrong action? What is self-governed action? My main thesis has been that debates about which are the best answers to these questions—the best moral theory and the best theory of self-governance, respectively—would do well to attend to the basic conceptions of what we are like as agents that ground these theories. Lack of attention to these conceptions has landed the debates in a dialectical stalemate. My other thesis has been that we can see our way forward in these debates by focusing on the basic conceptions behind rival theories. I have suggested a way of mounting a holistic argument in favor of a particular basic conception by showing that it can ground (or be properly related to the grounds of) various kinds of philosophical theory. My theses are related in that they take seriously the importance of basic conceptions of what we are like as agents, and I hope that the arguments I have provided in support of them have convinced the reader that these conceptions really are important in these ways.

I would like to conclude with some remarks about how I see the shape of the dialectic going forward. It is no accident that I have been sketching a holistic argument for the evaluator conception. In addition to this being the basic conception of what we are like as agents that I personally favor, it is also the

conception that best aligns with the tradition of Western philosophical thought. Aristotle, for instance, begins the *Nicomachean Ethics* with the assertion that "Every craft and every line of inquiry, and likewise every action and decision, seems to seek some good."[177] The notion that human agency operates under the guise of the good was accepted as doctrine on into the Twentieth Century. Rawls, for instance, affirms a conception of the person that includes both "the capacity for an effective sense of justice" and "the ability to form, to revise, and rationally to pursue a conception of the good."[178] Recent challenges to the evaluator conception may be seen as attempts to undermine philosophical orthodoxy. This adds to their interest because it adds to the significance that would attach to their success. But I am inclined to think that it also places them at a dialectical disadvantage.

I have argued that prominent contemporary challenges fail to provide convincing reason to abandon the evaluator conception. My diagnosis of this failure is that the arguments found in the literature are not apt to be fair or persuasive because they invoke conceptions of what is fundamental to human agency that are not shared in the context of the dialectic. And I have sketched a way of moving the dialectic forward with special attention to the basic conceptions behind rival theories. I am inclined to think, however, that the

[177] Aristotle (1999, 1).

[178] Rawls (1980, 525).

playing field going forward is not exactly level. Because challenges to the evaluator conception go against the grain of tradition, they shoulder an especially heavy dialectical burden.

I am inclined to think that the proponent of the evaluator conception can offer an argument from authority to shift the burden of proof onto the shoulders of proponents of rival views. This shapes the dialectic in the following way. Suppose that a proponent of the evaluator conception and a proponent of, say, the explainer conception were to muster equally strong holistic arguments in favor of their respective favored basic conceptions of what we are like as agents. The proponent of the evaluator conception is able to show that his preferred basic conception can ground (or be properly related to the ground of) various independently attractive philosophical theories, and the proponent of the explainer conception is able to show that the same is equally true of his preferred basic conception. I am inclined to think that, in a dialectical context such as this one, the proponent of the evaluator conception is in a more favorable position than the proponent of the explainer conception. This is because, in addition to his holistic argument, he may muster an argument from authority in favor of his view. That is, he may bolster his holistic argument in favor of the evaluator conception with an argument to the effect that this conception represents philosophical orthodoxy. To be clear, I do not think that this decisively tips the scales in favor of the evaluator conception. But I do think it shows that, in order

to convince us that we ought to break free of tradition, the proponent of the explainer conception should provide (at least) some reason to think that we should not take the fact that many brilliant thinkers of the past have accepted something like the evaluator conception as a consideration in its favor.

These reflections on the relationship between the Western philosophical tradition and the rival basic conceptions of what we are like as agents considered in this dissertation may be taken to provide further reason to accept the evaluator conception. But they may also be taken to provide a further specification for how proponents of the explainer and planner conceptions should seek to move the dialectic forward in ways favorable to their respective views. Not only should proponents of theories grounded in these basic conceptions seek to provide holistic arguments in favor of their preferred conceptions, but they should also seek to provide reasons to think that the tradition has been mistaken. Personally, I am not sanguine about the prospects of providing a convincing argument to this effect. But I register my conviction that this is indeed the challenge and my sincere hope that it be taken seriously by those with philosophical predilections different from my own.

Bibliography

Aristotle. (1999), *Nicomachean Ethics* 2nd Ed., Terence Irwin (trans.), (Indianapolis: Hackett).

Bratman, M. E. (1987), *Intentions, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press).

---. (2000), 'Reflection, Planning, and Temporally Extended Agency', repr. in Bratman 2007, 21-46.

---. (2003), 'A Desire of One's Own', repr. in Bratman (2007), 137-161.

---. (2004), 'Three Theories of Self-Governance', repr. in Bratman (2007), 222-253.

---. (2005), 'Planning Agency, Autonomous Agency', repr. in Bratman (2007), 195-221.

---. (2007), *Structures of Agency: Essays* (New York: Oxford University Press).

---. (2009), 'Intention, Practical Rationality, and Self-Governance', *Ethics*, 119: 411-443.

---. (2010), Agency, Time, and Sociality', *Proceedings and Addresses of the American Philosophical Association*, 84(2): 7-26.

Dancy, J. (2000), *Practical Reality* (New York: Oxford University Press).

Fischer, J. M. (2010), 'Responsibility and Autonomy', in Timothy O'Connor and Constantine Sandis (eds.), *A Companion to the Philosophy of Action* (West Sussex, UK: Wiley-Blackwell), 309-316.

---. (ms.), 'Responsibility and Autonomy: The Problem of Mission Creep'.

Fischer, J. M. and Mark Ravizza. (1998), *Responsibility and Control* (New York: Cambridge University Press).

Frankfurt, H. (1971), 'Freedom of the Will and the Concept of a Person', reprinted in Frankfurt (1988), 11-25.

---. (1977), 'Identification and Externality', repr. in Frankfurt (1988), 58-68.

---. (1988), *The Importance of What We Care About: Philosophical Essays* (New York: Cambridge University Press).

---. (1994), 'Autonomy, Necessity, and Love', repr. in Frankfurt (1999), 129-141.

---. (1999), *Necessity, Volition, and Love* (New York: Cambridge University Press).

---. (2006), *Taking Ourselves Seriously & Getting It Right* (Stanford, CA: Stanford University Press).

Grice, P. (1975), 'Method in Philosophical Psychology', *Proceedings and Addresses of the American Philosophical Association* XLVIII: 23-53.

Hieronymi, P. (2011), 'Of Metaethics and Motivation: The Appeal of Contractualism', in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (New York: Oxford University Press), 101-128.

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

Korsgaard, C. M. (1986), 'Skepticism about Practical Reason', repr. in Korsgaard (1996b), 311-334.

---. (1996a), *The Sources of Normativity* (New York: Cambridge University Press).

---. (1996b), *Creating the Kingdom of Ends* (New York: Cambridge University Press).

Rawls, J. (1980), 'Kantian Constructivism in Moral Theory', *The Journal of Philosophy* 77(9): 515-572.

Scanlon, T. M. (1982), 'Contractualism and Utilitarianism', in Amarty Sen and Bernard Williams (eds.), *Utilitarianism and Beyond* (New York: Cambridge University Press), 103-128.

---. (1986), 'The Significance of Choice', in Sterling M. McMurrin (ed.), *The Tanner Lectures on Human Values* (Salt Lake City: University of Utah Press), 151-177.

---. (1998) *What We Owe to Each Other* (Cambridge, MA: Harvard University Press).

(2008), *Moral Dimensions* (Cambridge, MA: Harvard University Press).

---. (ms.), 'Interpreting Blame'.

Schroeder, M. (2008), 'Having Reasons', *Philosophical Studies* 139(1): 57-71.

Velleman, J. D. (1992a), 'What Happens When Someone Acts?', repr. in Velleman (2000), 123-143.

---. (1992b), 'The Guise of the Good', repr. In Velleman (2000), 99-122.

---. (2000), *The Possibility of Practical Reason* (New York: Oxford University Press).

---. (2001), 'Identification and Identity', repr. in Velleman (2006a), 330-360.

---. (2004), 'Willing the Law', repr. in Velleman (2006a), 284-311.

---. (2006a), *Self to Self: Selected Essays* (New York: Oxford University Press).

---. (2006b), 'The Centered Self', in Velleman (2006a), 253-283.

Wallace, R. Jay. (1994), *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press).

---. (2011), 'Dispassionate Opprobrium: On Blame and the Reactive Sentiments', in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (New York: Oxford University Press), 348-372.

Watson, G. (1975), 'Free Agency', repr. in Watson (2004), 13-32.

---. (1977), 'Skepticism about Weakness of Will', repr. in Watson (2004), 33-58.

---. (1987), 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme', repr. Watson (2004), 219-259.

---. (2002), 'Volitional Necessities', repr. in Watson (2004), 88-122.

---. (2004), *Agency and Answerability: Selected Essays* (New York: Oxford University Press).

---. (2005), 'Hierarchy and Agential Authority', in John Martin Fischer (ed.), *Free Will: Critical Concepts in Philosophy*, vol. iv (New York: Routledge), 90- 97.

---. (2011), 'The trouble with Psychopaths', in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (New York: Oxford University Press), 307-331.

Williams, B. A. O. (1985), *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press).

Wolf, S. (1990), *Freedom within Reason* (New York: Oxford University Press).