March 2018

The Philosophical Value of Reflective Endorsement

Rachel Robison

2018

The Philosophical Value of Reflective Endorsement

Rachel Robison

THE PHILOSOPHICAL VALUE OF REFLECTIVE ENDORSEMENT

A Dissertation Presented

By

RACHEL DIANE ROBISON

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

February 2018

Philosophy

THE PHILOSOPHICAL
VALUE OF REFLECTIVE
ENDORSEMENT

A Dissertation Presented

By

RACHEL DIANE ROBISON

Approved as to style and content by:

_____
Hilary Kornblith, Chair

_____
Ernesto V. Garcia, Member

_____
Peter Graham, Member

_____
Andrew Cohen, Member

_____
Joseph Levine, Department Head
Department of Philosophy

DEDICATION

To my husband Richard Greene and my son Henry Greene, to whom I will be forever grateful for their endless support and encouragement.

ACKNOWLEDGMENTS

I'm happy for the opportunity to express my gratitude to a number of important people, without whom the completion of this dissertation would not have been possible. I'd like to thank the members of my committee: Ernesto Garcia, Pete Graham, and Andrew Cohen. I'd also like to express my appreciation for the guidance that my advisor, Hilary Kornblith provided through this process and through graduate school in general. He has contributed substantially to my growth as a philosopher.

I moved from the west coast to the east coast by myself to attend graduate school, and I survived it with a little help from some dear friends. I'll always hold dear long conversations with Brandy Burfield, Jon Rosen, and Kristian Olsen, and I look forward to lifelong friendships with each of them.

I wrote this dissertation in Utah, and these acknowledgments would not be complete without an expression of gratitude for our wonderful friends Joe and Christina Charbonneau and their son Simon. They've provided much needed levity for me at crucial times.

I'd also like to thank my parents, Susan and Neal Robison, and my siblings, Jason, Brooke, Becca, and Clint, for their consistent support and encouragement.

The gratitude that I feel for the love and care I've received from my husband, Richard Greene, is difficult to put into words. He's taught me so much about how to be a philosopher, from basic philosophical knowledge to professionalism and backbone. He has been someone to test ideas out on, someone to motivate me when

I can't find the motivation on my own, and someone to reassure me in moments of self-doubt. As a husband, he's often shouldered more than his share of the responsibilities while I worked on this project.

I'd also like to thank my wonderful son Henry Greene who believes in me so much that he built me a graduation gift out of Legos before I had written a single word of the dissertation.

ABSTRACT

THE PHILOSOPHICAL VALUE OF REFLECTIVE ENDORSEMENT

FEBRUARY 2018

RACHEL DIANE ROBISON, B.A., WEBER STATE UNIVERSITY

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

PH.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Hilary Kornblith

Through the years, many philosophers have appealed to reflective endorsement to address important philosophical problems. In this dissertation, I evaluate the merits of those approaches. I first consider Christine Korsgaard's appeal to reflective endorsement to solve what she calls "the normative problem." I then consider Harry Frankfurt's use of reflective endorsement as part of his account of "caring," which plays a crucial role in his accounts of agency, free will, and personhood. I then turn to Marilyn Friedman's use of reflective endorsement to explain autonomous action. Finally, I turn to Alan Gibbard's use of reflective endorsement as part of an account of what it is to make a normative judgment. I argue that each of these positions is subject to similar problems—they fail to provide a plausible account of the self. In the remaining chapters, I argue that empirical psychological studies suggest that reflective endorsement plays an important role with respect to psychological health, but that judgments made by using a process of reflective endorsement are generally not accurate. Ultimately, I argue that reflective endorsement is valuable, but only under certain circumstances.

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION

Many of history's deepest thinkers have identified our capacity to introspect as the source of some of the most unique features of human experience. It has been suggested that the capacity for reflection gave us the evolutionary advantage we needed to survive, that it makes free action possible, that it lends some sort of normative authority to an agent's actions, and that it gives rise to our deepest existential anxieties. Our ability to introspect allows us to decide what matters to us, to set goals and to make plans.

Not only are human beings capable of introspection, they are capable of taking evaluative positions toward their own inner states. They can adopt second order desires about their desires or second order beliefs about their beliefs. They can disavow their desire to take a drug or endorse their inclination to start to work on that novel they always intended to write.

If introspection and the capacity for endorsement set human experience apart, it is not surprising that philosophers have been motivated to appeal to that very feature to solve a host of philosophical problems that trouble humans—or beings sufficiently like humans. Introspection and endorsement are used as the lynchpins in accounts of philosophical topics like free will, autonomy, agency, normativity, morality, and in accounts of justification in epistemology.

One feature that most of these accounts share in common is a faith in the knowledge of the self that is gained through introspection. Most of these views

maintain that we come to see pretty clearly upon introspection the set of things that we truly care about, the real motivations for our actions, and the extent to which our beliefs, desires, goals and projects are internally consistent. The idea that we see our "true selves" clearly upon introspection seems to be, on many accounts, the source of the authority of its outputs. The value of endorsement is often tied to the value of authenticity.

Running counter to the sense of confidence philosophers seem to have on this topic are both philosophical arguments and evidence in empirical psychology that suggests that human beings don't understand themselves and their motivation for action anywhere near as well as they think that they do. For example, in their work, *The Person and the Situation*, social psychologists Nisbett and Ross survey a wide range of studies that support the conclusion that our attributions of stable personality characteristics, both in others and in ourselves, are often wrong, and that the motivation for a wide range of human behavior actually has to do with the particulars of the external situation in which people find themselves. If our attributions of enduring personality traits are misguided, then perhaps our affirmation of those very traits upon introspection is not actually the promising tool in the philosopher's toolbox that it is frequently taken to be.

It will be my project in this book to evaluate philosophical arguments that make use of introspection and endorsement to determine when, if ever, such accounts are successful. One major challenge to the completion of a project like this is that thinkers use very different language to describe a very similar concept. Throughout this work, I will be using Christine Korsgaard's terminology, "reflective

endorsement" to describe the phenomenon that interests me. Each of the accounts that I will survey uses different language to describe the process that is my primary focus. I take them to each have the same general processes in mind—the process of (1) introspection, and (2), taking an evaluative position toward the inner states upon which one introspects. For each account, this process will be key to solving some major philosophical puzzle.

Roadmap

In the next chapter of this work, I will look at Christine Korsgaard's use of reflective endorsement in her work *The Sources of Normativity.* I will ultimately argue that the procedure of reflective endorsement cannot do what she wants it to do—it cannot answer what she calls "the normative question" in a way provides a fundamental grounding for our normative judgments.

In the third chapter, I will look at Harry Frankfurt's use of the concept of endorsement in a wide range of his works. Frankfurt uses it to explain a number of key philosophical concepts. Underlying all of these explanations is Frankfurt's account of caring, which fundamentally relies on endorsement. In chapter two, I will argue that Frankfurt's endorsement-based approach to understanding caring doesn't capture the full range of cases in which human beings intuitively care about things.

In the fourth chapter, I will look at Marilyn Friedman's endorsement account of autonomy. She argues that autonomous actions are actions that issue from an agent's "true self." On this view, judgments that issue from an agent's true self have

authority over actions that issue from aspects of a person from which that person is alienated. I will argue that endorsed states do not always, or even often, reflect an agent's "true self." If this is so, then endorsed behaviors do not have the kind of authority that they need to truly be considered autonomous.

In Chapter Five, I will consider Allan Gibbard's expressivist account of the nature of normative judgments. Gibbard's project is very different from the projects of the other philosophers presented here. I include Gibbard in this work because he makes claims about the relationship between motivation, normative judgment, and reflective endorsement that are directly related to the concerns pertaining to moral psychology that I raise for the other thinkers we will consider. I will argue that the taxonomy of motivation that Gibbard provides is too narrow—his endorsement account of normative judgments leaves out other psychological events that are, intuitively, normative judgments as well.

The major theme of this book will be that reflective endorsement cannot do all the things that it is claimed that it can do. In Chapter Six, I will argue that one major shortcoming of the views considered here is that emphasis is put on *the value of the process* of endorsement itself, rather than how reliably it produces outputs that are truly valuable. In Chapter Seven, I will argue that reflective endorsement is sometimes valuable toward the end of psychological health. Finally, in the last chapter, I will sketch a virtue theoretic account of the way in which reflective endorsement, under a narrow set of conditions, could help to achieve a set of values that is more philosophical in nature.

CHAPTER II

REFLECTIVE ENDORSEMENT, CONSTRUCTIVISM, AND NORMATIVE GROUNDING

In her influential work *The Sources of Normativity*, hereafter 'SN,' Christine Korsgaard introduces the method of reflective endorsement as a way to ground normativity. She argues that the very idea of normativity gives rise to a general problem. Because we are reflective creatures, we can deliberate about our reasons for action. For any normative claim, we can always ask ourselves, "Why should I accept *that*?" This leads to a question about the foundations of normativity itself, that is, about its ultimate grounds or justification. For example, we find ourselves faced on a daily basis with all kinds of normative dilemmas. A person who craves chocolate might ask herself, "Should I eat this slice of cake?" If she decides that she shouldn't eat it because doing so is unhealthy, she might raise the further question, "But why should I care about being healthy?" We need an answer to these normative questions that provides some kind of satisfactory justification.[1] We also

[1] I've identified what might be understood as two distinct questions here. The first question is of the form "Why should I care about P?" and the second is of the form "What would provide satisfactory justification for P?" An answer to one of these questions doesn't necessarily constitute an answer to the other. For example it may be the case that we could provide a full account of the justification for a normative claim without thereby providing any reason for any particular agent to care about that normative claim or to be motivated by it. Korsgaard rejects the idea that these are two distinct questions. She contends that a satisfactory answer to the justificatory question will answer the "why should I care?" question, if it is answered correctly. As we will see, she grounds normativity in a feature of ourselves that we

find ourselves faced with distinctively moral normative dilemmas. For example, imagine that someone I love commits a serious crime. Should I turn them in? In these types of situations, morality might demand me to substantially sacrifice my own interests for the sake of some overriding moral concern. For this reason, moral claims require a special kind of answer that explains why we should be committed to morality in the first place. Call these various concerns about the justification of normative claims the *Normative Problem*.

In *SN,* Korsgaard claims that an adequate answer to the Normative Problem must be able to solve standard regress worries. For any normative claim, we can ask: "But why should I care about that?" On Korsgaard's view, if we can identify some reply about which it would be incoherent to raise this question, we have found a satisfactory answer to the Normative Problem.[2] In order to achieve this result, she argues that the type of value involved in justifying normative claims must be both 'intrinsic' and 'final'. Something has 'intrinsic' value if it has value in itself and does not derive its value from any outside source. Something has 'final' or 'non-instrumental' value if we value it for its own sake rather than for the sake of achieving some other end. As Korsgaard describes the difference between these two ideas, the former identifies the *source* of value whereas the latter explains *how* we value it.[3][4]

---

can't help but to care about—our capacity to care about or value things in the first place.

[2] Korsgaard, Christine. 1996. *The Sources of Normativity,* Cambridge University Press. 93.

[3] Ibid., 111.

[4] For further discussion of this distinction, see Korsgaard 1983 where she argues against the standard conflation of (1) something having 'intrinsic' value, which she

One final criterion Korsgaard sets forth for a general account of normativity is that it must be able to explain both moral and non-moral normative claims. It must explain why we are conditionally committed to various non-moral normative claims and unconditionally committed to moral ones.

In this chapter, I will argue that Korsgaard's commitment to the importance of reflective endorsement generates a number of problems. First, I will argue that her view generates counterintuitive results in non-moral cases. I will then argue that her view does not generate the results that she wants in moral cases either, because she has established neither that our humanity has intrinsic value nor that it has final value. I will argue further that, though she attempts to adopt only a procedural realism according to which reflective endorsement has value, her view actually commits her, after all, to substantive moral realism.

Korsgaard's Constructivism

Korsgaard defends a constructivist approach to the Normative Problem. While *realists* maintain that there exist normative truths that hold independently of our own judgments, attitudes, or beliefs, *constructivists* argue that what makes a normative claim true is that it is a result of practical deliberation of a certain sort. As Samuel Freeman explains, for constructivism, "moral principles are correct (or true or reasonable) when they are the outcome of a deliberative procedure that

---

contrasts with having 'extrinsic' value, and (2) something having 'non-instrumental' or 'final' value, in contrast to being merely' instrumentally' valuable for the sake of realizing some further end.

incorporates all of the relevant criteria for correct reasoning."[5] On a constructivist view, our normative judgments are correct insofar as they issue from the *right kind* of procedure of construction where appropriate constraints are placed upon our rational deliberation.

In SN, Korsgaard identifies the correct deliberative procedure for action in general in terms of conforming to what she calls our "practical identities." What is a practical identity? As Korsgaard defines the term, one's 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."[6] To act in accordance with a practical identity is to reflectively endorse reasons for action based upon some specific conception of ourselves. Most of these identities are contingent. For example, a person might conceive of herself as a mother, a daughter, a member of a profession, or a citizen of a state. Some identities are more central to our lives than others. The demands these practical identities make on us take priority over others that are less fundamental to who we are. All of our practical identities in general serve to ground normative claims. As Korsgaard argues, "Your reasons for acting express your identity, your nature; your obligations spring from what that identity forbids."[7]

For Korsgaard, however, we have one practical identity that is not contingent, namely, our 'humanity'. Humanity is, as she defines it, our "identity

[5] Freeman, Samuel. 2007 *Rawls,* Routledge, New York. 292.

[6] Korsgaard, Christine. 1996. *The Sources of Normativity,* Cambridge University Press. 101.
[7] Ibid., 101.

simply as *a human being*, a reflective animal who needs reasons to act and to live."[8] As she later explains, we are 'human beings' insofar as we "need to have practical conceptions of [our] identity in order to act or to live," where this most fundamental identity "stands behind" all the other particular identities we might have.[9] Because of its special status, Korsgaard concludes that we must value our own humanity—as well as the humanity of everybody else—unconditionally. We can reconstruct her argument as follows:

1. We need reasons in order to act or to live.[10]
2: Our reasons for action derive from our practical identities, or conceptions of ourselves under which we value ourselves.[11]
3. Our commitment to any specific practical identity arises from our humanity, or our need to act in conformity with practical identities in general.[12]
4. Therefore, if we value any specific practical identity, we must value our own humanity[13]
5. Therefore, if we are to act at all, we must value our own humanity.[14]
6. "To treat our human identity as normative, as the source of reasons and obligations, is to have… [a] 'moral identity.'[15]
7. "Valuing humanity in your own person rationally requires valuing it in the persons of others."[16]

This is meant to be a transcendental argument. Korsgaard begins with a claim that she thinks we all accept, namely, that we need reasons in order to act, where these reasons are grounded upon what we value. She then attempts to show that in order for this to be the true, we need to value our own humanity, which grounds the very

[8] Ibid., 121.
[9] Ibid., 129.
[10] Ibid., 121.
[11] Ibid., 101.
[12] Ibid., 129.
[13] Ibid., 123.
[14] Ibid., 123.
[15] Ibid., 129.
[16] Ibid., 121.

possibility of valuing in general. She builds upon this result to argue that, on pain of rational inconsistency, we must be moral. Morality requires that we must value humanity as it is found in other people as well.[17]

There are three main corollaries related to Korsgaard's argument. In the sections below, I show how all three lead to problems for Korsgaard's overall view. Call the first and most fundamental corollary the 'Reflective Endorsement Thesis' (RET):

> REFLECTIVE ENDORSMENT THESIS: Practical identities must be reflectively endorsed, that is, we must affirm them as being valuable.

Without the RET, Korsgaard's view doesn't get off the ground. Korsgaard embraces the RET because she thinks that our conceptions of ourselves spring from the fact that we are self-reflective creatures. The concept of reflective endorsement is a necessary, crucial feature of Korsgaard's account. When we come to regard a practical identity as expressive of ourselves, we reflectively endorse that identity as something we value. Call the second corollary the 'Source of Reasons and Obligations Thesis' (SROT):

> SOURCE OF REASONS AND OBLIGATIONS THESIS: Practical identities give rise to reasons and obligations.

Korsgaard offers the following considerations in support of SROT. The practical identities we adopt explain why we take some of our desires to be reasons for us and not others. For example, we might exclude some desires because they seem 'alien' to us, because we do not identify with them or want to be moved by them.

---

[17] For a thorough discussion of these kinds of transcendental arguments, see Skidmore 2001.

We view other desires instead as reflecting our commitments, as providing genuine reasons for action. Call the third corollary the 'Lexical Ordering Thesis' (LOT):

> LEXICAL ORDERING THESIS**:** Reasons for action are lexically ordered as follows:
>
> a. Reasons for action related to our practical identity qua moral agents always take priority over our other reasons for action.
> b. Other reasons for action are ordered by how central the specific practical identities related to such reasons are for our lives.

Some reasons for action are only prima facie obligatory. Others count as reasons we have all things considered. We regard some reasons for action as overriding. LOT explains why this is the case. Our practical identities are lexically ordered in terms of how central they are to our sense of identity, which is in turn explained by what we endorse. For Korsgaard, our most fundamental practical identity is our humanity, and it is this identity that gives rise to our specific moral obligations. As a result, we are unconditionally committed to morality. We have other practical identities, however, that, while more contingent, are conceptions of ourselves to which we are deeply committed. These might include our identities as parents, members of certain religions, and so on. Other practical identities are less fundamental. For example, someone might value being a soccer player, but they would easily give it up if it conflicted with a practical identity that they valued more.

In the remainder of this chapter, my argument proceeds as follows. First, in Section II, I discuss a standard objection to constructivism raised by Russ Shafer-Landau, which I apply to Korsgaard's own position. Second, in Sections III-IV, I will argue that it is Korsgaard's commitment to reflective endorsement in particular that makes it vulnerable to both horns of the dilemma raised for a more general account

of constructivism posed by Shafer-Landau. I will show how her analysis of both non-moral and moral practical identities ultimately fails to satisfy her own specified criteria for an adequate solution to the Normative Problem. Third and lastly, in Section V, I raise an objection to Korsgaard's overall approach in terms of her appeal to the idea of reflective endorsement in general.

General Objection to Constructivism

As Russ Shafer-Landau argues, there is a fundamental problem with constructivist views.[18]. He presents this as a problem for constructivism about morality, but it is also a problem for constructivism about normativity more generally. The worry is that constructivism faces a basic dilemma: either it lapses into a substantive moral realist position or else it fails to guarantee that, by following the constructivist procedure of construction, we will arrive at normatively valid outcomes. We can reconstruct Shafer-Landau's argument as follows:

1. Either the initial conditions for the procedure of construction are normative or they are not.
2. If the initial conditions for the procedure of construction are normative, then this entails that there are certain normative claims that exist independently of and therefore do not arise from the procedure of construction itself.
3. If the initial conditions for the procedure of construction are not normative, then such a procedure cannot guarantee normatively valid outcomes

[18] Shafer-Landau, Russ. 2003. *Moral Realism: A Defence,* Oxford University Press, New York. 41-42.

How do we determine how to set up the procedure of construction in the first place? On the one hand, we could be led by normative considerations that are independent of and prior to the procedure itself. By definition, such considerations cannot themselves be constructed from the procedure of construction. Instead, they must be normative principles to which we are antecedently committed, which is just to adopt a form of substantive moral realism. On the other hand, if this procedure of construction is grounded upon non-normative considerations, then it is unclear how this starting point can guarantee results that are normatively acceptable.

In the next two sections, I show that the two basic parts of Korsgaard's view—her account of both non-moral and moral practical identities—fall prey to the two different horns of this dilemma. The non-moral part of her view ends up being overly subjective. The moral part of her view fails by her own criteria, in part because it involves claims that can only be justified by some appeal to substantive moral realism.

Problems for Non-Moral Practical Identities

In this section, I focus on problems related to Korsgaard's account of non-moral practical identities.[19] Specifically, I argue that Korsgaard's normative framework, which involves acting in conformity with our various reflectively endorsed practical identities, does not provide us with the right kind of normative guidance in non-

[19] Hannah Ginsborg (1998) raises a different type of concern for Korsgaard's position on non-moral identities. She argues that Korsgaard's Kantian framework entails that actions that result from our non-moral identities are not really free actions.

moral cases. My first concern is that Korsgaard's account is not able to explain *why we should ever adopt any particular practical identity in the first place*. I will argue that she is not able to adequately explain this because she gives unwarranted normative authority to the practice of reflective endorsement, which can generate identities with no intuitive normative force. If this is true, then Korsgaard's account fails to adequately answer the normative question. My second concern is that, for similar reasons, her view does not provide a satisfactory account of *the kinds of identities that we shouldn't adopt*. If this is true, then Korsgaard's account is overly subjective and contradicts our fundamental normative intuitions. Third and lastly, Korsgaard's view provides no way to adjudicate among competing non-moral practical identities. If this is true, then Korsgaard's account fails to adequately answer the normative question *and* fails to capture our fundamental normative intuitions. I will suggest that all of these problems are brought about, at least in part, by the necessary role that reflective endorsement plays in her account.

The first concern is related to Korsgaard's commitment to the Reflective Endorsement Thesis or RET as follows. Our practical identities provide us with a first-person perspective from which to make choices regarding which desires we should act upon. They furnish us with reasons for choosing in the way that we do. But how do we come to adopt any particular practical identity? To take on an identity, we must reflectively endorse it or affirm it as valuable. In order to do this, however, we need reasons for valuing it. But on Korsgaard's view, our reasons are supposed to come from our practical identities themselves. So which comes first: (1) our practical identities, which are supposed to provide us with normative reasons

for action, or (2) our normative reasons for actions, on the basis of which it seems we should decide to adopt one practical identity over another? If we do not have a practical identity that provides us with reasons to act a certain way, then it seems we cannot have reasons for forming that specific practical identity. But if we do not have reasons for why we choose to adopt a certain practical identity in the first place, then our choice of this specific identity seems groundless or arbitrary. [20]

Of course, on Korsgaard's view, there is one practical identity that we are committed to if we are to value anything at all, namely, our identity as human beings. Perhaps this identity provides us with reasons to adopt particular contingent practical identities. We need reasons to act and to live, so we must adopt conceptions of ourselves as human beings that we value. This might explain why we adopt practical identities in general as opposed to none at all, but it cannot address why we endorse any *particular* practical identity. All our humanity requires is that to act and to live we must have *some* conception of ourselves. It doesn't provide any guidance in picking out what that conception should be.

Moreover, it is not clear that reflective endorsement has any normative force whatsoever, nor is there any feature of Korsgaard's account that establishes the necessary connection between reflective endorsement and the purported normative status of the practical identities that we take on through the process of endorsement. So, for example, there is no reason to think that the fact that someone

[20] See, for example, Brady 2003 who discusses somewhat similar concerns, although he fails to consider the possible reply Korsgaard could make to this objection as discussed below.

endorses a conception of themselves as, say, macho and strives to act as a macho person would, confers any normative status whatsoever on being macho.

My second concern is related to both RET and its corollary, the Source of Reasons and Obligation Thesis, or SROT. The problem is that there are some practical identities that we might reflectively endorse that do not intuitively give us reasons for action at all. Korsgaard's view is not in a position to tell us which practical identities we should *not* accept. There are some practical identities that, although they do not conflict with morality *per se*, still seem like identities we should not endorse. For example, consider Rawls' famous description in *A Theory of Justice* of a person "whose only pleasure is to count blades of grass in various geometrically shaped areas such as park squares and well trimmed lawns."[21] Or consider Nagel's example of "a man who wastes his life in the cheerful pursuit of a method of communicating with asparagus plants."[22] We are not inclined to think that a practical identity that provides one with reasons to count blades of grass or to attempt to cheerfully communicate with asparagus has much value at all, nor do those reasons seem to have any real normative force. As a result, if one takes reasons to be inherently normative, one might not count the outputs of the endorsement process to be genuine reasons at all. On Korsgaard's view, however, there is no reason why we shouldn't adopt these identities, so long as they don't conflict with other identities we have that are more fundamental to who we are.

---

[21] Rawls, John. 1971. *A Theory of Justice,* Harvard University Press. 432-433.

[22] Nagel, Thomas. 1970. Death. *Nous* 4 (1) 73-80.

Though we might never encounter people like those described by Rawls or Nagel, there are many real people who take on eccentric identities that don't intuitively have much value at all. People with obsessive compulsive disorders, for example, view it as an important part of who they are to be clean and meticulous, to always have a tidy house or to always deliberately step on every square of the sidewalk when they walk down the street. We find so little value in these identities that we when we see someone identify in this way, we encourage them to seek professional assistance in order to change who they are and value different things. We don't see the endorsement of identities of this type as providing genuine normative reasons for action.

There are also less extreme cases than these in which psychological disorders are not in play. In a world where the Internet has become so central to our lives, many people lose themselves online, adopting virtual identities in chat rooms that they place more value upon than almost anything else. Others strongly associate with specific characters that they create in various role-playing video games. When people spend excessive amounts of time engaged in these types of activities, our typical intuitions are that they should get their priorities straight and pursue things of real value. We don't typically take one's endorsement of one's identity as, for example, a level 97 wood-elf, as providing real normative reasons for action.[23]

[23] One might respond that the only way to understand the gamer's actions is to stipulate reasons for action. These, however, are explanatory reasons, and I take Korsgaard's position as an attempt to identify the source of justificatory reasons.

If our specific practical identities necessarily provide us with reasons for action, then we must conclude that in the cases described above, these people do have reasons to act in these ways. The idea that such people have reasons all things considered to engage in such activities, since these practical identities are central to who they are flies in the face of our ordinary intuitions. Our ordinary judgments would likely be that the person who suffers from obsessive-compulsive disorder should seek help. When it comes to the person who wastes their life playing video games at the expense of all other identities he could take on, we tend to think that such a person should learn to practice some moderation. He never has any reason to spend the amount of time that he does playing video games.

More specifically, if conforming with these practical identities does not necessarily violate or conflict with any overriding moral duties, and if such roles are most fundamental or central to a person's identity, then she has a reason all things considered to act upon the various reasons prescribed by such identities. There is nothing in Korsgaard's view that tells us why we shouldn't adopt these identities that don't seem to have much value. In this way, Korsgaard's account is objectionably arbitrary and fails to provide grounding for justificatory as opposed to mere explanatory reasons for action. It allows for all sorts of identities to be construed as normative, including those that our intuitions tell us we should not be valuing.[24]

[24]Korsgaard 1999 provides an account of what goes wrong when people perform actions that are immoral that may seem to provide a reply to this worry I have just raised. She argues that true actions are performed by agents acting in accordance with their "best constitution." An agents "best constitution" will be a one that is unified. Unified agents will act in light of practical identities that are consistent.

My third and final concern is related to the Lexical Ordering Thesis or LOT, specifically part *b* or the claim that, as stated earlier, "our non-moral practical identities are ordered in terms of how important they are to us." The problem is that Korsgaard's account of normativity provides us with no way of adjudicating among the various reasons which competing non-moral reflectively endorsed practical identities provide us. There may be any number of practical identities we have endorsed, all of which are of equal value to us, which can sometimes come into conflict with one another. Imagine the following situation. A person, Elizabeth, is CEO of a company she helped to build from the ground up. Throughout the years, she has put a tremendous amount of effort into the company and her identity as a CEO is very important to her. Her company is having an important meeting with its stockholders which cannot be rescheduled and that will dramatically affect the company's future. However, at the same time, her only daughter will be graduating from high school. Let's say that the identity she has as a mother and the identity she has as a CEO both matter equally to her. In this case, Korsgaard's account of normativity in terms of reflectively endorsed practical identities will not provide her with adequate normative guidance. For any reason she considers, she can ask herself "Why should I take reasons for action related to this reflectively endorsed

---

When we act immorally, we fail to reflect fully on the nature of our constitution. Such reflection would reveal that the value we place in our humanity provides us with obligations to act morally. Korsgaard might respond to my criticism by suggesting that when we take on identities that do not seem valuable, we make a mistake. We fail to see that adopting the identity prevents us from being fully unified agents. I don't think that this response is satisfactory. First, it is certainly possible that viewing oneself as a grass-counter does not conflict with one's moral identity or any other identity that one values. Second, one can raise the normative problem for this aspect of Korsgaard's account. We can ask "Why should I care about being a unified agent?"

practical identity to have more priority than reasons for action generated by the other practical identity that I have endorsed to the same degree?" Furthermore, Korsgaard's answer fails to capture our normative intuitions as well. A person might have two identities that she values equally. Someone might value her practical identity in terms of her job as much as her role as a mother. We tend to think, however, that at the end of the day, our relationship to our children is ultimately more important than achieving success in the corporate world. Korsgaard's view cannot account for these intuitions. Her position allows us to value different practical identities equally which are intuitively not of equal importance. It may also be the case that there *is* no fact of the matter regarding which of the two identities she values more. Her account provides us no way to resolve this issue. Here, again, her view seems objectionably arbitrary.

Problems for Moral Practical Identities

In this section, I argue that Korsgaard's account of moral practical identities does not provide an adequate answer to the Normative Problem. Recall that Korsgaard claims that the type of value involved in justifying normative claims must be both intrinsic and final. I shall argue that she fails to establish that our humanity has these types of value. I shall argue further that her commitment to the claim that humanity has intrinsic value exposes her to the first horn of Shafer-Landau's dilemma, that is, it commits her to substantive moral realism.[25] Lastly, I challenge

[25] I intend this as an internal critique of Korsgaard's view. I am not committed, at this point, to any particular position regarding substantive moral realism. The

Korsgaard's Lexical Ordering Thesis or LOT, a corollary of the reflective endorsement thesis, by arguing that she fails to show that our moral practical identities should have overriding priority over all our other identities. If Korsgaard has not shown that we really are unconditionally committed to morality, then she has provided no motivation to see reasons for action endorsed by our moral practical identity as something we should unconditionally value.

Let's consider Korsgaard's transcendental argument for the claim that we must value our own humanity—as well as the humanity of everybody else—unconditionally.

1. We need reasons in order to act or to live.
2: Our reasons for action derive from our practical identities, or conceptions of ourselves under which we value ourselves.
3. Our commitment to any specific practical identity arises from our humanity, or our need to act in conformity with practical identities in general.
4. Therefore, if we value any specific practical identity, we must value our own humanity
5. Therefore, if we are to act at all, we must value our own humanity.
6. "To treat our human identity as normative, as the source of reasons and obligations, is to have… [a] 'moral identity'"[26]
7. "Valuing humanity in your own person rationally requires valuing it in the persons of others."[27]

In what follows, I will be particularly concerned with claims 4, 5, and 7. I begin with an examination of how we should understand the notion of 'must' involved in premises 4 and 5. We can interpret this in two very different ways. Both interpretations create problems for Korsgaard's view. The first reading of 'must,' the way it is used in claims 4, 5 and 7, involves some kind of psychological necessity.

---

substantive moral realist position, however, is one to which Korsgaard is attempting to provide an alternative.

[26] Korsgaard 129.

[27] Ibid., 121.

That is, there is some feature of our human psychology that makes it the case that we cannot act if we do not simultaneously value our own humanity. But this claim seems to be just empirically false. Even if (3) is true, that is, even if our commitment to any of our specific practical identities arises from our humanity, or our need to act on practical identities in general, it may not be the case that we actually value our humanity itself. For example, imagine that I have a practical identity as a student, one that I reflectively endorse and value in my various pursuits. It does not seem to be the case that, psychologically speaking, I must be simultaneously valuing my humanity, that is, my identity as a creature who necessarily operates with some practical identities whenever I act. I may never even reflect upon the fact that I have such an identity at all. One additional worry is that, on this psychological reading of 'must,' the fact that we unconditionally value our humanity now rests upon merely empirical foundations. But if this is the case, then our commitment to morality is not unconditional. It instead rests upon a purely contingent psychological fact.

On the second reading of 'must', the way it is used specifically in claim 7, "must" instead involves a kind of logically necessity. That is, if we unconditionally value our own humanity, then, upon pain of rational inconsistency, we must unconditionally value the humanity of everybody else as well. The main problem with this interpretation is that it does not commit us unconditionally to morality either. Instead, it makes our commitment to the unconditional value of the humanity of others merely derivative of a more fundamental identity we have, namely, one of being a rationally consistent thinker; if we don't care about whether

our behavior is consistent, we no longer have an argument for why we should value the humanity of others in the first place.

On a third and more plausible reading of 'must', Korsgaard's claims in 4, 5, and 7 involve a kind of transcendental necessity. That is, in order to value anything, we must value our own humanity since it constitutes our very capacity to value anything at all. But understanding the value of our humanity in this sense—namely, that it represents the capacity for us to value anything else—only shows why we must see our own humanity as valuable in an instrumental sense, as the condition for the possibility of valuing anything else. This again, however, fails to show why we must unconditionally value our own humanity, since our humanity becomes merely instrumentally valuable for the sake of realizing any other ends we might value for their own sake.

None of these three readings of how Korsgaard uses the term 'must' in claims 4, 5, and 7 establishes the conclusion that we are unconditionally committed to morality. The fact that Korsgaard's argument fails in this way points to a more fundamental problem with her account, namely, that she fails to show that the type of value that we place upon our own humanity or upon the humanity of others must be understood as either intrinsic or final. Recall that Korsgaard claims that these features are essential to any adequate solution of the Normative Problem. She writes:

> The entity that brings a regress of justification to a satisfactory end must *combine* these two conceptions. It must be something that is *final,* good or right for its own sake, *in virtue of its intrinsic properties,* its intrinsic structure.

As mentioned before, Korsgaard understands the distinction between 'final' or 'non-instrumental' and 'instrumental' value in terms of how we value something. Things with 'final value' are valued for their own sake rather than for the sake of realizing something else. Korsgaard does not establish that this is true of humanity. All her argument shows is that we must value our humanity *if we want to act.* We must value our humanity for the sake of our being able to act at all. A similar point applies to the type of value that the humanity of other people has for us. As we have seen, Korsgaard's argument does not show that we are committed to unconditionally valuing others for their own sake. All that it shows is that we are required to value the humanity of others if we want to be rationally consistent. We don't value the humanity of others in a final or non-instrumental way; we value it for the sake of consistency. And if Korsgaard has not established that our humanity or the humanity of other people has final value, then she has not provided an adequate response to the Normative Problem. Recall that an answer to the normative problem must also provide a satisfactory response to the regress problem. That is, the source of our normative obligations must be such that it would be incoherent to ask "But why should I care about *that?"* But it does seem perfectly coherent to ask "But why should I care about acting, or about being rationally consistent, in the first place?"

Korsgaard also fails to show that our humanity has intrinsic as opposed to extrinsic value. As she writes elsewhere, this is a distinction "between things which have their value in themselves and things which derive their value from some other

source."[28] Korsgaard can establish each person's own humanity has intrinsic value. As she argues:

> 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.[29]

On Korsgaard's view, the value of our own humanity derives from the fact that we endorse it as a normative identity for ourselves. That is, the value of our humanity comes from the fact that we ourselves simply *are* human beings. Thus, the value of humanity is intrinsic insofar as it comes from no source other but our own self-reflexive valuing of ourselves.

But Korsgaard has not shown that the value of the humanity of others is similarly intrinsic. For Korsgaard, things have value because we value them. She writes: "It is the natural condition of living things to be valuers, and that is why value exists."[30] But if this is the case, then it seems that the humanity of other people has value for us *because* we value them. And if this is true, then the value that the humanity of other people has must be extrinsic. That is, the source of their value is *us*. It is the fact that *we* value humanity in others that makes it valuable.

On behalf of Korsgaard, we might reply to such concerns by arguing that the value of the humanity of others does not rest in the fact that we value it but that *other people value their own humanity*. In this way, their humanity has 'intrinsic'

---

[28] Korsgaard, Christine. 1989. 'Two Distinctions of Goodness', *The Philosophical Review,* vol. 92, no.2, pp.169-195.
[29] Korsgaard 1996 121
[30] Ibid.,161.

value in exactly the same way that ours does since they also self-reflexively value themselves. But even if this is true, this still does not show why *we* should unconditionally value their humanity in the way that morality requires. To be committed to this claim is to be committed to the idea that we must value all things which are intrinsically valuable simply because they derive their source from themselves. But it seems perfectly coherent to ask at this point, "Why should we care about *that*?"

In order to establish this, we need to invoke some additional normative principle, one that insists that we should value anything that has intrinsic value. But this principle does not follow from Korsgaard's constructivist procedure of construction itself, but instead seems to be an independent normative principle we are just brutely committed to. Indeed, it is simply a central tenet underlying all substantive moral realist views, a normative principle that is binding on us entirely independently of our beliefs or attitudes about it. Therefore, Korsgaard's argument for why the intrinsic value of the humanity of others unconditionally commits us to morality itself only succeeds if she abandons her constructivism and embraces a normative principle defended by substantive moral realism itself.

To summarize the results of the last two sections, we have shown how Korsgaard's account of normativity in terms of acting in conformity with both moral and non-moral practical identities that we reflectively endorse falls prey to both horns of the dilemma raised by Shafer-Landau. When it comes to non-moral normative questions, the framework of practical identities does not yield normatively valid reasons. It gives us no normative guidance when it comes to

taking on practical identities in the first place, identifying the practical identities that we shouldn't adopt, or adjudicating among competing identities that are endorsed equally. This is because reflective endorsement itself has no normative power.

Korsgaard's account of moral identities impales her on the other horn of Shafer-Landau's dilemma. She cannot explain why we must value the humanity of others without committing herself to the realist principle that we ought to value things that are intrinsically valuable.

Reflective Endorsement

As we have seen, Korsgaard is fundamentally committed to the Reflective Endorsement Thesis or RET, which claims that "practical identities must be reflectively endorsed, that is, we must affirm them as being valuable". This thesis also underlies her Source of Reasons and Obligations Thesis or SROT since our practical identities can only be the source of our reasons or obligations insofar as we have reflectively endorsed them in the first place. Indeed, Korsgaard goes so far as to identify our 'humanity' simply with our ability to reflectively endorse our desires as reasons for action in general. Broadly, this chapter has attempted to take issue with the importance Korsgaard places upon reflective endorsement. In this section I will focus on the particulars of her account of reflective endorsement. I take the account to have three key features:

1. The general stance an agent A adopts when pursuing her desire *x* to be a reason to act is the first person deliberative perspective.

2. In order for her desire *x* to count as a reason for action for A, A must see the desire x as a reason to act in light of some practical identity she endorses, i.e., some normative conception A has of herself.
3. In order for her desire *x* to count as a reason for action for A, *s* must rationally deliberate about and affirm her desire *x* as something she genuinely values.

Consider (1) first. In determining whether or not we should accept it, we should ask whether all of our actions necessarily involve or arise out of the first person deliberative perspective. This seems decidedly false. Consider the case of purely habituated actions. I have a close friend who always walks on the left side of the person he is accompanying. I spend a lot of time with this person and so have become accustomed to walking on the right. One day, another friend brought it to my attention that whenever I walk with her, I always make sure I am walking on the right. If I found myself on the left, I would move. I didn't even realize that this was the case since I did it routinely. My always moving to the right of whomever I am walking next to is an action I perform. Nonetheless, it is not something I self-consciously endorse from a first-person deliberative perspective. Rather, it is simply a behavior I have become habituated or accustomed to perform. Such cases are commonplace. We wake up in the morning and stumble into the bathroom to take a shower or brush our teeth. Often we do not adopt a first person perspective and consciously deliberate upon whether we reflectively endorse these reasons for action. We perform such actions out of sheer habit.

Consider (2), the claim that we must always act in light of some practical identity or other. A person may self-consciously adopt a particular identity, and yet the reasons he has for acting in a certain way may not come from that identity or

from any other identity that he explicitly endorses. In *Unprincipled Virtue*, Nomy Arpaly describes such a case as follows:

> Imagine Peter, who believes, in his own words, that "morality is for wimps." He advocates a quasi-Nietzschean view according to which one should be selfish and strive to increase one's own power. Yet Peter does not perform wrongfully selfish acts and he performs many unselfish acts for unselfish motives. When asked about this, he offers rationalizations as if he were rationalizing the breaking of a diet ("But I *was* being selfish, no *really)* or sometimes he blushes and says, honestly, "Well I guess I am a wimp." But he continues to act well nonetheless.[31]

The case Arpaly discusses here shows that a person can act even though the reasons they have for acting do not come from any conception of themselves that they value. They might be blamelessly ignorant about what their actual practical identity even is. Intuitively, cases like Peter's are not examples of weakness of will. Our intuition about Peter's case is not that he is being overwhelmed by some compulsion and so fails to act rationally. Rather, our intuition is that Peter is really a good guy but simply doesn't realize it. He is a moral person who just unfortunately misunderstands what his character is really like. Thus, he acts on reasons that are normative for him even though he never reflectively endorsed them in light of some practical identity he affirms.

Arpaly discusses many cases of this general sort, such as when a person has certain inclinations that they, for whatever reason, refuse to adopt as an identity. Consider the case of a lesbian who strongly desires to be heterosexual. Her homosexual desires appear to her as intruders in her own mind, impulses she wishes would go away. Her homosexuality can motivate her to act, but when it does

[31] Arpaly, Nomy. 2003. *Unprincipled Virtue*, Oxford University Press, New York. 8.

so, she views the act as an instance of weakness of will. In this case, our normal intuition is that when the woman acts on her homosexual desires, she is not acting out of weakness of will. Rather, she is acting on legitimate reasons, though she is unwilling to explicitly affirm the practical identity that such desires seem to imply.

These types of cases show that we can act even if our reasons for action are not desires that we ever accept from the first person deliberative perspective. Moreover, we may think a particular identity provides us with reasons for action when it turns out we are just *mistaken about who we are.* This will be a common critique of all of the philosophical positions in this book that employ the method of reflective endorsement.

Consider the case of somebody who thinks of himself as a serious film critic with impeccable taste. This is how he represents himself to all of his friends and family. He goes on about the virtues of various obscure foreign films, the merits of certain camera angles and lighting techniques, and he praises the effective use of artistic, nonlinear approaches to filmmaking. The man entirely embraces this conception of himself. Nonetheless, every time he suggests going to a movie with his friends, he always picks a vapid action film or romantic comedy. In this case, the man's conception of himself is simply wrong. He is utterly self-deceived and does not make decisions about which movies to see on the basis of the practical identity he endorsed as an astute film critic at all. All the cases described here seem like paradigmatic examples of action. But if so, then it seems that requirements (1) and (2) above are implausibly strong restrictions upon what counts as genuine reasons

for action. They rule out all sorts of things that we ordinarily take to be full-fledged human actions.

Counterexamples to (3) are abundant. We perform actions every day that do not involve any self-reflective affirmation or endorsement. Habituation, or some other form of less than conscious motivation produces many of our behaviors. The idea that reflective endorsement must, at some stage, be involved with our actions over-intellectualizes our everyday experiences.

Conclusion

In this chapter, I have argued that Korsgaard's attempt to solve the normative problem is not successful. Each of my criticisms pertains, fundamentally, to the value that Korsgaard places on the process of reflective endorsement. I have argued that reflective endorsement can't, on it's own, explain either our conditional commitment to non-moral obligations our unconditional commitment to moral obligations. The fundamental reason that this is so, or so I have argued, is that the process of reflective endorsement itself is not sufficient to generate outputs that are truly normative.

In the next chapter, I will turn to a thinker—Harry Frankfurt—who shares many of Korsgaard's commitments, but parts ways with her in many respects. One crucial point of similarity is their commitment to the importance of reflective endorsement. A crucial dissimilarity is the role that reflective endorsement plays in their respective accounts.

CHAPTER III

REFLECTIVE ENDORSEMENT AND THE CONCEPT OF A PERSON

Harry Frankfurt makes use of the concept of reflective endorsement to account for a number of different philosophical phenomena. He argues that we use reflective endorsement to identify with certain mental events and distance ourselves from others. Identification is crucial to his philosophical program because it is necessary, for personhood, freedom[32], and moral responsibility. Like Korsgaard, Frankfurt also maintains that identification and endorsement are the source of reasons for action, though his explanation for why this is so is quite different from the one that Korsgaard offers.

In this chapter, I will first provide an account of the role that reflective endorsement plays in Frankfurt's philosophical system. I will then argue that Frankfurt's account of caring, which relies on the notion of reflective endorsement, rules out intuitive cases of genuine caring.

Free Will and the Concept of a Person

In his groundbreaking paper, *Alternate Possibilities and Moral Responsibility*[33], Frankfurt argues that the contemporary debate on the issue of free will and moral responsibility rests on a mistaken assumption. He calls this assumption *The*

---

[32] The literature that comprises the free will debate is vast and is beyond the scope of this dissertation. I will not engage that debate here. I discuss Frankfurt's account of freedom here because of its connection to Frankfurt's account of personhood.
[33] Frankfurt, Harry "Alternate Possibilities and Moral Responsibility" *The Journal of Philosophy* Vol. 66, No. 23 (Dec. 4, 1969), 829-839.

Principle of Alternate Possibilities (PAP). PAP is the principle that a person is morally responsible for an action that they perform only if they could have done otherwise. Frankfurt thinks that PAP is mistaken because the question of whether an agent could have done otherwise is irrelevant to the question of whether that person is morally responsible for what they have done. He argues that, even if an action is overdetermined, that is, even if the same action would have been performed no matter what the agent did, the agent is both free and morally responsible if he performed the action because he *wanted* to perform the action.

On many occasions throughout his body of work[34], Frankfurt provides an example of a person he calls "the willing addict." The willing addict is addicted to drugs, but does not know that he is addicted. When the opportunity arises for the addict to indulge in his drug of choice, he takes the drug *because he wants to take it*. Unbeknownst to him, he would not be able to refrain from taking the drug even if he wanted to. If he didn't want to take it, his addiction would kick in and he would find himself incapable of refraining. Even if the addict would take the drug no matter what, when he takes it because he wants to take it, he performs his action freely *and* he is morally responsible for it.

Frankfurt's willing addict is free in this case because the decision to take the drug issues, not from anything alien or external to him, but from *the very features that make him a person.* Frankfurt contends that the concept of personhood has not received the philosophical attention that it deserves. The term "person" is appropriately applied, not to a particular *biological* type, but, instead, to a certain

[34] Frankfurt, Harry, *The Importance of What We Care About: Philosophical Essays* Cambridge University Press, 1988.

philosophical type. The criteria for personhood should capture, "those attributes which are the subject of our most humane concern with ourselves and the source of what we regard as most important and most problematical in our lives."[35] For Frankfurt, this will be the basic structure of our wills. He sets out, then, to provide an account of the basic structure of the will. To begin, he distinguishes between what he calls "First Order Desires," "Second Order Desires," and "Second Order Volitions." We'll begin with First Order Desires.

> *First Order Desire:* "simply [a] desire to do or not do one thing or the other."[36]

First Order Desires are the kinds of desires that we share in common with non-human animals. Both a non-human animal and a person may, for example, desire to consume sugar. We can imagine that it may actually be unpleasant for both the non-human animal and the person to have this desire. To alleviate the unpleasant sensation, the non-human animal has only one option—to satisfy the desire. Or, perhaps, they will be lucky enough to become distracted and forget the desire altogether. The person, by contrast, can alleviate the unpleasantness of the desire in more than one way. Like the animal, the person could satisfy the desire. Unlike the animal, however, they could also be rid of the desire by *changing the desire itself.* This can be done through the formation of what Frankfurt calls "Second Order Desires."

> *Second Order Desire:* A desire to have or not have a particular kind of desire.

---

[35] Ibid., 12.
[36] Ibid., page 12.

Consider the case of a lifelong smoker. The smoker has the first order desire to smoke the cigarette. He experiences this desire as a strong compulsion. He understands that smoking is bad for him. He finds the social stigma unpleasant. He has a second order desire about his first order desire—he desires to not desire to smoke. He desires his first order desire to be different from the one that he actually has.

A Second Order Desire is not enough, in itself, to motivate a person to action. A person can desire to desire all sorts of things. The smoker can desire to no longer desire to smoke. The reluctant student may desire to desire to work harder. The compulsive eater might desire to desire to eat healthier. But we can imagine that, in many of these cases, nothing ever comes out of having a second order desire for your first order desire to be different. Having a meta-level desire about a first order desire won't motivate action unless the agent desires that second order desire *to be their will.* Frankfurt calls these components of the will, *Second Order Volitions.*

> *Second Order Volitions: "*Someone has volitions of the second order when he simply wants to have a certain desire or wants a certain desire to be his will."[37]

The formulation of second order volitions requires introspection. As human beings, we take our endeavors and ourselves seriously. Frankfurt says, "Taking ourselves seriously means that we are not prepared to accept ourselves just as we come. We want our thoughts, our feelings, our choices, and our behavior to make sense. We are not satisfied to think that our ideas are formed haphazardly, or that our actions are driven by transient and opaque impulses or by mindless decisions. We need to

[37] Ibid., 16.

direct ourselves—or at any rate to believe that we are directing ourselves—in thoughtful conformity to stable and appropriate norms. We want to get things right."[38] To know whether we ought to identify with or endorse an action or a conception of ourselves, we need to check it against other things that we care about. To do this, we need to take a step back and take an evaluative stance toward our attitudes themselves. He says,

> We are unique (probably) in being able simultaneously to be engaged in whatever is going on in our conscious minds, to detach ourselves from it, and to observe it—as it were—from a distance. We are then in a position to form reflexive or higher order responses to it. For instance, we may approve of what we notice ourselves feeling, or we may disapprove; we may want to remain the sort of person we observe ourselves to be, or we may want to be different."[39]

The capacity for reflective endorsement is what makes taking this kind of a stance possible. We are now in a position to identify a central component of Frankfurt's position.

> *Identification as Endorsement Thesis*: We identify with a mental state if and only if we reflectively endorse that mental state.

And, relatedly, we can identify another crucial feature of his position pertaining to the nature of personhood itself.

> *Personhood as Identification Thesis:* We exhibit personhood when we identify with our mental states.

Now we are in a position to see the role that reflective endorsement plays in Frankfurt's account of personhood. Identification with a mental state is what makes

---

[38] Frankfurt, Harry, *Taking Ourselves Seriously and Getting it Right*, (Stanford University Press: 2006) e-book. Location 74.
[39] Ibid., Loc. 91.

a person a person. What it is to identify with a mental state is to reflectively endorse that state. Therefore, the capacity that we have to evaluate and endorse or reject a mental state is what makes us persons.

To return to Frankfurt's example, we can say that the willing addict's decision to take the drug is an expression of his personhood. The addict doesn't simply want to take the drug; he wants to want to take the drug. He has endorsed and has therefore identified with the desire to take the drug.

Frankfurt distinguishes his willing addict from addicts of two other types. The first type of addict is purely at the mercy of her addiction. She has formed no second order volitions about taking drugs. She goes where her addiction takes her. This type of addict is the type that Frankfurt would describe as "the wanton." Of the wanton, Frankfurt says,

> What distinguishes the rational wanton from other rational agents is that he is not concerned with the desirability of his desires themselves. He ignores the question of what his will is to be. Not only does he pursue whichever course of action he is mostly strongly inclined to pursue, but he does not care which of his inclinations are the strongest.[40]

Frankfurt thinks that human beings rarely if ever behave as wantons *all of the time*. However, it is theoretically possible that someone could. Such a person would not exhibit personhood. He or she does not reflect on his or her desires and, as a result, does not identify with any of them, and, as we have seen, endorsement and identification are essential to personhood.

The second type of addict knows herself to be an addict. She finds herself swept away by the strength of the addiction. She doesn't want to take the drug—it

[40] Frankfurt, Harry, *The Importance of What We Care About: Philosophical Essays* Cambridge University Press, 1988. 17.

is not a desire with which she identifies. Nevertheless, she takes the drug. An addict of this type, though she may exhibit personhood regularly, performs actions that do not issue from those aspects of her character that make her a person in this particular case. In this case, she does not behave freely and is not fully morally responsible for what she does.

The willing addict, by contrast, exhibits personhood because she endorses and therefore identifies with her desire to take the drug. Here, it will be useful to point out a way in which Frankfurt's account differs from Korsgaard's. Korsgaard maintains that reasons that are genuinely normative issue forth from the practical identities that individuals endorse. We are committed to the normativity of our endorsements because of our fundamental commitment to the value of choice that makes endorsing such identities possible at all. Imagine that the willing addict truly identifies with her conception of herself *as an addict.* On Korsgaard's view, it seems that this way of identifying gives the addict reasons with genuine normative force for taking the drug. Now, intuitively, we might think that she has no reason to take the drug and that, indeed, she has good reasons to *refrain* from taking the drug. It's not clear that Korsgaard's view can capture that intuition.

By contrast, Frankfurt is not committed to the existence of reasons that are somehow fundamentally grounded in anything. Indeed, Korsgaard is trying to answer a different question, The Normative Question. She is trying to provide an account of the fundamental grounding of normative claims. Frankfurt is attempting to provide an account of a different philosophical concept—the concept of

personhood.[41] Frankfurt's view is thus immune from the criticism that a decision to willingly take drugs cannot be fundamentally grounded. As we will see, he will agree with Korsgaard to some degree that our endorsements provide us with reasons for action, but he will disagree that those endorsements are fundamentally normatively grounded. His view admittedly contains a certain element of subjectivity when it comes to reasons.

Identification and Reasons for Action

Frankfurt thinks that motivating reasons cannot be external to the self. For something to count as a reason, an agent has to care about it in some way. However, not just any desire counts as a reason. Our second order volitions create reasons for us. That is, the desires that create reasons are those desires that we endorse or those with which we identify.

This account of reasons allows us to distinguish reasons from alien impulses. He discusses a case in which he feels an alien impulse to shoot his own son, whom he dearly loves. There is a desire involved here, though it strikes the agent as foreign and unpleasant. Frankfurt argues that, though there is a desire of *some* magnitude present in this case, it provides him with no reason whatsoever to shoot

[41] It is likely that Korsgaard would say that her account is an account of personhood also—a Kantian account according to which personhood consists in our ability to make autonomous choices. The point that I am stressing is that Korsgaard sets out to provide a solution to the Normative Problem, and to do so, she invokes a certain concept of humanity (broadly construed) and personhood. She thinks that this account can fundamentally ground normativity. Frankfurt is not attempting to arrive at fundamental grounding for normativity.

his son. The desire revolts him—he wholeheartedly disavows it. It is only the desires with which we identify that set for us real reasons for action.

In his later work, Frankfurt develops an account of caring and love that further explicates his notion of the relationship between identification and reasons for action. It is to that feature of his account that we will now turn.

Caring and Love

Frankfurt highlights the fundamental importance of *care* and *love* to human life. He emphasizes that there is a differences between merely wanting or desiring a thing and caring about that thing. An addict may want a drug, but it does not follow that he or she *cares* about the drug (though, of course, some addicts might). Caring about things is a matter of endorsing or identifying with them and is a position that we arrive at through introspection and reflection. Caring also has a temporal component. He says, "when we do care about something, we go beyond wanting it. We want to *go on* wanting it, at least until the goal has been reached. Thus, we feel it as a lapse on our part if we neglect the desire, and we are disposed to take steps to refresh the desire if it should tend to fade."[42] The caring entails, in other words, an ongoing commitment to the object of care. He says, "By our caring, we maintain various thematic continuities in our volitions."[43] Though all instances of caring involve second order volitions, the converse is not true. Not all second order volitions represent instances of caring because of the temporal component involved

[42] Frankfurt, Harry, *Taking Ourselves Seriously and Getting it Right*, (Stanford University Press: 2006) e-book.
[43] Ibid.

in caring. He says, "Caring about something implies a diachronic coherence, which integrates itself throughout time." I'll call this general requirement Frankfurt's *Dispositional Requirement for Caring.*

> *Dispositional Requirement for Caring*: If we care about something, we are "disposed to take steps to refresh the desire should it tend to fade."

Frankfurt also identifies a third category of evaluative attitude—there are some things that we come to love. When we love things, we often have very little, or no control at all, over whether we love them. The things that we love and that we can't help but to care about are what Frankfurt calls "volitional necessities." He says, "The objects of our love represent our most fundamental commitments and provide us with overriding reasons for action. When we love something, we see it as having value in itself, and we see the interests of the thing or the person that we love as worthy of pursuit for their own sake."[44]

Like Korsgaard, Frankfurt identifies the source of our most compelling set of obligations as the set of things that we would die rather than to give up. For him, these things are volitional necessities. Frankfurt, again, like Korsgaard, also has something of a lexical ordering thesis for reasons. The things that we love provide us with our most compelling reasons for action, followed by the reasons that are provided by the things that we care about. Like Korsgaard, then, Frankfurt thinks that what explains the value of the things that we care about is the very fact that *we value them.*

One of the most striking differences between Frankfurt's view and Korsgaard's is that Frankfurt does not think that our commitment to these identities

[44] Ibid., 229.

is tied in any way to the commands of rationality. We can perfectly well recognize that some thing or other is required by the demands of rationality, but we may find that the recognition that it so follows doesn't motivate us in any way. This may be because we simply don't care amount the demands of rationality. He says, "Explaining to a person that he has violated the requirements of rationality may lead him to regret and be ashamed of his error, but in itself provides him with no basis at all for feeling guilty about what he has done."[45] The simple understanding of the demands of rationality isn't enough, in itself, to generate the reactive attitudes associated with moral judgments.

He argues further that an analysis of a person's purely rational processes doesn't tell us anything at all about their character, but it is assessment of character that is relevant to moral assessment. He claims, "Our response to sinners is not the same as our response to fools."[46]

Unlike Korsgaard, then, Frankfurt does not maintain that our status as beings that need reasons to act and to live antecedently commits us to morality or to the demands of rationality. Rather, in order to live, we must be beings that *love* and *care about* things. The things that we love and care about that create reasons for us, nothing deeper or more fundamental, like a fundamental respect for humanity. He says,

> I do not believe that anything is inherently important. In my judgment, normativity is not a feature of a reality that is independent of us. The standards of volitional rationality and of practical reason are grounded, so far as I can see, only in ourselves. More particularly, they are grounded only

[45] Ibid.
[46] Ibid., Loc. 258.

> in what we cannot help caring about and cannot help considering important.[47]

Though nothing has fundamental value on Frankfurt's account, he does think that there are some things features of our external environment that human beings overwhelmingly tend to care about in common. These features include an intense commitment to the external conditions that provide for our own continued survival, or our commitment to the survival of our children.

Reflective endorsement plays a substantial role in Frankfurt's account of free will, moral responsibility, and personhood. It is also essential to his account of what it is to care about something and what it is to love something. Because caring and loving are the source of reasons for action, reflective endorsement is also essential to his account of reasons for action. In what follows, I will provide some challenges to the claim that reflective endorsement really should play such a prominent role in our understanding of all of these important philosophical concepts.

Caring Across Time

In this section, I will raise some concerns for Frankfurt's account of caring. Specifically, I will take issue with the idea that caring requires renewed endorsement across time. I will provide cases in which a person cares about something even though it is not a thing that they have endorsed consistently (or at all) through time. Instead, the caring in these cases will be demonstrated by the agent's *behavior* across time. I'll suggest that dispositions to behave consistently count as better evidence of caring than dispositions to realign one's second order

[47] Ibid., Loc. 363.

volitions. Agents that ignore their consistent behaviors in favor of their second order volitions seem to be ignoring something important about who they are.

As we have seen, Frankfurt thinks that reasons arise out of the things that we care about. If my arguments for the claim that some caring does not require identification are compelling, then it follows that, if all instances of caring provide us with reasons for action, some reasons are not generated by identification, and are, therefore reasons that have their source in something other than a process of endorsement.

In an earlier section, we looked at Frankfurt's *Dispositional Requirement for Caring:*

> *Dispositional Requirement for Caring*: If we care about something, we are "disposed to take steps to refresh the desire should it tend to fade."

To test this requirement, consider the case of Jane the aspiring lawyer. Jane identifies with her desire to be a lawyer. The causes of her identification with this desire are complex. She's been raised in a family that values high earning potential, she views being a lawyer as consistent with her general commitment to justice, she has respect for friends and family members that are lawyers, and so on. When she introspects, she affirms the value of a career in law. Accordingly, Jane graduates with her bachelor's degree and she enrolls in law school.

An additional fact that is true of Jane is that, throughout her childhood, she frequently created art in her free time. Her parents repeatedly emphasized throughout the years that interest in the arts was for entertainment only. Pursuit of a career in the arts or humanities simply wasn't practical or valuable. Jane took these lessons to heart, and whenever she reflects on her desire to create art, she

disavows it as frivolous—as almost a weakness. Nevertheless, all these years later, the basics components of Jane's character remain unchanged. She is now in law school, but her law notebooks are filled with fairly skilled drawings of people and landscapes. Her eyes glaze over when she's left to complete her readings for class, and she finds herself sketching or painting when she should be studying. She tries, repeatedly, to recommit herself to the study of law. Jane is disposed to try to refresh her desire to study law when it tends to fade. On many occasions, when she recognizes that she is sketching rather than studying, she tries to direct herself back to her proper course. Over time, however, she loses the ability to focus on her legal studies. By Frankfurt's lights, she has stopped caring about the law. She flames out of law school. Years later, however, she still finds herself sketching and painting.

If we apply Frankfurt's analysis of caring to Jane's case, the following seems true: (1) Jane once cared about studying law, (2) Jane no longer cares about studying law because she is no longer disposed to refresh the desire when it tends to fade. Frankfurt's analysis seems to get both of these cases right. But what would Frankfurt say about Jane's commitment to create art?

Jane's commitment to being a lawyer required constant refreshing. Indeed, she couldn't muster up the motivation to study *without* an active refreshing of her desires. By contrast, Jane *did* feel a persistent desire to create art. Her disposition to create art was so persistent that she *did not even need to refresh it*. It is appealing to say that Jane cares about creating art very deeply. The fact that a revival of her desire to create art *isn't necessary* suggests that she cares more about creating art

than she does about becoming a lawyer, not less. She cares about it so much that her commitment to it never even requires her conscious attention.

The intuition that Jane cares about creating art cannot be captured by Frankfurt's account of caring, but his dispositional insight is useful. If we think of caring in terms of being reliably motivated to behave, we can capture the intuition that Jane cares about art.

> Caring as Reliable Motivation Thesis: An agent, S, cares about something (P), if S is reliably moved to action by a desire to promote P.

This principle is able to capture all of the intuitions generated by the cases above. Early in her schooling, Jane is reliably motivated to pursue her interest in law. During that time, it seems correct to say that she cares about law. When she is no longer reliably motivated to study the law, it seems correct to say that she no longer cares about it. It also allows us to account for the apparent care that Jane has for creating art.

This motivational reliability thesis has several additional explanatory benefits. First, it seems like external observers are often in the position to know what people care about. Suppose that Congresswoman Smith is running for re-election in her district. She lives in a district in which a lack of clean air has caused much concern for the majority of her constituency. At town hall meetings, when distressed voters raise these concerns, she behaves as if she too is deeply concerned by lack of access to breathable air. She is re-elected. After the election, her actions demonstrate no commitment to solving problems of air quality. She is consistently motivated by other considerations—considerations that run counter to the interests

of those concerned by the health and safety risks posed by poor quality air. It seems reasonable for the voters in this case to say, "Congresswoman Smith misrepresented her position on air quality. She doesn't actually care about it." It would not be possible for external observers to determine whether Congressmen Smith cared about air quality if that caring was purely a matter of some state internal to the congresswoman and invisible to the citizens.

The motivational reliability account of caring also has a second explanatory virtue that Frankfurt's account does not have. People care about some things more than they care about others. This, too, is often obvious to external observers. Congresswoman Smith may care about clean air after all; she just might not care about it as much as she cares about getting re-elected in the future, which requires the financial contributions of major corporations who are interested in reducing environmental regulations. The motivational reliability account of caring can better explain how we are in a position to say that people care about some things more than they care about others. Congresswoman Smith may have a disposition to refresh her desire to pursue clean air should that desire happen to fade. She might also have a disposition to refresh her desire to earn money for her re-election campaign, should that desire happen to fade. She has then, in both cases, met the basic requirements of caring put forth by Frankfurt. But these two desires turn out, in the real world, to be mutually exclusive. Intuitively, her actions reveal which of these two things she cares about more, and it turns out to be her re-election. The motivational reliability account is in a better position to explain this fact.

This general discussion highlights what I take to be the main problem with accounts that implement reflective endorsement components to solve major philosophical problems or to account for noteworthy features of human experience. Reliance on reflective endorsement puts emphasis on what we *take* ourselves to care about, but what we *seem* to care about may be distinct from what we *actually* care about. There are studies in empirical science that support this contention. We'll turn to those in chapter six.

The Transparency of Caring Critique

Frankfurt's account of caring puts each individual in a privileged position with regard to knowledge pertaining to what they care about. In this section, I will argue that the level of privilege we are afforded is unwarranted. I take Frankfurt to be committed to all of the following epistemic claims:

1. An agent can't care about something without knowing that they care about it.
2. An agent can't be wrong in the judgment that they care about something.
3. An agent can't be wrong in their assessment that they *don't* care about something.

The reason that an agent has this kind of infallibility on Frankfurt's account is that caring is a state that is actually *constructed* by an internal process performed by an agent. It requires reflection on and avowal of one's attitude's toward the subjects of caring. If the agent performs the process, the agent, by definition, cares. I will argue that 1-3 are counterintuitive.

Let's consider (1) first. Imagine a man, Ted, who is not very self-assured. No one who knows him well fails to detect deep currents of insecurity surging through his personality. Some have even said, of him that "he would rather die than be humiliated." Ted's actions are constantly aimed toward avoiding humiliation and ensuring that everyone around him views him as intelligent and successful. Ted is not aware of this aspect of his own personality. He certainly has never reflected on his need to avoid humiliation. Yet, the desire to avoid humiliation motivates Ted so reliably, it seems implausible to say that it is not something that Ted cares about. His need is to avoid humiliation is so reliable that people can frequently predict his behavior in certain kinds of circumstances.

Similarly, imagine Linda, who hates crowds. She avoids shopping malls, never goes to sporting events or amusement parks, and does her grocery shopping in the middle of the night. Linda has never carefully reflected on her dislike of crowds, she has simply found herself feeling unpleasant when crowds are present. It seems like the only plausible way to describe her actions is to say that Linda cares about avoiding crowds. It seems like both Ted and Linda care about things without knowing that they care about them.

Now let's look at (2). Peter loves to appear well read at cocktail parties. To this end, he looks up the most noteworthy thinkers in history on Wikipedia to get a general sense of the nature of their philosophy, poetry, art, etc. He arrives at social functions ready with one-liners about Marx, Nietzsche, Klimt and Wagner. The foundation for certain important aspects of his self-esteem is that he cares about education and culture. The truth is, however, that Peter does not *actually* care about

either education *or* culture. What he cares about, though he doesn't recognize it, is that he *appears* both educated and cultured. It seems, in this case, that Peter is wrong in his assessment that he cares about education and culture.

Finally, let's look at (3). An agent can be wrong in their assessment that they *don't* care about something. Consider Julia, a lifelong student of philosophy. Julia has recently decided that her commitment to philosophy has been a waste of time. Her considered opinion, when she reflects, is that she has spent ungodly amounts of time considering questions without answers. She always dreamed of making a difference in the world and her official take on the matter now is that the study of philosophy will never provide her with anything she needs to really make concrete changes in the real world. Thinking long and hard about the matter, she disavows the study of philosophy. She decides to dedicate her time to other pursuits. And yet, Julia finds herself reading new philosophy books in her down time. She gets excited when her friends discuss a new response to the liar's paradox, and she is downright incapable of extricating herself from philosophical discussions on social media. Despite Julia's protestations, despite the "official position" that Julia advances on the subject, it seems that, after all, Julia does still care about philosophy. Her reliable behavior makes that clear.

The idea that reflective endorsement is crucial for caring grants a special epistemic status to an agent with regard to what they care about. This status ensures that an agent can't care about something without knowing that they care about it, an agent can't be wrong in the judgment that they care about something, and an agent can't be wrong in their assessment that they *don't* care about

something. I have argued that each of these claims is implausible. The question of what a person cares about should be resolved by appealing to a more diverse description of an agent than simply what they affirm upon reflection.

Guidance Critique

Frankfurt emphasizes the importance of the concept of guidance at various points throughout his work, and he might appeal to the notion of guidance to address some of the concerns that I've raised here. He might take issue with the motivational reliability thesis because, though the examples I have provided above are examples of persistent behavior, they are not examples of guided behavior. In the titular paper of his book of essays, *The Importance of What We Care About,* he highlights the importance of guidance for the notion of caring. He says, "As for the notion of what a person cares about, it coincides in part with the notion of something with reference to which the person guides himself in what he does with his life and his conduct."[48] The endorsement component of caring is what makes guidance possible. Guidance is reflexive. Motivational reliability is not enough, on Frankfurt's view, because behavioral patterns are "discernable…even in the lives of creatures who are incapable of caring about anything."[49]

He says, further, that "the notion of guidance, and hence the notion of caring, implies a certain consistency or steadiness of behavior; and this presupposes some degree of persistence. A person who cared about something just for a single

[48] Ibid., 82.
[49] Ibid., 83.

moment would be indistinguishable from someone who was being moved by impulse. He would not in any sense be guiding or directing himself at all."

There is much to discuss here. Let's begin with Frankfurt's idea that caring something for a single moment is indistinguishable from a person who was moved by instinct. This looks like a very external way of determining whether caring is taking place. It may well be true that the two acts are indistinguishable (though it may also, in many cases, be false—more on that later). More important for our purposes is that this test for whether the two things are indistinguishable does not seem to be consistent with what Frankfurt has to say elsewhere. One significant source that seems to be at odds with this account is Frankfurt's original view on free will and moral responsibility. Consider Frankfurt's own paradigm case of compatibilist free action and moral responsibility. Briefly put, Smith is contemplating the question of whether he should shoot Jones. Black wants Jones dead. He has planted a chip in Smith's brain that can detect whether or not Smith decides to shoot Jones. If Smith wants to shoot Jones, and acts on desire, the chip does nothing. If it detects that Smith will not shoot Jones, the chip activates and causes Smith to shoot Jones. Smith's decision to shoot Jones is overdetermined. It might even be indistinguishable from the way it would look for Jones to shoot Smith because he was forced, for whatever reason, to shoot Smith. But on Frankfurt's own view, If Jones shoots Smith because he wants to shoot Smith, then Jones shoots Smith freely and is morally responsible for his actions. It looks like the case that he is describing is, by hypothesis, one in which Jones is not merely acting—he is guiding his action. In this case, the question of whether he is guiding his action

seems to be entirely a matter of what is going on with him internally, regardless of how it appears, externally, to anyone else. The distinguishability test, then, doesn't seem like a consistent one for Frankfurt to appeal to when determining whether actions are guided.

Further textual evidence for this claim comes from Frankfurt's account of action in *Identification and Externality.* Here, Frankfurt argues that what distinguishes actions from mere bodily motions are their internality. He doesn't give a full account of the distinction between the internal and the external in that paper, but he does argue that one key difference is that identification is a necessary feature of those things that are internal. So, the determining factor for whether a bodily movement counts as an action or not has to do with whether identification is happening internally or not. This description of the distinction between actions and mere bodily movements seems to be a far cry from an external distinguishability test. In other words, why would it matter that the actions are indistinguishable?

It appears that the distinguishability requirement for guided action is something of a red herring. It may however, be correct to say that there is some value to behavior guided by internal states that are reflectively accessible. It doesn't follow, however, that *guided* behaviors are the only kinds of behaviors that demonstrate that caring is present.

The Importance of What We Don't Care About—First Order Volition Critique

From the reflective perspective, some things matter to us and some things don't. Frankfurt is right about that. But, even if there is such a thing as a human essence, I

will argue, he hasn't made the case that second order volitions, rather than first order volitions, are the source of that essence. In other words, I think it is perfectly conceivable that the essence can be established, in many, if not most cases, by looking at the first order volitions that she is inclined to have. Why, after all, should we identify an agent with the desires that she wants to have rather than the desires that she, in fact, has?

Consider the case of Paul. Paul is the son of an earnest but uneducated man who made his living working with his hands. Young Paul enjoyed school. He liked discussing politics and philosophy (to the extent that it came up early in his young life). Due to circumstances over which he had little control, Paul dropped out of school before he could graduate. He spends his adult life working with his hands for very little pay, and he has difficulty putting food on the table for his wife and the large brood of children that they eventually have together. Despite Paul's difficulties in life, he has, over time, become a staunch Republican. Though his family benefits from various types of welfare programs, he has no doubt that this state of affairs is only temporary. He believes that his family could survive without it, and he doesn't believe in governmental assistance. He believes that if people want to be successful in life, they should pull themselves up by the bootstraps and do something with their life. He doesn't believe that a person's origin story defines them; rather, he believes that every human individual is the sole author of their own story. When he votes, he votes Republican. Republican policies don't really benefit him (in fact, they actually harm him), but he believes that they will come to benefit him in time. He considers himself, as the saying goes, a "temporarily embarrassed millionaire."

Let's consider Paul's volitions. Paul is not a bad person. He is a decent guy who loves his family and tries his best. But his first order volitions do not line up with his second order volitions. Let's look specifically at the kinds of behaviors in which Paul is likely to engage when he is presented with the opportunity to better himself. Recognizing that blue-collar jobs increasingly require a working knowledge of computers and other technology, Paul's employer offers to pay for additional schooling for all employees. Though, on introspection, Paul believes that success in life involves embracing opportunities as they come, Paul does not take his employer up on the opportunity to go back to school. He doesn't attend trainings at work that would increase his overall skill set and make him more eligible for promotions and raises. He becomes increasingly frustrated as younger employees rapidly begin to advance up the career ladder, while he remains on an early rung. He does, indeed, strongly identify with his second order volitions about work ethic and merit based advancement. He simply fails to realize that his first orders desires are simply more telling about what is actually important to him than the set of things that he wants to have move him when he reflects on the matter.

Jean Paul Sartre's discussion of "bad faith"[50] will provide a useful framework to identify what has gone wrong with Paul here. To be in "bad faith" is to live inauthentically. This can happen in a variety of ways. One of these ways is to define oneself too much in terms of one's facticity. A person's facticity is the sum number of things that are true of that person. A person's facticity includes things about their past, such as who their parents were, where they were born, etc. If a person

[50] Sartre, Jean-Paul. 1966. *Being and nothingness; an essay on phenomenological ontology*. New York: Washington Square Press.

identifies too strongly with their facticity, they are rejecting the broader range of possibilities for themselves—they are rejecting what Sartre calls their "transcendence." Just as a person can identify too strongly with their facticity, they can also identify too strongly with their transcendence. They may reject descriptions of themselves that are true in favor of descriptions of themselves that might someday, be true. But then again, those descriptions of themselves may never be true.

I think that Sartre's way of describing bad faith is useful for understanding Paul's case here for a number of reasons. Paul identifies strongly with his second order volitions. He believes, along with Frankfurt, that the set of desires that he wants to be effective in motivating him are the desires that best represent who he is as a person. But, as our example illustrates, Paul is very rarely, if ever, actually motivated by his second order volitions. His first order volitions, though he may not ever even fully realize it, are the desires that actually motivate him to act. In this way, Paul behaves inauthentically. Paul defines himself too much in terms of the second order volitions and fails to acknowledge the facts that actually apply to him—that his motivations are almost always supplied by his first order desires.

Indeed, it might be quite accurate to say of Paul that other people in his life understand his set of motivations much better than he understands them himself. Frankfurt might be tempted to say that what has happened to Paul is just a garden-variety instance of weakness of will. Paul wants his desires to lift himself up from his own bootstraps to be effective, but he simply can't get it to turn out this way at the end of the day. This isn't the only possible description of what is happening with

Paul in this case. He doesn't think he is in bad faith. He simply fails to realize that, in a vast majority of cases, he fails to be motivated by his second order volitions. He thinks that his endorsements control his behavior, but they actually don't.

Conclusion

I have argued in this chapter that Frankfurt's account of personhood paints a picture of persons and of human motivation that is too shallow. I have argued that too much value has been assigned to reflectively endorsed states, and too little has been assigned to other potential demonstrations of personhood. In the next chapter, I will consider the role that reflective endorsement plays in Marilyn Friedman's account of autonomy. I will draw similar conclusions.

CHAPTER IV

REFLECTIVE ENDORSEMENT AND AUTONOMY

In her book *Autonomy, Gender, Politics,*[51] hereafter, AGP, Marilyn Friedman uses a philosophical approach that resembles, in its key features, the other approaches we have considered so far. Like the other thinkers we have considered, Friedman also identifies the capacity to reflect on our own inner states as crucial to some fundamental aspect of our humanity. In AGP, she outlines an account of *autonomy* as reflective endorsement.

For Friedman, Autonomy is normative. Autonomous actions are more valuable than actions that are not autonomous. Regardless of the particular action that is being performed, the action, if autonomous, is valuable for its own sake (at least to the degree that it is autonomous. Other considerations could, potentially, override the value of autonomy). She suggests that autonomy is normative in two ways. First, autonomous actions have some special status over non-autonomous actions because autonomous actions reflect a person's *real self*. The *real self*, for Friedman is the *reflectively endorsed* self. She says, "For choices and actions to be autonomous, the choosing and acting self as the particular self she is must play a role in determining them."[52] Second, autonomous actions are normative in the sense that they provide the justification for our other social practices. So, as we'll see, honoring the capacity for autonomy is what it is to show respect for other human

[51] Friedman, Marilyn, *Autonomy, Gender, Politics*. (Oxford University Press, 2003).

[52] Ibid., Loc. 79-80.

beings. The possibility of social and political participation and the potential for justified public policy is contingent on the capacity to exercise autonomy possessed by the individuals involved. Friedman discusses at great length the implications that her account of autonomy has for political philosophy and social policy.

In this chapter, I will outline and evaluate Friedman's view with an eye toward determining whether a reflective endorsement component provides an effective account of autonomy.

Autonomous Behavior and the Process of Reflective Endorsement

There are several similarities between Korsgaard's account and Friedman's account. One central similarity is their respective commitments to procedural realism rather than substantive realism. As we have seen, Korsgaard thinks that the substantive realist position cannot, even in principle, hope to answer the normative problem because it will always be subject to Moore's open question argument. For any explanation offered for why one ought to perform or value an action, one can always ask, "why should I care about *that*?" In her account of autonomy, Friedman also rejects substantive accounts in favor of a procedural account. Substantive accounts of autonomy maintain that there are some constraints on the kinds of behaviors that count as autonomous. For example, a person cannot autonomously consent to an arrangement that deprives them of their very capacity for autonomy. Friedman thinks that accounts of this type define autonomy too narrowly. She thinks that people are capable of autonomous behavior in a wider range of cases than those which substantive accounts of autonomy would be willing to grant. For Friedman,

an appropriate account of autonomy would identify the procedure that makes an action an autonomous action. As we will see, this procedure, for Friedman, is a procedure of reflective endorsement.

Another key similarity between Korsgaard's account and Friedman's account is that they both emphasize the importance of the different identities that a person takes on from the standpoint of reflective evaluation.[53] For Friedman, in order for an agent to behave autonomously, she must judge her actions to be consistent with one of her "perspectival identities." *Perspectival Identities* are the set of wants, desires, beliefs, and related traits that we embrace from a reflective, evaluative position. This element of her account highlights what we will call the *Reflective Endorsement Requirement for Autonomy:*

> *Reflective Endorsement Requirement* "To realize autonomy a person must first somehow reflect on her wants, desires, and so on and take up an evaluative stance with respect to them."[54]

On this view, engaging in truly autonomous behavior is more than just a matter of acting intentionally. Autonomous action must issue from states that the agent has reflected upon, evaluated, and approved as part of her perspectival identity.

Friedman contrasts our Perspectival Identities with what she calls our "Trait-Based-Identities."[55] Our trait-based identities are comprised of features like our

---

[53] In one sense, Korsgaard and Friedman are engaged in two distinct projects. Korsgaard is providing an account of the source of normativity and Friedman is providing an account of the nature of autonomy. In another sense, however, insofar as Korsgaard's view is Kantian, autonomy is crucial to her account.

[54] Ibid., Loc. 80.

[55] Friedman's discussion of traits centers on identity groups to which a person might belong. In chapter Six, I will be used the term "traits" differently, to connote enduring personality characteristics (or *perceived* enduring personality characteristics).

race, gender, socio-economic status, sexual orientation, and so forth. These traits may dictate the kinds of lives we lead in both obvious and non-obvious ways. For Friedman, however, our trait-based identities are only relevant to the question of autonomous behavior to the extent that we choose to take them on as elements of our perspectival identities. She says,

> On my view, what counts for autonomy is someone's perspectival identity, her wants, desires, cares, concerns, values, and commitments. The non-perspectival kinds or traits she instantiates or exemplifies are relevant to her autonomy only if they matter to her, only if they are features of herself she cares deeply about. Otherwise they do not ground her autonomous choices or actions.[56]

Friedman also proposes a condition that is similar to Korsgaard's Lexical Ordering Thesis. She claims that actions are most autonomous when they reflect what is most important to the agent.

> *Importance Requirement,* "Autonomous actions and choices also stem from what an agent cares deeply about. They stem from wants and values that are relatively important to the acting person."[57]

Our Perspectival Identities are comprised of those desires, wants, and values, that, upon reflection, we realize are the most important to us. The actions that demonstrate the greatest degree of autonomy are those that issue from those most important associations. Autonomous decisions can arise from states that we care about only trivially, but, as we will see, those actions are less autonomous than actions that issue from endorsed states that are more important to the agent. Friedman says,

---

[56] Ibid., 196-198.
[57] Ibid., 109.

> Wants and values are "deep" when they are abiding and tend to be chosen over other competing wants and values. Wants and values are also deep when they constitute the overarching rationales that an agent regards as justifying many of her more specific choices. Wants and values are "pervasive" when they are relevant to a great many situations that a person faces. They are frequently salient in someone's life and she chooses in accord with them often. When someone reflectively reaffirms wants or values that are important to her in the sense just described, they become part of the perspective that defines her as the particular person she is. They embody the "nomos" of her self: relatively stable, enduring concerns and values that give her a kind of identity as the person she is. Someone is self-determining when she acts for the sake of what matters to her, what she deeply cares about, and, in that sense, who she "is."[58]

Friedman claims that autonomous actions are "reflective" in two distinct senses of the word:

1. "They are partly caused by a person's reflection on, or attentive consideration of, wants and desires that already characterize her."[59]
2. "They must reflect or mirror, the wants, desires, cares, concerns, values, and commitments that someone reaffirms when attending to them."[60]

Friedman doesn't spend a lot of time unpacking these two conditions, but it is not difficult to get a sense of what she has in mind. Let's first consider (1). It is not enough that a person performs an action and that action just so happens to comport with the wants and desires that categorize the agent. It also must be the case that the actions are *brought about* by states on which the agent has reflected. There are at least two senses in which this is true: (a) the external world must be such that it allows the agent to make autonomous decisions.[61] As we will see, some external

---

[58] Ibid., 114-120.
[59] Ibid., 92.
[60] Ibid., 103.
[61] If determinism rules out free will, then the external world is not one in which actions can ever be autonomous. Friedman provides a compatibilist answer to this problem. She says, "A compatibilist answer to this criticism, to which I subscribe, is that autonomy s a matter of degree and requires agents simply to harbor the capacities for certain sorts of reflection and agency, however these are acquired or

conditions eliminate the possibility of autonomous behavior entirely. The second sense in which (1) is true is: (b) The mental states that cause an action must be states that have been reflected upon, rather those that have not been reflected upon.

Let's consider some examples in which condition (a) fails to be satisfied. One category of cases in which (a) is violated are cases in which conditions in an agent's external environment are not favorable to autonomous choice. Friedman makes a distinction between factors that are constitutive of autonomy itself, and the causal conditions that make the exercise of autonomy possible. She says,

> The nature of autonomy itself consists of the conditions that choices actions must meet in order to be autonomous. These conditions constitute autonomy. These are distinct from the causal conditions, both past and present that must obtain for choices and actions to manifest the constitutive conditions in virtue of which they are autonomous. The distinction between the constitutive and the causal conditions required for autonomy will be particularly important for appreciating the role that social relationships and cultural context play in the realization of autonomy."[62]

This first category of cases that we will discuss involve cases in which autonomy is not realized, not because the constitutive conditions for autonomy are lacking, but because the necessary causal conditions are lacking that would ordinarily provide the right kind of environment for autonomous action to take place.

For example, imagine that Sarah has, as one of her most fundamental concerns, a desire to feed her family only the healthiest food. While at the

---

are interconnected with the agency of others. Those reflective and practical capacities together with wants and desires must constitute a self who, as a self, plays a determining role in the process leading to her behavior. Self-determination may, ontologically speaking, be merely an intermediate causal process in a causal sequence extending backward and forward to infinity. Such causal connectedness does not undermine its character as the kind of causal stage in the process that it is: the part determination by a self of her own behavior." (AGP Loc 647).

[62] Ibid., 75-78.

supermarket, Sarah picks up what she takes to be her favorite breakfast cereal. In fact, it is a different brand displayed in a similar box. A supermarket employee accidentally placed the imposter in the wrong location on the store shelf because of its resemblance to Sarah's preferred brand. When she arrives home, Sarah realizes her mistake. Happily, it turns out that the cereal she brought home is actually *healthier* (and more delicious) than the one she normally buys. In this case, Sarah is the beneficiary of some good luck, but she did not behave autonomously because the acquisition of this particular cereal was not caused by states that she endorsed. In a case of this type, factors in Sarah's epistemic environment made it the case that Sarah could not ensure that values that she had reaffirmed from a reflective perspective could actually play a causal role in her action.

This insight might be useful to explain why autonomy is lacking in cases in which sufficient information is not provided. It might, for example, be useful for understanding the concept of free and informed consent. An individual cannot truly behave autonomously if they aren't in the right kind of epistemic environment; they can't exercise their autonomy if they don't have the right kind of information. It might also provide us with tools for understanding the difference between autonomous actions and merely intentional actions. Not all intentional actions are truly autonomous. Depending on how the action is individuated, Sarah *intentionally* picked up the breakfast cereal. She did not, however, *autonomously* pick up the breakfast cereal. Picking up this breakfast cereal in particular, given that she didn't know anything about it, was not something she would have endorsed from a reflective perspective.

There are also cases in which an agent's endorsements are barred from being causally efficacious, but do not involve epistemic challenges. For example, if a person is being kept as a prisoner, they will be prevented from acting autonomously. I'll discuss this type of case in more detail later in the chapter.

Now let's consider cases of type (b) above. These are cases in which a person's own mental state deprives her of full autonomy. Cases of type (b) are cases in which the correct sort of external causal conditions are in place for the exercise of autonomy, but the constitutive conditions are not met. Suppose that Sarah, as a matter of routine, does the dishes every night after dinner. She engages in this behavior without reflection on what she is doing or on whether she values what she is doing, while the rest of her family goes on to enjoy their evening, unburdened by chores. Sarah has a husband and several older children who are perfectly capable of doing the dishes and helping around the house. Her behavior in this case would be far from fully autonomous, since she has never really reflected on the question of whether norms or values that require her to do the dishes every night rather than spreading the chores out among the other members of her family is consistent with values and norms that she would actually endorse. Again, for Sarah's behavior to be fully autonomous in this case, it must, be "partly caused by a person's reflection on, or attentive consideration of, wants and desires that already characterize her." Sarah hasn't met this requirement.

Because Friedman's account of autonomy is *procedural* rather than *substantive*, Sarah need not come to some objectively *correct* answer concerning what behavior would actually count as autonomous. What matters for autonomy is

the *process* that Sarah goes through. Upon reflection, she could conclude that the rest of her family ought to be taking care of their fair share of the dishwashing responsibilities. It would also be open to her to come to the conclusion that she really is the one, after all, that ought to do all the dishes. What really matters here, when it comes to the question of whether Sarah is autonomous, is whether her actions are caused by norms that she endorses. Pre-reflection, Sarah is not fully autonomous. If she recognizes that her dishwashing habits are consistent with norms that she endorses, then she behaves autonomously subsequent to that recognition.

Now let's consider (2)—the second sense in which autonomous actions are reflective. Recall that Friedman says, "They must reflect or mirror, the wants, desires, cares, concerns, values, and commitments that someone reaffirms when attending to them."[63] This condition requires that the actions that are candidates for autonomous action must *in fact* mirror the states that the agent has endorsed. Consider another choice that Sarah makes at the supermarket. When choosing breads, it matters very much to Sarah that she feeds her family the recommended daily amount of whole grains. She has certain beliefs about what this means, chief among them that the requirement is never going to be met by purchasing white bread. She knows very little about how to select the healthiest bread, however, so she selects the loaf that has *Heart Healthy* printed in big letters on the bag. She learns later that the bread she selected was actually not particularly heart healthy at all. In fact, *Heart Healthy* was just the name of the product. In this case, though her

[63] Ibid., 103.

intention was to purchase healthy bread for her family, her choice was not fully autonomous because the decision she made did not actually, at the end of the day, mirror the states she reflectively endorsed.

To summarize, Friedman provides a procedural account of autonomy according to which autonomous behavior is behavior upon which the agent has reflected and has endorsed in light of her deeper wants and desires. Autonomous behavior is reflective both in the sense that a process of reflection is at least partially *causally* responsible for an agent's action, and in the sense that the action *actually reflects* or mirrors the agent's wants and desires.

Additional Requirements for Autonomy

The reflective endorsement requirement of Friedman's account is its most fundamental component. There are number of corollaries to the reflective endorsement requirement that are worth mentioning here. First, for Friedman, action is not autonomous if it is coerced.

> *Absence of Coercion, Deception, or Manipulation Requirement* "For self-reflection to be effective in practice, it must not be impeded by interfering conditions. Coercion, deception, and manipulation by others are the paradigm examples of conditions that interfere with the practical effectiveness of someone's self-reflections."[64]

If autonomy requires reflective endorsement, conditions that make reflective endorsement difficult or impossible will also make the exercise of autonomy difficult or impossible. The Absence of Coercion requirement provides at least one of the reasons that oppression and unjust social conditions are problematic—they

[64] Ibid.

prevent those people who suffer under them from being fully autonomous. Some oppressive social conditions will be so restrictive that they will prohibit agents from exhibiting much autonomy at all.

Even when a person is being coerced or manipulated, she may still be capable of exercising some minimal level of autonomy, under certain conditions. Friedman maintains that autonomy admits of degrees. I will call this the "Continuum Thesis."

> *Continuum Thesis:* "Autonomy is a matter of degree. No finite being is thoroughly self-determined. Even self-reflection itself can range along a continuum. The more extensively one reflects on one's values and commitments, the greater is one's autonomy with respect to them."[65]

Human lives, even under oppressive conditions, are often reasonably self-determined. Even in conditions in which options for what to value or how to act are severely restricted, an agent can reflect on the options available to her and endorse those that are most consistent with her overall character. Friedman maintains that autonomy is valuable, even when exercised minimally. Ideally, however, societies will construct social policies that allow for the optimal exercise of autonomy by its members.

The matter of whether an agent can act autonomously will be determined by whether she has developed what Friedman calls "autonomy competency."

> *Autonomy Competency* "…autonomy competency is the effective capacity, or set of capacities, to act under some significant range of circumstances in ways that reflect and issue from deeper concerns that one has considered and reaffirmed."[66]

---

[65] Ibid., Loc. 133.
[66] Ibid., Loc. 231.

Sadly, it is possible for a person to be in a position in which there is no possibility for the development of autonomy competency. For example, individuals that were sold into slavery as children and who remain slaves through adulthood, their lifestyles tightly monitored and controlled, might not have developed autonomy competency. Autonomy competency is developed by repeated opportunities to choose from among various options in a way that mirrors the things that you have reaffirmed are deeply important to you. Those that have been barred from choosing among options could, theoretically, never have developed this competency at all. This kind of situation would be exceedingly rare.

Searching for the "True Self"

My main objection to Friedman's account of autonomy is her identification of the *true self* with set of wants and desires that are affirmed through the reflective endorsement process. Friedman thinks that autonomy is valuable because it involves self-determination. Indeed, one of the main reasons she gives for advocating a procedural rather than a substantive account of autonomy is that a procedural account emphasizes the importance of reflection on the set of affirmed traits that constitutes an individual's "true self."

The objection that I want to raise to this is a standard self-knowledge problem that I have raised throughout this book and that should be familiar by now. People aren't always the best judges of what is truly important to them. Friedman anticipates this objection. She acknowledges that agents may not be in an epistemic

position to know which states are really the most representative of what they sincerely value. She puts the point in the following way,

> A second version of the criticism that self-determination is impossible challenges the presumption that selves can reliably understand their own wants and values, that they are transparently accessible to themselves. To the extent that autonomy depends on self-understanding, it is vulnerable to the vicissitudes of self-knowledge.[67]

She identifies two main sources for the claim that barriers to self-knowledge pose a problem for her account of autonomy. She first looks at psychoanalysis. The premise of psychoanalysis is that, through psychoanalytic therapy, we can come to know facts about ourselves about which we were previously unaware. If psychoanalysis really can serve this function, then it seems like it is not uncommon for us to harbor wants and desires about which we are unaware. If this is the case, it would make autonomy of the type that Friedman outlines difficult, if not impossible to achieve.

Friedman points out that there is some reason to be dubious about psychotherapy, and I agree with her on this point. But she identifies another field that gives us reason to doubt the extent to which we can really know the wants and desires that motivate our actions—empirical social science. She discusses studies done by Nisbett and Ross[68] that suggest that the particulars of our practical environments have much more to do with the decisions that we make than the

---

[67] Ibid., Loc. 669.

[68] Nisbett and Ross, *The Person and the Situation: Perspectives of Social Psychology.* (McGraw-Hill, 1991)

character traits, or sets of wants and desires, that we take ourselves to have. If this is true, when we take ourselves to be acting in a particular way because of some set of wants and desires that we have endorsed, we may be entirely wrong about our own motivations.[69]

She thinks that none of this is inconsistent with her account of autonomy. Reflecting on our behavior in practical contexts and noticing the things that Ross and Nisbett point out may serve as an important test to check to see if a choice made in any given situation is, truly, autonomous. She says, "Thus, rather than undermining the practical significance of attempts at self knowledge, psychoanalysis and empirical social psychology *support* autonomy by enabling us to correct our misconceptions and improve our prospects for self knowledge."[70] I think she gives this objection short shrift, as I will explain in what follows.

Friedman acknowledges that, if what autonomy requires is reflection on a person's internal states, that might sometimes pose a problem for her account of autonomy. She thinks that reflection on our internal states is sometimes, but not always effective. It is possible for an agent to meet Friedman's requirement for agency if, rather than reflecting on her internal states, that agent reflects on her own past behavior. After all, that behavior might be a more reliable source when it comes to the issue of what the agent's beliefs and desires actually are. She puts the point in the following way,

> Self-reflection needs to be reflection on oneself as an agent, but it does not need to be reflection on a private inner realm. It can equally be reflection on

[69] I will discuss studies in empirical social science in much more detail, including those conducted or provided by Nisbett and Ross, in chapter seven.
[70] Op. cit., Loc. 704.

> one's past behavior. It can also be cautiously and narrowly linked to clear-cut evidence. As long as accurate self-knowledge is at all possible, even the frequent occurrence of self-misunderstanding would not undermine accounts of autonomy based on reflective self-understanding.[71]

In this passage, Friedman acknowledges that reflection on one's own past behavior rather one's internal states may sometimes prove to be more telling when it comes to the issue of what the agent really values. The language that she uses here, however, seems to indicate that she does not take the problem to be a serious one. Notice that she uses, in one part of this passage, the language that self-reflection *need not* constitute reflection on one's inner states, but might "*equally* be reflection on one's past behavior." This suggests that reflection on one's behavior is not superior to reflection on one's inner states, but is, rather, equally good. That she chooses to use the language "It can *also be* cautiously and narrowly linked to clear cut evidence" is also telling. It suggests that her account of autonomy can be expressed as a disjunction—that reflection on *either* one's internal states *or* reflection on one's past behavior would be sufficient for the exercise of autonomy. I will argue that there are problems for understanding autonomy as a disjunction of this type for a number of reasons.

1. If the reason that autonomy is valuable is because autonomous actions issue from an agent's *true self,* then if a certain type of reflection doesn't actually accurately identify one's true self, then that reflection isn't valuable, or at least it is not valuable *for that reason.*

[71] Ibid., Loc. 703.

2. (1) seems to suggest a *substantive* truth about what the real self consists in. If this is the case, the answer to the question of what autonomy consists in may be substantive rather than procedural.

(3) There are cases in which reflection on one's past behaviors and reflection on one's inner states generate different, contradictory results.

To illustrate the tension present in Friedman's disjunction, I'll present three test cases. First, let's consider a case that works out well on Friedman's account. Imagine that Martha is a woman who is attempting to make a choice about whether or not to have her own biological child. She is single and is not getting any younger. She is worried that, if she waits any longer, the child will be susceptible to health problems. Martha has read Friedman's account of autonomy and is on board with the theory. She recognizes that, historically, women often feel pressure to become mothers. She wants to make sure that her decision to become a mother is not coerced either explicitly or implicitly by societal expectations. She wants to be sure that her choice in this very important matter is, truly, autonomous.

In scenario one, Martha reflects on her internal states and finds a strong maternal instinct and a passionate desire for children. As it turns out, her past behavior also indicates that she has a strong desire to be a mother. She recalls numerous occasions in the past in which she has either expressed to others or thought privately about how her picture of a flourishing human life fundamentally includes raising another human being. She remembers feeling pangs of emptiness when observing the relationships that her friends enjoy with their children. It is not difficult for her to recall, in the past, exhibiting a certain attentiveness to books and

articles that describe the healthiest practices for raising children. Martha chooses to go through in vitro fertilization so that she can give birth to her own biological child. If there truly is such a thing as an autonomous choice, it seems likely that Martha has made one in this case. Importantly, Martha arrives at the same result, both when she reflects on her inner states and when she reflects on her past behavior.

Now consider Scenario Two. In Scenario Two, Martha's internal states indicate that she wants to be a mother, but her past behavior suggests that she does not. Martha does not respond well to children. Patience is not her strong suit. She has a history of becoming irritable easily when things do not go precisely as she wants them to go. Historically, she has strongly valued the ability to do exactly what she wants to do, exactly when she wants to do it. When she is seated next to young children on an airplane, she feels frustrated and is not empathetic when they begin to cry on the flight. She has nieces and nephews and, though she expresses love for them, truth be told, she finds spending time with them to be something of a burden. She simply doesn't like doing the types of things that children tend to like to do. However, when Martha reflects on her internal states, she finds that she strongly desires to be a mother. It is unclear to Martha whether she should view her past behavior as representative of her "true self" or whether her inner states demonstrate her "true self."

Finally, let's consider Scenario Three. In Scenario Three, if and when Martha reflects on her internal states, they indicate to her either that she simply does not have positive attitudes toward the idea of being a mother, or that she even has a strong aversion to the idea of being a mother. Her past behavior, however, is exactly

the same as that described in Scenario One, and strongly suggests that she does want to be a mother. In this scenario, Martha doesn't know which action is truly autonomous. Should she act in a way that is consistent with what she finds when she reflects on her own internal states? Or, should she recognize that she is really self-blind when it comes to this issue and that the truly autonomous thing to do would be to act in a way that is consistent with her past behavior?

Friedman's Importance Thesis may, perhaps, provide some guidance in the resolution of this question. Recall that, for Friedman, actions that are maximally autonomous issue from wants and desires that are relatively important to the agent. How instructive this turn out to be, however, depends on how we understand the word "important" in this context. If what is meant by "important" is that, upon reflection, the agent values it highly, then it seems like the agent should defer to what she affirms in her reflection on her inner, private realm. After all, it is easy to imagine that the agent's past behavior simply doesn't matter to her at all. If an action is more autonomous to the extent that it is important to an agent, then it doesn't much matter whether she is acting in accordance with what constitutes the best evidence for what really constitutes an agent's "true self."

If, by contrast, if what is meant by "important" is that a consideration or set of considerations frequently moves an agent to action, then it seems that the action that is truly autonomous is the action that is consistent with the agent's past behavior.

The criticism that I am raising here can be understood as a dilemma argument as follows:

A Dilemma Argument Against Friedman's Reflective Account of Autonomy

> P1: Autonomous behavior, issuing from the "true self" is either born from the act of endorsement itself, or it arises out of consideration of the particulars of one's past behavior.
>
> P2: If autonomous behavior, issuing from the "true self," is born from the act of endorsement itself, then there are no substantive considerations (including past behavior) that need to be taken into account in order for an action to count as autonomous.
>
> P3: If autonomous behavior, issuing from the "true self," arises out of consideration of the particulars of one's past behavior, then the act of endorsement isn't sufficient to render an action autonomous.
>
> C: Therefore, either there are no substantive considerations (including past behavior) that need to be taken into account in order for an action to count as autonomous or the act of endorsement isn't sufficient to render an action autonomous.

If the argument is sound, these two possibilities for achieving autonomy aren't really as consistent as they seem.

The first premise of the argument comes from Friedman herself. As we have seen, she thinks that an agent's behavior counts as autonomous if it either comes about as a result of endorsement *or* if it includes careful consideration of one's past behaviors. Textual evidence does not support the view that consideration, on the part of the agent, of their own past behavior is *necessary* for autonomy, but, rather, that such behavior *could be* reflected upon in order to arrive at an autonomous decision.

Textual evidence from most of the work suggests that Friedman thinks that the act of endorsement alone is sufficient for autonomous behavior. For example, she wants to provide space for people who might be entirely blind to their own past behavior, to be understood as autonomous agents. She thinks that, so long as an

agent has autonomy competency, or the ability to reflect and endorse or reject different options and values, they count as autonomous. If there are substantive requirements for what they must consider upon reflection, that starts to look like a *substantive* rather than a *procedural* account of autonomy. She explicitly argues against substantive accounts of autonomy as too narrow in a substantial portion of the first sections of her book.[72]

The justification for the third premise has been provided already in examples above. There are cases in which behavioral evidence is at odds with what we might be inclined to endorse upon reflection. If reflection on the right kind of evidence, namely, behavioral evidence, is what makes an action autonomous, then the endorsement itself isn't doing any work, and, again, under these conditions, the account starts to look substantive rather than procedural.

Self-Blindness

The idea that introspection might not reveal an agent's "true self" is problematic for Friedman's view in a fundamental way. One of the reasons that she favors a procedural account over a substantive account of autonomy is that the procedural account captures the value of self-determination, even in circumstances in which the conditions that might be required for substantive autonomy do not exist. On a procedural account of autonomy, an agent's decisions that issue from reflection are valuable because they issue from an agent's "true self." If decisions made from that

[72] In a later chapter, I will argue that Friedman is onto something when she argues that behavioral evidence should be reflected upon. I will argue that her account can't give the weight to behavioral evidence that such evidence deserves because her account is procedural rather than substantive.

perspective do not truly issue from the "true self," then there is no longer any value in those "autonomous" decisions.

One might find what Nisbett and Ross say about the true motivations behind our actions compelling. One might also agree with Friedman that it is valuable for an agent's actions to issue from their "true self." As a result, one think that the most reliable way to ensure that decisions do, in fact, issue from the "true self" is to reflect on past behaviors. One's "true self" in this case would be revealed by what they consistently *do* or *refrain from doing.* The way to truly act autonomously would, in this case, be to behave in a way that is consistent with the way that one's "true self" would behave.

There are several problems with proceeding in this way. First, it isn't a strategy that is consistent with Friedman's *Importance Requirement*. After all, as I mentioned earlier, the way that an agent has behaved in the past may not be important to the agent at all. An agent might take any number of pro attitudes toward their own past behavior.

Noamy Arpaly's example of the reluctant Nietzschean is applicable here. The example, as you will recall, is the following,

> Imagine Peter, who believes, in his own words, that "morality is for wimps." He advocates a quasi-Nietzschean view according to which one should be selfish and strive to increase one's own power. Yet Peter does not perform wrongfully selfish acts and he performs many unselfish acts for unselfish motives. When asked about this, he offers rationalizations as if he were rationalizing the breaking of a diet ("But I *was* being selfish, no *really)* or sometimes he blushes and says, honestly, "Well I guess I am a wimp." But he continues to act well nonetheless.[73]

[73] Arpaly, Nomy. 2003. *Unprincipled Virtue*, Oxford University Press, New York.

Imagine that Peter is in the grips of an important decision. He wants to behave autonomously. He is sympathetic to the idea that one's past behavior is the best way of determining their "true self." He proceeds to reflect on his past behavior. When he does so, he sees that, time and again, he has performed unselfish acts for unselfish motives. Others have witnessed Peter's behavior and they believe him to be a generally selfless person. In this case, the evidence suggests that others know Peter better than Peter knows himself. For when Peter reflects on his past behavior, he is ready with his standard rationalizations. He believes that, though he performs far more selfless actions than selfish ones, all of this is due to weakness of will. His *real* values are selfish values, he's just too weak to consistently, or even often, demonstrate those values in action.

Peter denies that his past behavior is illustrative of his "true self." Peter's perception of himself—that he embraces Nietszchean values—is a self-deception. Friedman suggests that reflections on our past behaviors can help us refine our conceptions of ourselves and that, as a result, we will behave more autonomously. But what if we reject what we find when we reflect? Why should we understand the actions that follow from Peter's self-deception as more autonomous than those actions that issue from his consistent character?

In a less extreme case, a person may not outright reject their past behavior when they reflect on it, they may simply fail to care about it. Imagine that Jane is about to embark on a new career as a chef. Suppose further that this new career involves spending a lot of money on related supplies—a set of quality knives, high-end cookware, a set of fresh herbs, etc. Jane's financial situation is somewhat dire,

but she thinks that she ought to buy the supplies because they are necessary for her new career. As Jane reflects on her past behavior, she realizes that she has really never followed through with a long-term plan in her life. At the moment, she is full of enthusiasm for her new plan, and though she doesn't reject her past behavior as unconnected to her "real self," she simply fails to care about that past behavior. Jane lives in the moment. She buys the cookware. A month later, she is broke, and she has lost the drive to be a chef entirely. Again, though past behavior may be relevant to who an agent really is, there is no guarantee that an agent will care about it. So it may be the case that the notion of who an agent really is—how we determine her "true self" and the notion of what the agent cares about come apart. Which one is really relevant to autonomy, who one truly is, or what one takes oneself to care about?

Conclusion

I have argued that, like Korsgaard and Frankfurt, Friedman puts too much emphasis on the value of reflective endorsement. She provides some valuable insights about the relationship between endorsement and analysis of past behavior, but, because her account is procedural rather than substantive, she doesn't seem entitled to claims about substantive considerations that must be considered upon reflection. I will share some thoughts on the weaknesses of procedural accounts in a later chapter, but not before considering Allan Gibbard's endorsement-based account of the nature of normative judgments.

CHAPTER V

DO NORMATIVE JUDGMENTS EXPRESS STATES OF ENDORSEMENT?

Normative judgments, when expressed, sound quite a bit like ordinary expressions of belief. "Murder is wrong," structurally, looks quite a bit like the sentence "the book is on the table." When we express the later, we attribute a property to the book, namely, the property of being on the table. Similarly, it may be that when we express the former, we attribute a property to the act of murder—the property of being wrong. If this is what we are really doing, moral judgments and, more broadly, normative judgments in general, are ordinary expressions of beliefs about objects and properties.

There is disagreement over the type of mental state that is expressed when people make normative judgments. *Cognitivists* believe that the similarities that appear to exist between normative judgments and expressions of belief are not deceptive. They maintain that normative judgments *are* expressions of beliefs. *Non-cognitivists* argue that, though normative judgments and beliefs seem to have similar properties, normative judgments express a type of mental state that is distinct from belief. There are a number of different proposals on offer for the mental state that normative judgments express. In this chapter, I will look specifically at one proposal—that normative judgments express the mental state of endorsement.

The proposal that I will consider is the one offered by Allan Gibbard in his book, *Wise Choices, Apt Feelings.*[74] The motivation for his non-cognitivism is the observation that normative judgments are intrinsically motivational. In this sense, Gibbard is what I'll call a *motivational internalist.*

> *Motivational Internalism:* The view that, when an agent, S, sincerely judges that she ought to perform action x, S is motivated, at least to some degree, to perform action x.

In the tradition of Hume and subsequent non-cognitivists like Ayer and Stevensen, Gibbard contends that cognitivism can't account for the motivational nature of normative judgments. He argues that endorsement *can* explain this necessary connection.

In this chapter I will argue that, if normative judgments don't express beliefs but, rather, some other mental state, it is not plausible that the state in question is endorsement or first personal acceptance of norms. I will not argue either for or against non-cognitivism in general. Instead, I will focus specifically on the plausibility of the claim that normative judgments express states of endorsement.

Gibbard is straightforward about the fact that the mental state that he sketches, that of endorsement, or accepting a norm, is not a state that we can define precisely. His analysis, he admits, is psychologically speculative. The evidence to which he points in order to substantiate the existence of such a state, "consist[s] in part in commonsense belief and vocabulary, and in part of observation, both systematic and casual."[75] I will argue that the evidence does not, in fact, support the

---

[74] Gibbard, A., 1990, *Wise Choices, Apt Feelings*, Cambridge: Harvard University Press.
[75] Ibid., 56.

existence of such a unified state of "endorsement" or "acceptance of a norm," and, as a result, Gibbard's norm expressivist account is not successful as an account of the state that normative judgments express.

The Non-cognitivist Tradition and the Move to Expressivism

Hume famously argued that reason alone is not motivational. Our moral judgments, by contrast, by their very nature *do* motivate. Moral judgments, then, cannot be judgments of reason. Hume says,

> Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason.[76]

If we accept what Hume has to say on this point, we are left with the conclusion that normative judgments function more like desires than beliefs. Michael Smith provides a useful way of thinking about the difference between beliefs and desires. We can understand the difference between the roles of these two mental states by evaluating them in terms of what he calls their "direction of fit."[77] Smith points out that when a belief is successful (when it is a true belief), what is in an individual's mind comes to match what is in the world. He calls this a "world–to-mind" direction of fit. If I am in the process of forming a belief about, say, the number of students my logic class, I strive to get the number in my head to match the number of students that really are in the class.

---

[76] Hume, David, David Fate Norton, and Mary J. Norton. 2000. *A treatise of human nature*. Oxford: Oxford University Press.
[77] Smith, M., 1987. "The Humean Theory of Motivation," *Mind*, 96: 36–61.

Desires have a different direction of fit. They have what Smith calls a "mind-to-world" direction of fit. When we have desires, our goal is to get the world to match what is in the mind. When I desire to successfully run a marathon, my goal is to get the state of the world to match what is in my mind—I want to change the world such that it is one in which I run a marathon.

Normative judgments appear to have more in common with desires in this regard because they are concerned with the way the world ought to be. The goal, then, is to get the world to match what is in the mind, not to get the mind to match what is in the world.

If normative judgments structurally seem to have more in common with desires than they do with beliefs, then perhaps such judgments don't actually express beliefs. This would go a long way toward explaining the connection between normative judgments and motivation.

This insight alone, if accurate, is just a starting point. If Hume is right and normative judgments are not expressions of belief, what, exactly, do they express? Gibbard's Expressivist account is quite recent. It was developed, in part, as an alternative to other non-cognitivist accounts that suffered from seemingly insurmountable difficulties.

An early suggestion was that normative judgments express emotions. This is the approach that AJ Ayer took in *Language, Truth and Logic.*[78][79] Ayer's view, called

---

[78] Ayer, A.J. 1936, *Language, Truth, and Logic*, London: Gollancz, 2nd Edition, 1946.

[79] Ayer's motivation for adopting Emotivism had much to do with his logical positivism. He thought that normative judgments had no literal significance because, in his view, they are neither empirically verifiable nor analytic. As I have

emotivism, was that ethical judgments do not express propositions at all. He argues that normative judgments add nothing to the proposition being expressed. He says, "If I say to someone, "You acted wrongly in stealing the money," I am not stating anything more than if I had simply said "You stole the money." In adding that the action is wrong, I am not making any further statement about it I am simply evincing my moral disapproval of it. It is as if I had said, "You stole the money," in a particular tone of horror, or written it with the addition of some special exclamation marks."[80]

Moral judgments often are intended as commands. When this is the case, Ayer argues that the strength of the emotion being expressed varies with the perceived degree of the importance of the command. For example, a person expresses a stronger emotion when they assert, "It is your duty to tell the truth," than they do when they assert, "It is good to tell the truth. " In the first case, the moral emotion expressed is stronger than it is in the second case.

In *Ethics and Language*, Charles Stevenson proposes an alternative candidate for the mental state that moral judgments express. He says, "Moral judgments are involved with *recommending* something for approval or disapproval."[81] Like Ayer, Stevenson thinks that normative judgments add nothing new to the proposition being expressed. Normative judgments are not truth-apt. Instead of adding emotion, however, Stevenson thinks that we express recommendation for or against

---

indicated, there are independent reasons to think that normative judgments express states other than beliefs. Emotivism is, therefore, a non-cognitivist view that is worthy of consideration regardless of the plausibility of logical positivism.

[80] Ibid. 107.

[81] Stevenson, C.L. 1944, *Ethics and Language*, New Haven: Yale University Press.

something when we make normative judgments. In this way, Stevenson is able to account for the phenomenon of disagreement about normative issues in a way that it is not clear that Ayer can. For Stevenson, there are two different types of disagreement: disagreement in belief and disagreement in attitude. Two people disagree in belief when they disagree about the facts. They disagree in attitude when one favors something that the other disfavors. If Bob and Sally disagree about where the ice cream shop is located, they have a disagreement of belief. If they disagree about whether the action of going to get ice cream is favorable or disfavorable, they have a disagreement in attitude.

Traditional non-cognitivist accounts like those provided by Ayer and Stevenson are subject to a problem raised in different places by both Frege and Geach.[82] The problem has come to be known as "the Frege-Geach Problem." It is also frequently referred to as "the problem of embedded contexts." The concern is that, though attempts to understand normative judgments as expressions of something like emotion or recommendation seem to work reasonably well in straightforward cases of normative judgments such as "murder is wrong", they don't fare as well when normative language is used in other ways. For example, normative expressions can be used as the antecedents in conditional statements (e.g., if murder is wrong, it should be illegal). They can be used in questions (Is it wrong to provide tax breaks for the wealthy?) They can also be used as premises in arguments. If normative judgments express emotions or recommendations, then, since such judgments do not have truth-values, they cannot contribute to the

[82] Geach, P. 1960: Ascriptivism. *Philosophical Review 69.*

validity or invalidity of an argument and yet there seem to be straightforwardly valid arguments making use of normative premises. Stevenson is thus unable to account for the obvious validity of arguments such as the following:

(P1) Murder is wrong.

(P2) If murder is wrong, then what John did is wrong.

(C) What John did is wrong.

In light of this problem, if the non-cognitivist position is to be maintained, a satisfactory theory regarding the mental state that normative judgments express should be able to explain how we can use normative language meaningfully in ways other than straightforward expression of moral judgments.

Gibbard's Account

Gibbard proposes that we start from scratch. Instead of beginning with the assumption that normative judgments express beliefs, we should instead assume that they express states of endorsement and then see how much explanatory power that assumption has. He suggests that we "fix on the dictum 'To call a thing rational is to endorse it.' And search for a sense of 'endorse' for which the dictum holds true."[83] He argues that any analysis will strain the concept being analyzed. When we are comparing competing analyses, then, we need to look at the ways in which the concept is being strained. Does the analysis require us to abandon intuitions that are fundamental? His strategy is to show that his own, non-cognitivist, analysis strains the concept of normative judgment less than do the cognitivist alternatives.

[83] Op. Cit., 6.

As a first approximation, Gibbard offers the following account of what it is to say that something is "rational."

> *Expressivist Endorsement Thesis*: "to call something rational [to say that it is "the thing to be done][84] is to express one's acceptance of norms that permit it."[85]

Gibbard's project is to provide a general account of normative guidance and control. He argues that the account he provides makes sense from an evolutionary perspective. Normative systems, like the one that allows us to endorse norms, would be fitness promoting in the human species—they would make members of the species more likely to survive to pass along their genes—because such systems would aid in coordination, both simple and complex. It would have been evolutionarily advantageous for humans to develop different types of motivational systems. Taken together, these systems of motivation provide a general system of normative guidance and control. Context determines which of the systems will be controlling in the circumstances. The first system he identifies is the most basic. He calls it the *Animal Control System*.

> *Animal Control System (ACS)*: The system of motivation that involves bodily appetite, such as hunger, craving, or addiction. Non-human animals also have a control system of this type.

Again, from an evolutionary standpoint, it makes sense that animals of all sorts would develop this kind of motivational system. The animal control system guides us in the pursuit of our basic survival needs.

---

[84] Throughout *Wise Choices, Apt Feelings,* Gibbard uses the language of normativity and the language of rationality interchangeably.
[85] Ibid., 7.

Not all of our motivations are of this type. We are also strongly motivated toward pursuits that don't have to do with our basic survival needs. For example, I might be motivated to engage in certain hobbies, to pursue career goals, or to advance my education. What kind of motivational system can explain motivations of this type? Gibbard proposes a system that he calls "The Normative Control System. " It is this system that he thinks generates normative judgments, so it is this system that will be most important for our discussion in this chapter.

> *Normative Control System (NCS):* The linguistically infused motivational system that involves the acceptance of norms.

To understand how the NCS works, we need to understand what it is to accept a norm.

> *Acceptance of a Norm*: To accept a norm is to be prepared to avow it in normative discussion. When we are prepared to avow a norm, we expose ourselves to demands for consistency.

The notion of Acceptance of a Norm, and, by extension, the notion of the normative control system, has three main corollaries:

> 1*) Imaginative Rehearsal and Consistency Thesis* (IRCT)*:* To effectively enter into normative discussion, one must be prepared to respond to demands for consistency. In order to do this, an agent must engage in imaginative rehearsal to ensure that the things she endorses and is prepared to avow are consistent with one another.

When an agent truly accepts a norm, she will have gone through a process where she has reflected on the norm in absent situations (in a cool hour, in situations in which she is not tasked, at the moment she is considering it, with actually acting on the norm). In these moments, she has concluded that the norm is something she would be willing to avow in normative discussion because she has concluded that it

is not inconsistent with any other norms that she also accepts. Gibbard describes the process in the following way,

> "From this imaginative rehearsal, then, a kind of imaginative persona may emerge, an "I" who develops a consistent position to take in normative discussion. It is then that we can speak most clearly about what the person accepts, he has a worked out normative position to take in unconstrained contexts."[86]

The next corollary is a claim about the connection between imaginative rehearsal and motivation. We'll call this the *Normative Governance Thesis*.

> 2) *Normative Governance Thesis* (NGT)*:* Working out what to do, what to think, and how to feel in absent situations will influence what we do, think, and feel in actual, like situations.[87]

The NGT follows from the IRCT, taken together with Gibbard's commitment to motivational internalism. If we have reflected on and formed judgments with regard to what we ought to do, then, if motivational internalism is true, we will be motivated, at least to a certain degree, to do the thing that we have judged that we ought to do.

The final corollary is a claim about how normative governance works to allow for coordination in groups.

> 3) *Normative Discussion Thesis* (NDT)*:* Working out in groups what to do, what to think and how to feel will involve a move toward consensus through mutual influence and responsiveness to demands for consistency.

Normative discussion is crucial for cooperation and cooperation makes the human species well suited to survive in its environment. Language acquisition was crucial to our ability to coordinate. It provided us with the capacity to convey our needs to

[86] Ibid., 75.
[87] Ibid., 72.

others and to understand others when they conveyed their needs to us. We became more capable of working together to satisfy common goals. The Animal Control System served us well when it came to motivation of a certain sort, but the Normative Control System, made possible by language acquisition, provided us with a unique new form of motivation that effected both individual and group behavior, and further contributed to the fitness we had to survive in our environment.

The guidance that the ACS provides will sometimes be at odds with the guidance that the NCS provides. This conflict, Gibbard argues, accounts for many common cases of weakness of will. If I am on a diet, but I crave a slice of the birthday cake, my ACS and my NCS are at odds with one another. My ACS provides me with a strong craving for the cake. My NCS is more reflective. Though I recognize the strength of my craving, it isn't a craving that I want to have. My NCS recognizes that eating the cake is not consistent with other things that I endorse (e.g., that it is important to eat healthy food, that I desire to maintain a certain weight, etc.). If I succumb to my desires and eat the cake, my weakness of will is explained by the fact that the motivation provided by my ACS was stronger in this particular case than the motivation provided by my NCS.

Not all cases of weakness of will are explainable in terms of a conflict between the ACS and the NCS. Some cases of weakness of will are cases in which nothing like the sensation of craving is involved. These cases often involve competing social motivations.

Gibbard uses the famous Milgram[88] experiments as examples of this kind of weakness of will. In these experiments, subjects were asked to administer electric shocks to other subjects. The subjects that were asked to administer shocks (group A) were made aware that the shocks could be extremely painful and might, potentially, even be lethal. Unbeknownst to group A, shocks were not really being administered. As group A administered what they understood to be shocks, they witnessed subjects hooked up to the device (group B) exhibit signs of pain and distress. Some of the group A subjects were troubled by the fact that they thought they were causing the people in group B pain, but the researchers insisted that the experiment continue. Two thirds of the members of group A complied with the demands of the researchers.

Gibbard argues that the conclusion to draw in these cases is not that the subjects administering the shocks lost all sense of the difference between right and wrong. On some level, even while administering the shocks, they still believed it was wrong to hurt other people. They continued to *accept* or *endorse* the norms against inflicting pain. Despite that recognition, complex social motivations that demand politeness and a willingness to comply with authority were controlling in this particular case. Gibbard argues that the explanation for this kind of behavior has to do, again, with competing motivational systems, but this time there is a conflict of a different sort. He makes the distinction between *acceptance of a norm* and the mental state he refers to as being *in the grip of a norm*. To understand what

[88] Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371-378.

it is to be in the grip of a norm, we first have to understand the notion of *internalizing* a norm.

> *Internalizing a Norm:* To internalize a norm is to have a motivational tendency of a particular kind to act in a way that is consistent with a set of rules that come about as a result of social training, but are not made explicit.

Internalization of a norm works independently of normative discussion. A sophisticated observer might be able to formulate principles by watching the patterns, but no such patterns needs to be explicitly recognized by the person who internalizes the norm. Consider the case of speech volume. We speak softly when we are in a library, movie theater, church, or cemetery. When we are at parties, speaking before an audience, or participating in most kinds of outdoor activities, we speak more loudly. Contingent social circumstances may alter the unarticulated rules, so the standards of appropriateness may vary from culture to culture. Children learn the appropriate level of volume for their voice in certain circumstances just from watching the behavior of others. Children follow these rules, though they might not be able to articulate them.

Emotional responses might also be internalized. There are certain, unwritten rules that govern responses of this type. Those responses, again, vary from culture to culture. For example, different cultures might have different unwritten rules regarding the appropriate conditions under which a male should show emotion that might be different from the circumstances under which a female can show emotion. In many cultures, men are socialized in such a way that they feel it is not appropriate to cry or express distress in public. The men themselves might be

unaware that they have internalized these norms, but they are motivational all the same.

In some cases, when a person has internalized a norm, they are also in a state that Gibbard calls being "in the grip of a norm."

> *The State of Being in the Grip of a Norm*: A person is in the grip of a norm when they find themselves motivated by a norm they have internalized, despite the fact that the norm in question is at odds with norms the person accepts.

Gibbard suggests that the participants in the Milgram study exhibited weakness of will because they were "in the grip" of norms that call for politeness and obedience. Ordinarily it is not bad to be motivated by these types of norms, but in this case, participants were motivated by norms that they did not accept.

For Gibbard, to endorse a norm is not just to express approval of a particular proposition in isolation. When an agent endorses a norm, they take that norm to be consistent with an overall plan, which means that understanding endorsement of norms involves understanding total systems of factual-normative worlds, where factual-normative worlds are complete descriptions of the facts that the agent knows to be true and the set of norms that the agent endorses.

Imagine that Smith is considering whether he should run a marathon. Let *r* stand for the proposition "I ought to run the marathon." We'll say that *W* stands for the factual-normative world that is an expression of Smith's total system of norms and factual judgments.

There are three relations that *r* can stand in with respect to *W*: *r* can be *W* forbidden, *W* permitted, or *W* required. The question of which of these relations *r*

bears to *W* is a factual question. Because it is a factual question, people can agree about whether *r* is W forbidden, permitted, or required without thinking that the total system of norms has anything, whatsoever, to speak for it.

Consider Dani, who is so athletic that it is safe to call her a fitness addict. Dani participates in every athletic activity that she can. In the past, she has participated in athletic events that are so strenuous that they cause her toenails to fall off. Dani is considering whether to participate in an athletic event that involves running through rough, mountainous terrain barefoot. Let's call this norm that she ought to do so, *b* (for running barefoot). Dani's friend, Jon, may agree that, given the total system of norms that Dani accepts, *b* is *W* permitted, and is perhaps even *W* required. Jon can agree with Dani on this point, even while rejecting the idea that *W* ought to be endorsed. Jon may think it crazy to subject oneself to that kind of pain. So the normative aspect of *W* has to do with whether one *ought* to accept it. The factual aspect has to do with which norms are forbidden, permitted, or required relative to the overall system *W*.

The various systems of norms that most people accept will be incomplete. In the real world, no one is fully informed when it comes to facts, or fully opinionated when it comes to norms. Logically, however, there will be a set of completions for each factual-normative world. A system of norms is complete when every possible factual-normative option that an agent might consider in a situation is categorized as either permissible, impermissible, or required. As Gibbard puts it, "The content of a normative statement is the set of factual-normative worlds in which it holds."[89]

[89] Ibid., 97.

Concerns for The Imaginative Rehearsal and Consistency Thesis

Many of the concerns that I will raise for Gibbard's view below are related to one simple point. That point is that according to his position, there are multiple motivational systems, but only the outputs of the NCS count as true normative judgments. I hope to show that this way of looking at human motivational systems gives too much authority to norms that have been reflected upon and are, as a result, capable of being easily articulated.

As we have seen, Gibbard makes a distinction between internalizing a norm and endorsing a norm. Internalizing a norm is important for the purposes of coordination, but all that is required for internalization to occur is for creatures to be socially conditioned to engage in certain patterns of behavior. Endorsement requires more. Creatures become capable of endorsement when they acquire language. Language gives us the ability to formulate our endorsements and allows us to test them against one another for consistency.

The first concern I have for Gibbard's view has to do with the relationship he establishes between imaginative rehearsal and acceptance of a norm. It is often a good thing when a person has worked out in advance the position that they are prepared to avow in normative discussion. The discussion, if and when it happens, may run more smoothly if a person has worked out what they plan to say ahead of time. My concern is that this process rules out far too many cases that should, intuitively, count as normative judgments. In what follows, I will provide some cases

in which an agent has not engaged in imaginative rehearsal but still, intuitively has made a normative judgment.

Consider a case of what I'll call spontaneous normative avowal. Suppose that I survey my Introduction to Ethics Class regarding their positions on a variety of controversial ethical issues. At the point at which I poll them, I have not given any lectures on the topics of the poll. I ask them about the moral status of practices like euthanasia, capital punishment, abortion, and so on. The students provide responses. Some of the students may have considered the issues before, but many of them may not have. This is even more likely to be the case if we make the issues under consideration issues that the students are less likely to have encountered or thought about in the past such as problems in ethics of technology, medical ethics, or environmental ethics. They have had no time to check their endorsements for consistency. Nonetheless, students are able to report their initial attitudes on these topics. Intuitively, the judgments that they offer are, indeed, normative judgments. Of course, throughout the duration of the class the same students will consider a wide range of perspectives on the various topics. The students may find that their positions change as they are exposed to different facts and arguments. In part, their positions will change because they are forced to see that inconsistencies exist between their various judgments. It seems to me that the right thing to say here is that the normative judgments of the students have now changed, not that their initial judgments were not genuine normative judgments to begin with because they had never gone through a reflective process involving imaginative rehearsal.

A change in normative judgment may not always come about as a result of reflection. Some changes might involve spontaneous radical shifts in attitude. Consider the case of a woman who has been a strong, life-long, supporter of the death penalty. She has close friends who have been the victims of terrible crimes. She believes in retributive justice and also believes that the death penalty is an effective deterrent. She advocates publically for the death penalty, seeking to increase the extent to which it is pursued in cases involving the most horrendous crimes. On one occasion, her activism puts her in the position to actually attend an execution. When she does so, however, she has a reaction that she is quite surprised by—she is appalled by what she witnesses. In that moment, her view on the issue changes dramatically. She comes to view the death penalty as inhumane. She is driven to this conclusion, not as a result of reflection on the consistency of her judgments, but, instead, as a result of a strong empathetic response to the person being put to death. The judgment that was born in the moment did not come about as a result of imaginative rehearsal. She did not prepare herself to deal with demands for consistency. In fact, we can stipulate that the attitudes that she holds are not, in fact, actually consistent. She has not had an opportunity to consider the consistency of the things that she endorses. If she had, she would have some hard choices to make. She would have to consider whether her beliefs about the deterrent nature of capital punishment serve to justify the practice, even if it is inhumane. She would have to consider whether she thinks the act of inhumanity against the victim or the pain felt by the victim's families is sufficient to justify the inhumane act. When she gets around to reflecting (if she ever does), she will

inevitably find that she will have to dispense with one or more of the norms that she endorses. Intuitively, however, in the moment where her attitude abruptly changes, she *has* made a genuinely normative judgment. But this intuition is not consistent with Gibbard's view.

One response that Gibbard might offer is that it is unsurprising that in each of the cases that I have described above, we are responding to what is clearly a motivational impulse of some type. But the motivational impulse at play in each case is a motivation generated by a motivational system other than the NCS. In both cases, he might diagnose the judgment involved as response generated by internalized norms, but not necessarily by endorsed norms. If we haven't subjected our judgments to demands for consistency, they are not really normative judgments.

At the outset, Gibbard asked us to start with the assumption that normative judgments express endorsements. He asked us to see how much explanatory power that proposal had and to consider the extent to which the concept of normative judgment is strained by the proposal. When an analysis would have us rule out cases of mental states that seem, intuitively, to be genuine normative judgments simply because the mental states in question have not been reflected upon, the analysis strains the concept. If spontaneous avowals and spontaneous radical changes in avowals intuitively strike us as genuine normative judgments, those types of judgments shouldn't be considered differently simply because of the way Gibbard has carved up our motivational psychologies.

The consistency aspect of Gibbard's account is also problematic for a number of reasons. First, there are consistent judgments that will withstand societal

demands for consistency, but are not judgments that are intuitively "genuinely" endorsed (if there really is such a state of endorsement at all). These examples also illustrate that normative judgments are more complicated than Gibbard suggests. There may be cases in which we are self-blind when it comes to what we "really" think it makes the most sense to do. These are cases, which are not instances of weakness of will, in which norms that have not been explicitly endorsed are more compelling to the agent than the norms that have been endorsed. I contend that in these cases, intuitively, the agent in question is not "in the grip of a norm." If these examples are compelling, they show that the response to a demand for consistency is not a necessary condition for a mental state to express a normative judgment.

Consider the case of Peter, a college student who wants to enter a fraternity. He strongly desires to feel that he belongs in this particular group of peers. He knows that the fraternity has certain attitudes toward women, certain attitudes about what manliness consists in, and so on. He develops an internally consistent position to take in discourse with the members of his fraternity. He creates an imaginative persona that is consistent with the position held by the other members. In normative discussion, he gets on with them well. On this basis, he is admitted into the fraternity and is in good standing with its members. In the course of time, however, he witnesses the other members of the fraternity engaging in behaviors to which he has a strong emotional response—behaviors like hazing, harassing women, and making sexist and racist remarks. He feels a strong sense of disgust. As a result, he never actually engages in any of these behaviors. He doesn't recognize this about himself, however, and in normative discussion; he continues to avow the

attitudes that are similar to the other members of his group. There are at least two directions that we could take here. One might maintain, with Gibbard, that the linguistically infused norms that Peter endorses *are* his normative judgments because they come about as a result of the right kind of process. The disgust that Peter feels comes about as a result of being in the grip of norms of politeness and social responsibility. The other response is to say that, though Peter avows one set of norms, his disapprobation and his subsequent behavior most accurately reflect his "true" normative judgment.

It is noteworthy that in Gibbard's discussion of imaginative rehearsal, he talks about developing an "*imaginative persona."* It is noteworthy because, when a person reflects on the things the set of norms they endorse, the persona they construct may *truly be* nothing more than a construction of their imagination, with no real connection to how they behave in real world circumstances. If this is so, then that is a problem for the NGT. Though we have worked out what we take to be best to think or do in a cool hour, that reflection has very little or nothing at all to do with what we think or do when the hour is not so cool.

Consider the character John Falstaff in Shakespeare's *Henry IV Part I* and *Henry IV Part II.* Famously, Falstaff is a good for nothing, cowardly, opportunistic, drunk. In addition to providing the plays with most of their comic relief, the character of Falstaff advances the plot by tempting prodigal Prince Hal away from his father and his responsibilities to the kingdom with the lure of a lifestyle of debauchery.

Falstaff's many character flaws are obvious, but that doesn't keep him from talking a big game. When he avows norms, they are always virtuous. He swears "a plague on all cowards," though he, himself, runs from every battle. He claims that liars are "villains and the sons of darkness," though he only tells the truth himself when other, more reliable witnesses are present to call him out on his falsehoods. Nonetheless, the set of endorsements that Falstaff is willing to avow *is* consistent. Not only are his avowals consistent, but they also seem sincere. Viewers can't help but to like Falstaff, not just because he is funny, but also because there really seems to be *some* goodness there, which is why it tugs at our heartstrings when Prince Hal renounces him.

One of the reasons that Falstaff's character is so comical is that his imaginative persona, consistent as it may be, is so *inconsistent* with his actual behavior. So, contrary to the wisdom offered by the NGT, Falstaff's reflections on norms don't tend to guide his behavior at all. This is, then either a problem with IRCT or a problem with the NGT. That is, it is either a problem with the view that normative judgments require attention to demands for consistency, or it is a problem for the view that when we form normative judgments, we are motivated, at least to some degree to act on them. We can imagine a scenario in which Peter the frat boy is not motivated in any way to engage in hazing or like behavior. Falstaff is certainly in no way motivated to be brave.

If the problem is the consistency requirement, the concern is similar to a one that is frequently raised against coherentism as a theory of justification in epistemology. It is possible for a person to have a set of beliefs that all cohere with

one another, but are all, in fact, false. The coherence theory of justification is problematic, in part, because mere coherence of beliefs doesn't necessarily provide good reason for thinking the beliefs under consideration are true. Similarly, the fact that there is a consistent set of endorsements present does nothing to establish that the set is a set of the agent's normative judgments or that the agent will be in any way motivated to act on that consistent set. There may well be many other factors that figure into the set of endorsements one is willing to avow, and self-blindness limits the norms an agent may be capable of avowing. The fact that a set of endorsements is consistent doesn't provide us with any reason to think that, among all the mental states we could select as the mental states that normative judgments express, the state must be a state of endorsement.

Gibbard acknowledges that our total plan, our total system of norms, is almost always incomplete. There is some information that we don't have. There are some instances in which we will be ignorant of information that could help us settle certain normative questions. Some incomplete or inconsistent systems aren't inconsistent or incomplete because facts are missing. We frequently, perhaps even always, conduct our lives while maintaining endorsements that are inconsistent with one another. It would be some sort of miracle if we could pick out all the inconsistencies given the number of norms we would have to endorse just to get through our daily lives. We don't have that kind of time to engage in imaginative rehearsal. Consider the case of Jane. Jane endorses norms that call for being kind to others. She thinks that friendliness can make a huge difference, even in the most insignificant circumstances. She therefore endorses the following:

Greet: If I see someone that I know on the street, I ought to greet them.

Now, consider two different cases in which she might follow *Greet.* In the first case, she sees someone at the supermarket with whom she is acquainted. She does not know this person well but, in keeping with her general commitment to *Greet*, she says hello. This is an easy case for Gibbard's view to analyze. Jane acts in a way that is consistent with her endorsements, and as a result, she acts on her normative judgments. There is no weakness of will present.

The second case is not as obvious. Suppose that Jane accepts *Greet*, but she also accepts the following:

> *Toxic:* I ought not to interact with people who tend to be a drain on my emotional well-being.

Often, acting on both *Toxic* and *Greet* will be possible at the same time. After all, the vast majority of people that she knows do not have a substantial effect on her overall happiness. Suppose, though, that Jane sees one such toxic person in the supermarket. She passes her in the aisle and, despite Jane's distaste for the woman and her desire to refrain from interacting with toxic people, she nevertheless, says hello.

What do we want to say about this case? Do we want to say that, as in the Milgrim case, Jane has succumbed to weakness of will? Do we want to say that Jane was in the grip of a norm? This all depends on the norm that is "really" endorsed. If the norm I endorse is *Greet,* then we aren't dealing with a case of weakness of will. If the norm I *actually* endorse is *Toxic*, then it is a case of weakness of will. I am in the grip of a norm rather than acting in accordance with norms that I endorse. There is really no non-arbitrary way of determining which of the two states Jane is

in, if the two distinct states really even exist at all. In fact, it isn't clear that the distinction is non-arbitrary in the Milgram case either. A person may take himself or herself to endorse norms that call for non-violence, but what reason is there to believe that the avowals that they are willing to make in a cool hour are better indicators of their genuine moral judgments than their behavior in an actual situation? Pointing to norms of non-violence endorsed in a cool hour when it comes to Milgram-type experiments as true normative judgments may be more palatable, but it isn't necessarily more accurate.

Perhaps our endorsements really aren't that stable or consistent with one another. Again, the judgments that actually determine what we will do in a particular case may have very little to do with what we have reflectively endorsed. We often use non-reflective processes when adjudicating disputes between two inconsistent competing endorsements. Those non-reflective processes, intuitively, may themselves be forms of normative judgments, and, again, it seems ad hoc to rule them out.

One response might be that the norm that Jane has really endorsed is a norm that specifies a preference order of endorsements. For example, she might actually endorse a norm that says, "*Greet* always outweighs *Toxic*. Follow *Toxic* when it comes to avoiding voluntary interactions with toxic people, but, if forced into the situation, follow *Greet.*" It is, of course, possible that Jane really has endorsed some norm like that, but it is equally, if not more likely, that she has never spent any time imaginatively rehearsing what she would do in a situation like the one she finds herself in in the supermarket. It is implausible that we have, not only first order

endorsements, but also second order endorsements about the rank ordering of our endorsements.

Conclusion

Gibbard speculates that the mental state of endorsement exists and that it is the state that normative judgments express. In this chapter, I have argued that endorsement is not a plausible candidate for the state that normative judgments express. It fails to capture the range of our normative experiences.

CHAPTER VI

IS REFLECTIVE ENDORSEMENT GOOD FOR ITS OWN SAKE?

We have now seen reflective endorsement used to solve many philosophical problems. Korsgaard uses it to ground normativity. Frankfurt uses it as part of his accounts of freedom, personhood, care, and love. Friedman defines autonomous action as action that has been reflectively endorsed, and Gibbard uses it to account for what human beings are doing when they make normative judgments.

Each of these thinkers suggests that reflective endorsement is *procedurally* valuable. On one interpretation of what this means, the *procedure itself* is valuable, regardless of what it brings about. This seems to suggest that reflective endorsement is intrinsically valuable. In this chapter, I will argue that, if reflective endorsement is valuable at all, it must be instrumentally, rather than intrinsically valuable, In the chapters that follow, I will argue that reflective endorsement is, indeed, instrumentally valuable, but only under certain conditions.

Is Reflective Endorsement Good for It's Own Sake?

The question that I will be addressing here is whether, absent any other good-making features, reflective endorsement is, itself, a good. I will argue that it is not and that, in fact, in many cases it actually contributes to a negative state of affairs.

As we've surveyed the literature on the various uses of reflective endorsement, a common thread has been that it provides a kind of normatively superior motivation. For example, for Friedman, a Muslim woman behaves

autonomously when she affirms the principle that wearing her hijab is consistent with deeply held values that she has endorsed. Emboldened by her reflection, she puts the hijab on once more when she wakes in the morning. Her running sense of self, established by her continued reflective endorsement of the values that matter to her, motivates her to behave in similar ways in the future. For Friedman, this kind of behavior is normatively superior to behavior that is not motivated by reflective endorsement or that is less motivated by endorsement.

Something similar is going on in the case of Korsgaard. When we reflect on what really matters to us, we endorse certain practical identities, and those practical identities provide motivations that are genuinely normative to engage in behaviors that are mandated by those identities.

We have also seen, especially in our discussion of Gibbard, that reflective endorsement is just one source of motivation. Its proponents seem to suggest that endorsement is normatively superior to other forms of motivation. For Friedman, it is because we are on the far end of the autonomy spectrum when we engage in endorsement. For Korsgaard, it is because we are acting on reasons that are genuinely normative for us when we do so. For Frankfurt, it is because we are expressing our wills—exhibiting our personhood, when we take an evaluative stance toward our own inner states.

I'll attempt to demonstrate that motivation that comes from endorsement is not properly understood as superior and should not, in many cases, be granted any sort of normative authority. To make this case, I'll first outline just a few sources of motivation that are not instances of endorsement. I'll consider cases that are

motivated in the ways I have described and compare them to cases in which behavior is motivated by endorsement. I'll consider behavior of each of the following types:

1. Instinctual Behavior
2. Behavior that results from habituation
3. Behavior that results from shaming

First, I'll provide a sense of what I have in mind in each case. First, let's consider instinctual behavior. This is the kind of behavior that we are, essentially, hardwired to engage in. For example, if I notice that my child has tripped and is beginning to fall, I will, without thinking, reach to catch him.

The second type is behavior that has been habituated. These types of behaviors are often those in which we have been socialized to engage. Gibbard calls motivation of this type motivation on norms that have been *internalized.* These behaviors might include things like maintaining a certain social distance from others, holding the door open for others as they enter public places, and so on.

Habituated behaviors can also be more complex. For example, imagine a mother, Sarah, who has internalized appropriate parenting behaviors based on the way that she was raised. When she was a child, her mother had a swear jar that everyone contributed to when they uttered a profanity. Sarah has one as well. Sarah's mother took her children to church every Sunday. Sarah does the same thing with her children. She has not reflected on the behavior, she has simply assumed, without challenge, that the way that her mother raised her was the way that she should raise her children.

The third kind is shame-based behavior. Again, consider a woman who is a mother. We'll call her Martha. Martha uses only cloth diapers for her baby. She buys only organic food. She won't let her family go near anything that is genetically modified. Though these behaviors seem to suggest some consistency with a certain kind of world-view, in Sarah's case, that is entirely a coincidence. Truthfully, she doesn't care much about the environment, she has no attitudes about the health benefits or lack thereof when it comes to eating organic food, and, frankly, she doesn't even know what it means for something to be "genetically modified." Her motivation for these kinds of practices come entirely from the fact that her friends engage in the same behaviors and she doesn't want them to think that she is an inferior parent.

On the face of it, it may seem as if these sources of motivation are somehow inferior to behaviors that are motivated by reflective endorsement. In what follows, I will make a case against that intuition. I will argue that the good-making and bad-making features of a case of human action have nothing to do with endorsements, but, rather, are determined by states of affairs. To establish this, I provide a case where, initially, the other types of motivation that I listed above (instinctual, habituated, or shame based behavior) are in play. I will then introduce endorsement into the situation and check to see if, intuitively, endorsement has added anything of value to the situation.

Let's begin with a case of instinctual behavior. Betty and John are on a subway platform. John trips and begins to fall into the gap. Instinctually, Betty grabs him by the arm and pulls him up, preventing him from falling. John's life is

saved. That's a good thing. Betty's actions, though not motivated by any reflection at all, let alone reflection *and* endorsement, have produced a good state of affairs.

Let's modify the case slightly. Betty and John are, again, on the platform. Again, John begins to fall, and, again, Betty catches him. But, before she catches him, she reflects briefly, finds that she endorses norms that dictate that she should help people when she can, and then she proceeds to catch him. What do we want to say in this case? One might be inclined to say that there is no normative difference between the case that involved endorsement and the case that merely involved instinct. After all, the action was the same in both cases—Betty saved John. On the other hand, one might be inclined to say that the case in which reflective endorsement is involved is superior because Betty's decision in that case is somehow reflected in her very character. That intuition may be justified but, if it is, reflective endorsement might not be doing all of the work in generating that intuition.

Let's consider yet another case of a similar type in which John and Betty, yet again, find themselves on the platform. Again, John loses his balance, and, again, begins to fall. Betty takes a split second to reflect on her values, realizes that she hates John and that she endorses norms that would best be promoted by his death. She lets him fall to his doom. This seems like a bad state of affairs. What does the endorsement contribute to this state of affairs? Had Betty not reflected on allowing John to fall, her actions still might have produced a bad state of affairs. But it is not clear that her actions would have been the same had she not stopped to reflect on her values. Her instincts may well have kicked in, and she might have done

something to help him. In this case, the presence of reflective endorsement seems to have made matters worse. It's not obvious that there is value in the mere fact that Betty went through the reflective process.

Sometimes, habituated behaviors may also be more valuable than behaviors motivated by norms that an individual endorses. Let's return to the case of Sarah, who has been habituated to engage in all of the same parenting behaviors as her mother. She's never really reflected on her parenting decisions, she's simply assumed that all of the decisions that her mother made in raising her were the right decisions to make. This is not uncommon. Let's imagine that one of the things that Sarah routinely does is see to it that her child eats green vegetables as part of her dinner each night. She has never reflected on the advisability of this course of action, she is merely habituated to engage in it. She does it without fail, and, as a result, her child routinely gets the nutrition that she needs.

Now imagine that Sarah, realizing that she has never really given much thought to the practice of feeding her child vegetables, thinks it through. She consults a number of different websites. She concludes that she has been doing the right thing all along, and she continues to feed her child vegetables. She has now endorsed norms that support feeding vegetables to her child. Again, one might have two different responses to this state of affairs. One might think that Sarah's investigation into the nutritional value of vegetables has added nothing of value to the situation. She was feeding her child healthy food before and there was no reason to believe that behavior was going to stop anytime soon. After consulting sources on the issue, she simply reaffirms the action in which that she was already

inclined to engage. On the other hand, one might think that Sarah is better off in the case in which she has considered carefully her reasons for feeding vegetables to her child and has endorsed norms that support the practice. Again, however, we can question whether it is the endorsement *itself* that adds value to the state of affairs, or, instead, it is the *substance of what is endorsed* that lends value to that state of affairs.

Let's consider another case. Again, Sarah realizes that she has never really reflected on her reasons for feeding her child vegetables. Once more, she takes some time to reflect on it. She consults various websites, but, sadly critical thinking skills are not her strong suit and she checks all the wrong sites. Perhaps she assigns too much credence to websites that suggest that toxic chemicals have leaked into all the earth's soil, rendering all vegetables unsafe for consumption. She concludes that her mother had been mistaken about the advisability of feeding children vegetables and that now, she, Sarah, had been doing the wrong thing too. She stops feeding her child vegetables.

Feeding one's children vegetables is, of course a good thing to do regularly. The reflective endorsement helped when it led to a good state of affairs, but not when it led to a bad state of affairs. But, it seems that it was the goodness of the state of affairs itself, and not the endorsement that made the action good. It doesn't seem like the habituation motive is inferior to the endorsed motive, and again, the endorsed motive can actually be worse.

So far, we have been considering cases in which reflective endorsement makes a state of affairs worse. However, there are cases when reflective

endorsement seems to be ideally suited to the task of changing bad behavior. Imagine that Sarah spanks her children. When she was growing up, her parents spanked her. She never viewed her parents as abusive, but she never really reflected on the matter, she simply assumed that it was the right thing to do. A friend of Sarah's mentions to her a study pertaining to spanking that concluded that spanking was actually harmful to the proper development of children. This new information brings about a change in Sarah's behavior. Instead of simply continuing to believe that what her mother did was best, she looks into the issue and learns that repeated studies confirm that spanking children is harmful and that other forms of punishment are actually more effective. She now endorses norms that speak in favor of refraining from spanking one's child. She no longer spanks her children. This seems like a good thing. Again, we must ask ourselves, is it the endorsement involved in this situation that makes the state of affairs good, or is it the brute fact that Sarah's children are no longer being spanked? Again, one might think that the act of endorsement is a good thing. One might assert, as mentioned earlier, that engaging in these kinds of practices routinely as part of one's practical reasoning helps one to endorse the right norms often. One might also argue that reflective endorsement of this type helps to develop Sarah's moral character. This is a consideration that we will explore in greater detail in the final chapter of this book.

On the other hand, in many cases, our intuitions might be that it is best to ignore one's endorsements. Consider the case of a deeply religious Christian man. The man is a fundamentalist, and he believes that homosexuality is a sin. When he

reflects, he endorses the norms that a fundamentalist Christian would endorse. His daughter comes out to him as a lesbian, and informs him that she has a partner whom she loves and is planning to marry. Despite the set of Christian norms that, upon reflection, the man endorses, he treats his daughter with love and support. If he had to describe his own behavior, he would suggest that he is experiencing weakness of will. It seems, however that what it is that he endorses, if it is playing any role in this situation, is actually making the situation worse. It would be better if he didn't experience the cognitive dissonance that his religious beliefs are pretty clearly causing.

What all of this seems to suggest is that reflective endorsement itself is not intrinsically valuable. These cases also suggest that reflective endorsement is not entirely without value. In the nest two chapters, I will explore possibilities for the type of value that reflective endorsement might have.

Authenticity and Authority

Above, I provided some cases that were meant to generate the intuition that reflective endorsement on its own is not intrinsically valuable. To show this, I offered cases in which motivation coming from "the animal control system" generates results, that are, intuitively, equally valuable or even more valuable than those generated by the process of reflective endorsement. There are, however, a different set of arguments for the claim that the process of reflective endorsement has intrinsic value. These arguments concern the values of authenticity and authority, respectively.

Some people think that authenticity has intrinsic value. It is important to be genuinely, who one is. Authenticity is a trait that is highly valued by existentialist thinkers, among others. Some people who value authenticity might think that what it is to be an authentic person is to engage in the process of reflective endorsement. One might think that the reflective endorsement process constantly checks some values that you endorse against other values that you endorse, scanning for consistency and ensuring that you remain true to yourself.

Similarly, some might think that commands issued from one's true self have authority, and that authority makes those command normative for us. This is similar to the view that Korsgaard holds, and also to the view that Kant holds. One might think that the way that one identifies or constructs one's true self is through the process of reflective endorsement. Therefore, because of the identity relationship that obtains between normative authority and reflective endorsement, if the authority one has over oneself has intrinsic value, so, too, does reflective endorsement.

Both of these arguments rely on identification of states that are accessible through endorsement as states that represent a person's "true self." If a person's deepest hopes, wants, desires, fears, and reasons for actions are not transparent to her, if they are not accessible to her through the process of endorsement, then it is not clear that the states that are being accessed really are representative of the individual's "true self." At this point, I've provided a number of arguments for the conclusion that we don't get authenticity or normative authority if we don't have access to knowledge about who we really are. In fact, there are substantive

impediments to mere *attempts* to determine who we really are. I will turn to a discussion of those impediments, and a discussion what I take to be the real value of reflective endorsement, in the next two chapters.

CHAPTER VII

HUERISTICS AND THE PSYCHOLOGICAL VALUE OF REFLECTIVE ENDORSEMENT

In this chapter, I will present one of the ways in which reflective endorsement can be valuable. Ultimately, I will contend that the endorsement process is valuable in more than one way, and that the types of value involved are substantially different from one another. In fact, I will argue that the value that reflective endorsement has in one domain may diminish the type of value that it has in another domain.

The clash of values in play here tracks the debate concerning what the central role of philosophy should be. The history of philosophical thought is rife with philosophers who have argued that the proper goal of philosophical inquiry is to arrive at truths about the world. Plato and his philosophical descendants attempted, to the extent possible, to find the way out of Plato's cave—to come to know the true nature of things. In his *Meditations*,[90] Descartes cast all of his beliefs into doubt, retaining only those that he could know with certainty. The goal of his philosophical program was, in part, to become better able to distinguish truth from falsity, and, in so doing, provide a firm foundation for the sciences—contributing to the general project of philosophy conceived as an endeavor aimed at arriving at truths about the world.

Others think about the aims of philosophical inquiry differently. Instead of engaging in a search for truths, these philosophers ask, "How ought I to live?" Of

[90] Descartes, R. and Cress, D. (1993). *Meditations on first philosophy*. Indianapolis: Hackett Pub. Co.

course, these two approaches need not be entirely independent from one another. After all, the "truth seekers" described above could answer, "I should live my life in pursuit of truth." The views are distinct from one another, however, in the sense that the answer to the question "how ought I to live?" needn't have the pursuit of truth as any part of its answer. There are different interpretations of the question. Some, like Aristotle, think that it can only be understood functionally, as a question about what it is for human beings to excel.[91] Others see the question as pertaining to how to live ethically or virtuously, independent of any understanding of human function. Still others see the question as a religious one. Finally, and most crucially for our purposes in the next section, some thinkers understand the question as one of how to exist in this world as a human being without succumbing to the seemingly inevitable existential pitfalls that accompany a condition such as ours. Is it possible to live a happy, or at the very least, a tolerable human life? Should we take seriously Camus' claim that "there is only one truly philosophical problem, and that is suicide?" If a life with limited existential anguish is possible, how can it be achieved?

My analysis of the value of reflective endorsement will have bearing on both of these conceptions of living a philosophical life. As I've noted, these conceptions can sometimes be divergent. I will argue that the diagnosis of many of the problems that I have identified for the process of reflective endorsement as a method for solving philosophical problems is that, at times, the reflective endorsement process has value in one domain or in one respect, while contributing to *disvalue* in another

[91] Aristotle, W. D. Ross, and Lesley Brown. 2009. *The Nicomachean ethics*. Oxford: Oxford University Press.

domain or respect. I will address one of these sources of value in this chapter, and the other type of value in the following chapter.

Reflective Endorsement and Psychological Health

When we reflect on our own internal states and avow or disavow things that we find upon introspection, it is often psychologically satisfying. It may seem as if the satisfaction comes about as a result of the fact that our introspection has led us to become more authentic, more internally consistent, and more committed to the things that we, as persons, take to be *truly* valuable.

The feeling that, through introspection, we have become a more unified and principled human being is, undoubtedly, a nice one, even if it doesn't track the truth. The *appearance* of coherence, consistency, and, ultimately, authenticity, in many cases turns out to be just as satisfying, if not more satisfying, than *actual* coherence, consistency, or authenticity. In this section, I will employ studies from social psychology to establish how reflective endorsement might be valuable toward important psychological ends. The value that it has toward these ends, as we will see, may well be in virtue of that fact that taking the reflective stance puts us in an ideal position to make use of heuristics that are really good at producing favorable psychological states. What I have in mind by "favorable psychological states" are pleasant sensations or, at least, the lack of a minimal level of unpleasant sensations, such as fear, pain, or anxiety.

Studies in social psychology suggest that human beings are guilty of various kinds of attribution errors. We misjudge the impact of situational, contextual factors

on the behavior of others, particularly in situations in which we perceive those others to be behaving badly. In these cases (and a broader range of others), we are quick to blame the behavior of others on enduring personality characteristics, rather than on the unique features of the situation. In their book *The Person and the Situation,* Nisbett and Ross provide several examples of cases of this type.[92]

In a 1967 study of college students, participants were asked to listen to or read speeches that they were told were prepared by other students.[93] They were informed further that the students who prepared the speeches or essays were under substantial constraints. The essays took positions on social issues, and the students were assigned a side of the issue to argue for. So, the participants in the study knew, before being asked to provide an evaluation that the students could not argue for any other side of the social issue than they in fact did. They were constrained by the situation. Even so, participants in the study overwhelmingly assumed that the students who constructed the speeches or essays were sympathetic or even fully committed to the side of the issue for which they were arguing. That is, students were more likely to make attributions of consistent personality characteristics or dispositions to the speakers than they were to explain the behavior primarily in terms of the situation in which the speakers or writers found themselves.

---

[92] Nisbett and Ross, *The Person and the Situation* Pinter & Martin Ltd; 2 edition 2011

[93] Jones, E.E., and Harris, V.A., The attribution of attitudes. *Journal of Experimental Social Psychology, 3,* 1-2.

A different study in the 1970s[94] led researchers to a similar conclusion. Females were asked to provide entertainment for board members of "The Human Development Institute" and a group of their financial backers. Some of these women were offered $.50 per hour to volunteer, and others were offered $1.50 per hour. Observers were then asked to predict the future behavior of the women who volunteered. They were asked "How likely is it that the subject would also volunteer to canvas for the United Fund?" The assessment of the observers was that the women who volunteered to provide the entertainment would be more likely to Canvas for the United Fund than non-volunteers, regardless of who much they were being compensated for providing the entertainment. In other words, observers were likely to attribute the behavior of the volunteers in this case to concrete and enduring features of their behaviors, rather than to the particulars of the circumstances under which the volunteers were asked to participate.

Studies also show that we are far more likely to do the opposite when it comes to our own behavior.[95] We are more likely, at least when it comes to negative behavior, to attribute the things that we do to the circumstances that we are in rather than concrete, stable, personality characteristics. As we will see, however, the same is not true when it comes to what we might deem to be our own *favorable* personality characteristics.

---

[94] Nisbett, R.E. Caputo, C., Legant, P., and Maracek, J. (1973) Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology, 27, 154-164.*

[95] See Moskowitz, 2005.

The Need for Control

In his book, *Social Cognition*, Gordon Moskowitz argues that human beings long for a sense of control over what happens in their lives, and this desire helps to explain some of the attributions that they are apt to make.[96] In particular, assigning dispositional attributions to others helps to satisfy this desire. The world strikes us as a safer, more predictable place when we think that we can rely on the behaviors of others to stay the same, as it might if enduring personality traits or characteristics caused the behavior.

Often, we make character attributions in the case of our own behavior as well. Moskowitz argues that doing so contributes to a sense of control over our own lives. However, that the sense of control that we take ourselves to have as a result of such attributions is often illusory. He supports his point with a 1975 study by Langer.[97] In the study, participants were given a can that contained two marbles. They were told that, if they picked one marble in particular, they would be given a desirable prize. Some were told in advance which marble would win the prize and others were not. Some participants were handed a marble and others had to reach into the can and select one. In any event, the process by which the participants obtained a marble, and which marble they obtained, was entirely random. Participants who randomly selected the prize-winning marble from the can were more likely to see themselves as responsible for selecting the winning marble than

[96] Moskowitz, Gordon B., *Social Cognition,* The Guilford Press New York 2005.
[97] Langer, E.J., (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311-328.

were participants from any other group, even though they randomly selected that marble.

The idea that we have control over the positive events that take place in our lives is psychologically helpful. Our brains tend to lead us to favorable conclusions about the degree to which we are responsible for the good things that happen to us. We take the good things, and we attribute them to enduring traits of our own characters. Of course, evidence suggests that we are making a mistake when we do this. Nonetheless, it appears to be psychologically healthy.

Here, it seems that reflective endorsement has value of a certain sort. If we are inclined, as empirical evidence seems to suggest that we are, to affirm conceptions of ourselves according to which we take ourselves to have control, and if the perception of control is valuable (even in the absence of the existence of *actual control),* then, prima facie, reflective endorsement is instrumentally valuable toward that psychological end. This type of value is not insignificant. A sense of control, even if it is illusory, can easily prevent a slip into existential despair.

It clearly isn't this type of value that thinkers we have considered in this book have in mind, at least, not primarily. In fact, if it is really a need for a sense of control that motivates our self-attributions, that fact might undermine some of the conclusions drawn by reflective endorsement theorists. Consider, for example, Korsgaard's view of the normative authority of practical identities. On her account, reasons with real normative force are generated by the conceptions of ourselves that we value. It is an open possibility, of course, that those conceptions of ourselves that we value, that we would rather die than give up, are not actually

persistent character attributes at all. Instead, our brains develop a psychologically healthy heuristic that allows us to perceive the world in such a way that it makes psychological sense for us to understand ourselves as having control over what happens to us. If the practical identity is illusory, does it really generate reasons that have normative force?

The important thing to note here is that a person might, quite involuntarily, construct a picture of themselves, that doesn't exactly match the facts, but affords them the optimal level of control. Having a psychological sense of control may require that the person in question, in certain cases, increases her reflective attention on positive features of her character, while decreasing reflective attention on the more negative features of her character.

Consider Stephanie. Stephanie conceives of herself as a health nut. She runs several miles a day and never skips weight training. She eats a balanced diet, never scrimping on fruits, vegetables, or protein. This is the conception of herself that she endorses upon reflection. Consider further, however, that there is a description of Stephanie that is true, but that she will not accept. She has pushed the truth of this description to some subconscious place where she never has to look at it. Stephanie is an alcoholic. After engaging in healthy behavior all day, she relieves stress at night with excessive amounts of alcohol to the point that her physical health, in particular, the health of her liver and her kidneys, are in serious jeopardy. The psychological mechanism that kicks in to help Stephanie feel in control of her life and her health, despite her clear deficiency is aided by her act of reflective endorsement. She endorses a conception of herself according to which she is a

health nut, while ignoring, or perhaps even rejecting the conception of herself as an alcoholic—a conception of herself that might leave her feeling powerless.

The value of reflective endorsement in this case is that, absent a sense of control, Stephanie may be plunged into a deep despair, or, at the very least, she might feel tremendously anxious. However, this example also highlights a way in which reflective endorsement can also be harmful. When the heuristic kicks in that allows Stephanie to feel in control of her own health despite her alcoholism, it overrides the concern that, objectively, she should be feeling in order to change her habits and solve her drinking problem.

The Self-Serving Bias

Reflective Endorsement has instrumental value in a different psychological way as well. Empirical science supports the conclusion that we are motivated to both attain and to maintain a positive sense of self.[98] The process of reflective endorsement can help us to develop a positive conception of ourselves that is psychologically healthy.

It is a truism that most people are average. After all, that's what it is to be average. Despite the obvious truth of this description, however, most people *believe of themselves* that they are above average. In a paper on the topic, Alicke and Govorun highlight the results of a well-documented case of the phenomenon:

> Data collected in conjunction with the 1976 College Board Exams provide on of the earliest, most striking, and most frequently cited demonstrations of the better-than-average effect. Of the approximately one million students who took the SAT that year, 70% placed themselves above the median in

[98] See also Cross, P. (1977), Svenson (1981),

> leadership ability, 60% above the median in athletic ability, and 85% rated themselves above the median in their ability to get along well with others.[99]

Of course, it is obvious that many of the participants were wrong about how they compared to the average with respect to each of these characteristics. What this study, and many others like it highlight is that we have a psychological need to think highly of our good traits—in many cases, more highly than appraisal of our traits honestly deserves. It seems that we do this, in part, because thinking highly of ourselves is an important part of psychological health.

Importantly for our purposes here, people also have a tendency to view themselves as better moral agents than their peers. Codol asked study participants to assess how often they conformed to socially desirable norms.[100] Most participants indicated that they conformed to such norms more often than average.

As a result of the better-than-average phenomenon, or our self-serving bias, we are inclined to view ourselves in more favorable ways than might actually be warranted by the best available evidence. Viewing ourselves in the best possible light, or, at least, a better-than average light, contributes to psychological health. Reflective Endorsement can, and probably often does, contribute to this conception of ourselves. We introspect, we consider ourselves and our traits, and we employ a heuristic that emphasizes the positive.

---

[99] Alicke, Mark D., Govorun, Oleysa, "The Better Than Average Effect." In *The Self in Social Judgment.* Ed. Alicke, Mark D., Dunning, David A., Krueger, Joachim I. Pyschology Press New York 2005.

[100] Codol, J.P. (1975) On the so-called "superior conformity of the self" behavior: Twenty Experimental Investigations. *European Journal of Social Psychology, 5,* 457-501.

Consider the following case. Mark considers himself to be an honest person. He would report that he views honesty as a very important trait. When he introspects, he endorses a picture of himself as an honest person, excusing those cases in which he told little white lies, and even exempts some whoppers. He was, after all, justified in lying on those occasions, or so he tells himself. The act of rationalizing his past lies is, perhaps, a best-case scenario. Research done by Alicke suggests that when making character attributions, people do not survey their past behaviors.[101] Instead, Alicke et al. concludes that people employ a better-than-average heuristic. This heuristic has an "automatic tendency to assimilate positively-evaluated social objects toward ideal trait conceptions, and does not assume that people routinely review their behaviors to make self-other judgments." The idea that past behaviors are frequently not factored into self-assessment is strongly suggested by an Alicke study.[102] In phase one of the study, participants were asked to reflect on their past behaviors. They were asked, for each trait dimension, what percentage of the time they had acted in a way that was consistent with that trait. For example, a person might say that they were helpful to others, 80% of the time. At a later stage in the study, the same participants were given numbers that they were told represented the average percentages of how often, on average, their peers reported engaging in behavior along those same trait dimensions. Unbeknownst to them, the students were provided with the numbers

[101] Alicke, Mark D., Klotz, M.L., Breitenbecher, D.L, Yurak, T.J., and Vrendenburg, D.S. (1995) Personal contact, individuation, and the better-than average effect. *Journal of Personality and Social Psychology, 68, 804-825.*

[102] Alicke, Mark D., Vrendenberg D.S., Hiatt, M., And Govorun, O. (2001) The "better than myself" effect. *Motivation and Emotion, 25,* 7-22.

that they, themselves, provided at the earlier stage in the study. At this second stage, they were asked to compare their traits to the numbers that they falsely believed were provided by their peers. The results were consistent—students on average said that they fared better for those traits than the average college student (even though these were the numbers that they, themselves provided as an accurate summary of their past behaviors!) Alicke calls this the "better-than-myself paradigm."

Moreover, we can imagine that Mark's perception of himself as an honest person is never seriously challenged, for when he compares his honesty against others with whom he is either acquainted or just generally familiar, he, perhaps even unknowingly, compares himself with the least honest people that he knows.[103]

As I have pointed out, however, the practice involved here is not truth-conducive. Many philosophers seem to want to maintain that the value that reflective endorsement has comes from the connection it has to something like authenticity. Friedman, for example, wants to identify autonomous actions with actions that are reflectively endorsed, because such endorsements represent choices made by an agent's true, authentic self. The true, authentic self here is just defined in terms of the kinds of things that they would, consistently avow or disavow. The conclusion that reflective endorsement ensures authenticity or even makes authenticity more likely than not appears to be undermined by heuristics like the self-serving bias. If we are inclined to endorse pictures of ourselves that are flattering, pictures of ourselves according to which we are better than average, even

[103] This type of explanation of self-serving bias, or better-than average effect, is suggested by work done by Perlfloff and Fetzer (1986).

when we are not, then it looks like we are not truly motivated by a desire for authenticity.

Ambiguity heightens our tendency toward this kind of bias. People might have very different conceptions of what it means to be "trustworthy," "kind," or "respectful." When we reflect, we might create trait abstractions, and the self-serving bias is ready and able to help us construct those abstractions in ways that are favorable, rather than unfavorable, to us. It may be, then, that we are more quick to think about those self-biased abstractions as being representative of our selves rather than our actual, concrete, past behaviors.

Self-Verification

Reflective Endorsement can also be psychologically healthy in a third way—to satisfy a need we have to verify those components of the self that we already take ourselves to have. Some social scientists have suggested that, in addition to satisfying a need for predictability and control, our impulse to self-verify also satisfies a psychological need that we have to have "consistent and balanced cognitions.[104]" Too much change is unsettling to us. Therefore, in addition to attempting to maintain a positive view of ourselves, we also try to maintain a stable one.

In support of the claim that human beings have a tendency toward self-verification that is distinct from their tendency toward self-serving affirmations,

---

[104] Moskowitz, Gordon B., *Social Cognition: Understanding the Self and Others*. The Guilford Press New York 2005.

studies have been conducted that indicate that human beings prefer to be judged negatively with respect to those features of their life or behavior that they already view as being deficient. In one such study,[105] participants were asked to describe their worst feature (e.g., their weight). Other participants were then brought in to comment on that feature. Some of these participants said something positive about the feature and others said something negative about the feature. The original participants were then asked which of the two assessors they would like to participate with in a later stage of the experiment, and people chose the assessor who agreed with their assessment of their negative feature rather than the one who disagreed with it. It may be, then, that, to achieve a sense of consistency, we accept evidence from others that supports the way in which we already want to view ourselves, and that we reject disconfirming evidence.

Of course, the process of reflection on one's own attributes and then disavowing one or more is a paradigm case of engaging in a reflective endorsement process (or, in this case, an instance of reflective disavowal). These studies suggest that we like confirmation of the ways in which we are already inclined to view ourselves.

Three Conclusions

There are three conclusions that I would to draw here. The first has to do with the value of endorsement and the second has to do with why I think value of this type is

[105] Swan, W.B., Jr. (1990) To be adored or to be known?: The interplay of self-enhancement and self-verification. In E.T. Higgens and R. Sorrentino (Eds.) *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 2 pp. 527-561). New York: Guilford Press.

neither what the philosophers we have discussed in this book have in mind, nor does this type of value benefit them in any way. The third specifically addresses Gibbard's use of reflective endorsement.

The first conclusion is, as we have seen in this chapter, it looks like the process of reflective endorsement can help us to satisfy at least three important psychological goals. First, it satisfies a need that we have for psychological stability and security. It would, after all, be quite frightening to feel as if you don't know what to expect, even from *your own behavior.* When we reflect on our own inner states, our traits, and the things that matter to us, we take ourselves to arrive at a reasonable enduring, stable conclusions. Sure, this sense of stability might be arrived at, in part, by a set of heuristics that allow us to confirm what we already believe. We might be more inclined to accept evidence that confirms the existence of the set of attributes that we already take ourselves to have. It is likely that those traits developed because it was fitness enhancing for them to do so. Human minds can be consumed with anxiety and general existential dread. Mechanisms that confirm a certain intrapersonal stability can help with that.

The process of reflective endorsement also psychologically healthy to the extent that it helps us to feel good about ourselves. We have seen that endorsement of positive traits is a common practice, guided by self-serving heuristics. So, this chapter has highlighted the ways in which reflective endorsement can be psychologically valuable.

The second conclusion that I want to draw is this: the type of psychological value of reflective endorsement that I have outlined is pretty clearly not what the

thinkers we have discussed have in mind when they claim that endorsement has value and can solve all sorts of philosophical problems. In fact, the heuristics that lead to psychological health in these cases appear to be at odds with the kinds of philosophical aims these positions are trying to achieve. Most of the views that we've considered are after the kind of authority or normative force that emanates from the "true self" in some sense. The existence of these heuristics that aim toward psychological health rather than truth or authenticity raise the specter of self-deception on a massive scale. This can't be the kind of value they're after.

The conclusion that I have drawn so far may not appear to apply directly to Gibbard's project. The third conclusion that I want to draw, then, applies specifically to Gibbard's project. Gibbard is doing something descriptive rather than normative. He isn't trying to tell us that we should or shouldn't treat the outputs of reflective endorsement as if they have authority over judgments of different types. Instead, he is simply providing a descriptive account of what it is that normative judgments express.

I think that operations of the heuristics I have described in this chapter are problematic for Gibbard's descriptive position as well. In my discussion of Gibbard, I argued that his moral psychology—specifically his taxonomy of human motivation, is too narrow, and that it rules out other types of motivation, that seem, intuitively, to count as normative judgments. The discussion of heuristics in this chapter highlights the idea that what goes on when we endorse something is not as simple as he makes it out to be. Some of the heuristics that we employ when we endorse things are really not much different from motivational impulses that take place at

the level of what he called "the animal control system." If these mechanisms are really not much different, that fact lends further credibility to the claim that there is no real, significant line to be drawn between an "animal control system" and a "normative control system."

In this chapter, I have raised some empirical studies pertaining to self-evaluation. I have argued that, reflective endorsement is valuable for psychological reasons, but that this isn't the form of value that philosophers who employ reflective endorsement in philosophical accounts are looking for. In the next chapter, I will provide an account of the conditions under which reflective endorsement might be valuable in a philosophical sense.

CHAPTER VIII

THE PHILOSOPHICAL VALUE OF REFLECTIVE ENDORSEMENT

In the previous chapter, I argued that reflective endorsement is instrumentally good in at least one respect—it contributes to psychological well being in several forms. It fosters a sense of control and predictability and it contributes to a positive self-image. I will now argue that, for a variety of reasons, it is implausible that the *only* value that reflective endorsement has is psychological value. After all, people who dedicate their lives to the study of philosophy do so because they think engaging in reflective activity is of fundamental importance. Many of us agree with Socrates that, "The unexamined life is not worth living." We tend to think that evaluation of the self is a crucial part of what it is to live an examined life.

That we take self-evaluation to be important is well illustrated by the quotidian ways in which we judge ourselves and other people. We tend to value the behavior of those who do not simply take life as it comes, but who, instead, stop to challenge or reaffirm the beliefs and values that they hold. When we know that one of our local hiking buddies, who finds great joy in the beauty of nature, supports the reduction of environmental regulations or the seizure and sale of state park lands to private entities, we want him to stop, reflect, and get his values in order. It is common for people to get frustrated when members of their communities claim to embrace one set of values, while voting for candidates who support an entirely different set of values. We want them to reflect, not just on the justification they have for their beliefs, but also on the consistency and strength of their values.

We value this kind of reflection. We tend to think it is bad when it does not occur. This fact simply cannot be explained by the kind of psychological value that I have appealed to above. We may want our hiking buddy to be in good psychological health, but that isn't the reason that we want him to reflect more carefully on his values.

It seems, then, that there is *some* value in the reflective endorsement process, above and beyond the mere psychological value that it has under the conditions described in the last chapter. When an abusive friend reflects on her abusive nature and resolves to change it, we don't think that the value of what she has done lies simply in the fact that she has introspected. Though the process of introspection might be psychological valuable *for her*, that psychological value isn't what we are appealing to when we praise her thoughtfulness. After all, she could have introspected *incorrectly.* She could have employed a self-serving heuristic like the one described in the previous chapter that motivated her to pay more attention to her positive traits than her negative ones, or she might have focused only on evidence that supported the picture that she already had of herself as a loving, caring friend. We've spent much time in this book discussing what it might look like for a person to introspect incorrectly.

One of the main difficulties for the approach taken by the thinkers addressed in this book, in my view, is that they maintain that an agent's introspection and affirmation of a value or set of values is enough to determine what is, in fact, actually valuable to that agent. What I have intended to suggest, throughout this book, is that the idea that values are constructed by reflective endorsement renders

individuals infallible on the issue of what it is that matters to them. I maintain that such a view is implausible for several reasons. First, we already have reason to believe that many of the heuristics that are employed when we introspect are aimed, not at achieving a more coherent, well developed picture of the self, but are, instead, aimed at maintaining control, stability, and overall psychological health. Second, the matter of what types of things a person truly cares about might be better addressed by looking at factors other than what a person would affirm upon introspection. This is so because the question of what an agent cares about is only internally accessible to a certain degree. Nevertheless, as we have seen, many thinkers on this topic tend to think of it as an entirely internal matter. At best, the views that we have considered provide accounts of what an agent *thinks* matters to them. The question of what *actually* matters to them might be one that needs to be resolved very differently.

Getting this right is important for many reasons. I'll briefly highlight one reason here, and will discuss it in more detail at the end of the chapter. Reflective Endorsement, if done right, is valuable because of the potential it provides for character refinement and moral improvement. I've argued throughout this book that reflective endorsement is not always valuable, and that, in certain cases, it actually has negative value. Despite the fact that the process can often go wrong, it can also go right. In fact, it may be the case that there are certain valuable states of affairs that can *only* be achieved through introspective avowal or disavowal. For example, if a person can't recognize negative character traits in themselves or dispositions to behave in ways that are harmful to either themselves or others, then

they may never be in the position to change those behaviors. Without critical self-evaluation, we may never be in a position to live flourishing human lives.

Reflective Endorsement and Socratic Virtue

I have suggested that, when we take reflective endorsement to have value, it is because we view it as an important part of a commitment to living a philosophical life. For the purposes of our discussion here, I will identify the project of living a life in which one's values are frequently assessed and revaluated in an attempt to live a more integrated, coherent, well reasoned, life, as a project with the aim of attaining *Socratic Virtue.* Going forward, I will use the term *Socratic Virtue* to pick out a phenomenon that has real philosophical value. Reflective Endorsement plays some role in the achievement of that goal, but only under highly specified conditions.

An Analog in Virtue Epistemology

The approach to the attainment of Socratic Virtue that I will argue for is motivated, in part, by Ernest Sosa's approach in his work *A Virtue Epistemology.*[106] I will argue that, though our interest here is in a different philosophical concept (Sosa is providing an account of reflective knowledge), many of the structural elements of his view will help us in our analysis as well.

Sosa thinks of belief attainment as a performance. He uses an example of an archer shooting at a target to describe how this is supposed to work. There are

[106] Epistemic questions, though relevant to much of what I am discussing here (especially insofar as knowledge is, itself, normative), are beyond the scope of this project. I discuss Sosa's view here to highlight some of its structural elements, which I will employ in my own account.

skilled archers and there are unskilled archers. When it comes to hitting targets, there are good shots and there are bad shots. Under certain conditions, skilled archers can take bad shots and unskilled archers can hit the bull's eye.

Sosa identifies three ways in which we can evaluate an archer's performance. First, we can determine whether it hits the target. The standards for evaluation here are straightforward. The arrow either hits the target or it does not. Second, we can determine whether the shot exhibits some skill on the part of the archer. This standard of evaluation admits of degrees. Some archers are better than others. Skilled archery, as with skill in many other performances, requires knowledge of various types. The archer must know, for example, how to shoot in less than favorable conditions (e.g., in the wind). They must know enough about their machinery to be able to operate it successfully in various conditions. Sosa calls evaluation of this sort evaluation of whether a shot is *adroit.*

Finally, Sosa considers whether an archer's shot is *apt*. An arrow can hit a target for any of a number of reasons, and not all of them are attributable to the archer. An arrow might be on course to miss the target entirely, but a sudden wind might pick up, blowing the arrow right where it needs to go. The arrow might be headed off course, but someone might move the target such that the arrow hits right where it should. An archer's shot is *apt* when it hits the target, not because of some external condition or set of conditions, but because of the skill of the archer.

Aptness, Reflective Endorsement, and Socratic Virtue

The use of Reflective Endorsement to achieve Socratic Virtue is also a skill, and is valuable in the Socratic sense *only* when it is being used skillfully. Some people are more skillful than others, and this is a respect in which I think accounts like those I have addressed in this book get things wrong. All of the accounts that we have considered (with, perhaps, the exception of Gibbard) seem to maintain that reflective endorsement is a process that we all engage in with roughly equal degrees of skill. So long as we have reflectively endorsed *something* we have reasons for action that are truly normative (Korsgaard), we counts as free persons (Frankfurt), and/or we are autonomous (Friedman). This is so because, for the most part,[107] on these views, the question of whether an agent has reflectively endorsed something or not is an on/off question, and there are no further standards to employ.

Extending the analogy with what Sosa has to say about belief production as a skill, I'll now lay out some of the conditions that need to be met in order for a person's assessment of what they value to be *adroit.* Recall that in the case of the archer, attainment of skill involved the attainment or possession of a certain body of knowledge. The archer needs to know how to operate his or her instrument. They also need to know background information about how their bodies work in conjunction with the bow. We will now consider what it would take for a person to be skillful with respect to the use of reflective endorsement.

1. Reflection on Past Behaviors.

[107] Friedman's account of autonomy does admit of degrees, but she seems to suggest that, though people might be *more* autonomous than others under certain conditions, merely engaging in the process of endorsement is enough to make a person minimally autonomous.

We've said that in order for an archer to be a good archer, he or she must have a certain sort of knowledge base. This knowledge base would include not just a substantial understanding of how bows and arrows work, but also an understanding of the archer's own strengths and weaknesses. Perhaps the archer is stronger on one side of his or her body than the other. Perhaps the archer is healing from an injury and needs to plan in advance for certain kinds of scenarios in which the injury might prove to be an issue.

Reflective Endorsement is not used skillfully when we simply affirm whatever we choose to affirm. A crucial piece of the puzzle of what we really care about involves looking at our own past behaviors. We can look back at some of the examples presented in earlier chapters to highlight this phenomenon. Recall Jane, from our discussion of Frankfurt, who thought she wanted to be a lawyer, but who really wanted to be an artist. In the past, she put off or avoided her law studies at all costs but was always eager to sketch or engage in other forms of art. The evidence in this case suggests that she cares about art but does not care about the law. As we have seen, human beings do not frequently take their own past behaviors into account when making character attributions about themselves.

To achieve Socratic Virtue, past behaviors must be taken into account. We must use them to inform us about who we are as people. Imagine that a person conceives of himself or herself as having a commitment to truthfulness, or to timeliness. He or she has a disposition to reaffirm those commitments when asked. They are neglecting very important information if they ignore their own past

behaviors. If, upon reflection of those behaviors, they find that they lie frequently or that they are always late, they must take these observations into account.

It is also important to recognize that key evidence pertaining to these issues might well come from sources other than our own self-assessments. Others with whom we are close may often be in a better position to call to our attention our more reliably performed, persistent behaviors. These may be behaviors that, for whatever reason, we are unable or unwilling to see in ourselves.

A willingness to pay attention to relevant evidence is part of what it is to employ reflective endorsement skillfully. When an archer behaves skillfully, he or she moves in a particular way because of their possession of the appropriate kind of background information. For a person's self-evaluation to be adroit, it must be similarly responsive to the right kind of information. Appreciating the importance of reliable behavioral dispositions is essential to skill of that type.

Knowledge of How Behavior is Associated with Value

For an agent to determine what really matters to them, it will not be enough to simply look at their past behaviors. After all, past behaviors, especially those that occur when people are very young, often demonstrate a lack of understanding about how a behavior is connected to a value or set of values. Coming to have an understanding of the connection between behaviors and values is also important to engaging in reflective endorsement skillfully.

This is a concept that is familiar from Aristotle's ethics. On Aristotle's view, very young people are consumed with pleasure, and it is only after looking to the

virtuous and becoming habituated that the agent recognizes the connection between virtue and action. He says:

> Actions are called just or temperate when they are the sort that a just or temperate person would do. But the just and temperate person is not the one who merely does these actions, but the one who also does them in the way in which just or temperate people do them.[108]

So, for Aristotle, the just or temperate person doesn't simply engage in the appropriate behavior given the situation. The person has been habituated to recognize the connection between the action and the virtue.

Similarly, skillful endorsement must be more than simply empty avowal. An agent's reflective endorsement is adroit when they recognize the range of behaviors that certain endorsements entail. A tendency toward ignorance or misinformation represents diminished skill or, in the more extreme cases, a lack of skill.

Consider the case of a Michelle, who professes to care about the environment. Perhaps this attitude was instilled in her in a somewhat clumsy way in grade school. Despite her claim to care about the environment, she engages in many practices that have serious negative environmental impacts. She eats meat, doesn't recycle, drives her car to work rather than taking accessible public transportation, and allows her car to idle while she waits for her kids to come out of school so that she can keep the air conditioning running in the summer. Her behaviors don't actually bear out her claim to care about the environment.

I'm not denying, here, that Michelle experiences strong reactive attitudes when she introspects about the environment. She meets an internal component for

---

[108] Aristotle, W. D. Ross, and Lesley Brown. 2009. *The Nicomachean ethics*. Oxford: Oxford University Press.

caring. What I am suggesting here is that skillful endorsement has external components as well as internal components. It is not enough for the archer to simply understand how to move his body in order to get the arrow to fly in one direction or another. If he really wants to be skilled at archery, there are other things that he must know. Similarly, if a person really cares about something, whether that thing is a thing, a person, or a practical identity, there is an external obligation to become knowledgeable about the thing in question if a person wants to care well.

What I've said about people who lack information also applies to people who possess inaccurate information. Let's go back to our example of the woman in the Friedman chapter who desires to feed her family and children only healthy food. She mistakenly comes to believe that all and only food that is organic is healthy food. Here, I am reminded of a scene from *Harold and Maude* in which Maude offers young Harold some champagne. When he informs her that he doesn't drink, she replies, "It's alright, it's organic." Let's imagine that our health food minded mother reasons according to similar principles. Of course, not all things that are healthy are organic, and not all things that are organic are healthy. If the woman in this case were to follow the advice offered in the previous section and reviewed her own past behaviors, she could do so in more than one way. The first way is to look at whether she was always motivated by a desire to feed her children healthy food. She could also evaluate whether she has, in the past, been the best judge of what constituted healthy food. So, one point of analysis may be not simply what an agent's past behavior happened to be, but also the extent to which that past behavior was

actually an expression the values that the agent claims to care about. This is an external condition.

Should we, then, look at our past behaviors themselves, or at the intentions behind our past behaviors? I think that the answer is *both.* Sometimes the external world conspires against us in the sense that an action we desired to be of one type actually, in fact, turned out to be a behavior of an entirely different type. The woman intending to feed her children healthy food, but being mistaken about what counts as healthy food is a good example. By contrast, sometimes our actions are quite telling when it comes to what we actually care about, but we our own intention is not transparent to us. I may act in such and such a way because I'm afraid of stepping outside my comfort zone, all the while I am totally unaware that I even have a comfort zone and don't know what it would look like to be outside of it.

Reflective Endorsement and Protecting Against Hueristics

We saw in the previous chapter that behavior that we take ourselves to be endorsing for one reason, we might, in fact, be endorsing for different reasons altogether. For example, Melinda might introspect and confirm to herself that she is a good student. She might do this, even in light of the fact that it is roughly midway through the semester and she has entirely given up on attending class regularly. Instead of paying much attention to her attendance, she focuses on her ability to score decently well on exams by simply searching the Internet for information listed as weekly topics on the syllabus. Why is she capable of engaging in such impressive mental gymnastics? It may well be that it is because self-serving heuristics have

come to save the day, if not with respect to her grade at least with respect to her self esteem.

Michelle, a previously overweight person who has been on an exercise kick for years might conceive of herself as overweight. She might discount the opinions of those who tell her otherwise and may choose to spend time with those who confirm the negative image that she has of herself. A validation heuristic might be to blame.

When reflective endorsement is adroit, those engaging in it are aware of the possibility that the mind's use of heuristics might masquerade as authentic self-evaluations. That such a heuristic is operative is always a possibility—serving as persistent local skeptical hypothesis. We may never be able to rule it out entirely, but, if we want to be skillful, we must keep the possibility in mind and do what we can to protect against it. Like an archer making apt shots, the apt reflective endorser is familiar with their equipment and knows how to protect against its potential deficiencies.

Apt Reflective Endorsement

Finally, we can paint a picture of what it would look like for reflective endorsement to be apt. Recall that an archer's shot is apt, not simply under the conditions that he or she makes the shot, or simply under the condition that the shot is skillful. In order for the shot to count as apt, the archer must make the shot *because* he or she is skillful.

Apt reflective endorsement may look very similar. When a person introspects aptly, they introspect by making use or the best, most reliable information available to them. If they arrive at a position of value, they must do so, not as a lucky result of useful heuristics, but because they are reflecting correctly.

BIBLIOGRAPHY

Alicke, Mark D., Klotz, M.L., Breitenbecher, D.L, Yurak, T.J., and Vrendenburg, D.S. (1995) "Personal contact, individuation, and the better-than average effect." *Journal of Personality and Social Psychology, 68,* (1995): *804-825.*

Alicke, Mark D., Vrendenberg D.S., Hiatt, M., And Govorun, O. (2001) "The "better than myself" effect." *Motivation and Emotion, 25,* (2001): 7-22.

Alicke, Mark D., Govorun, Oleysa, "The Better Than Average Effect." *The Self in Social Judgment.* Edited by Mark Alicke, David A. Dunning, and Joachim I. Krueger. New York: Psychology Press, 2005

Aristotle, W. D. Ross, and Lesley Brown. *The Nicomachean Ethics*. Oxford: Oxford University Press, 2009.

Arpaly, Nomy. *Unprincipled Virtue*. New York: Oxford University Press, 2003.

Ayer, A.J. 1936, *Language, Truth, and Logic*, London: Gollancz, 2nd Edition, 1946.

Brady, Michael. 2002. "Skepticism, Normativity, and Practical Identity," *The Journal of Value Inquiry,* vol.36, (2002): 403-412.

Camus, Albert. 1965. *The Myth of Sisyphus, and Other Essays*. London: H. Hamilton, 1965

Codol, J.P. "On the So-Called "Superior Conformity of the Self" Behavior: Twenty Experimental Investigations." *European Journal of Social Psychology, 5,* (1975): 457-501.

Dancy, J. *Normativity.* Oxford: Wiley-Blackwell, 2000.

Descartes, R. and Cress, D. *Meditations on First Philosophy*. Indianapolis: Hackett Pub. Co., 1993.

Fitzpatrick, William. 2005. 'The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism, and the Nature of Normativity," *Ethics,* vol. 115, (2005): 651-691.

Freeman, Samuel. *Rawls.* New York: Routledge, 2007.

Frankfurt, Harry. *The Importance of What We Care About.* Cambridge: Cambridge University Press, 2006.

Frankfurt, Harry, *The Reasons of Love.* New Jersey: Princeton University Press, 2006.

Frankfurt, Harry, *Taking Ourselves Seriously and Getting it Right.* California: Stanford University Press, 2006.

Friedman, M. *Autonomy, Gender, Politics*. Oxford: Oxford University Press, 2003.

Geach, P. "Ascriptivism." *Philosophical Review, 69* (1960).

Gibbard, A., 1990, *Wise Choices, Apt Feelings*, Cambridge: Harvard University Press.

Ginsborg, Hannah. 1998. "Korsgaard on Choosing Non-Moral Ends," *Ethics,* vol. 109, No. 1, (1998): 5-21.

Hume, David. *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, 2nd ed. revised by P. H. Nidditch, Oxford: Clarendon Press, 1975.

Jones, E.E., and Harris, V.A., The Attribution of Attitudes. *Journal of Experimental Social Psychology, 3,* (1967): 1-2.

Korsgaard, Christine. 'Two Distinctions of Goodness', *The Philosophical Review,* vol. 92, no.2, (1989): 169-195.

Korsgaard, Christine. *The Sources of Normativity,* Cambridge University Press, 1996.

Korsgaard, Christine. "Self Constitution in the Ethics of Plato and Kant," *The Journal of Ethics*, vol. 3, (1999): 1-29.

Langer, E.J., "The Illusion of Control." *Journal of Personality and Social Psychology, 32,* (1975): 311-328.

Miller, A., *An Introduction to Contemporary Metaethics.* Massachusetts: Polity Press, 2013.

Moore, G. E. *Principia Ethica*. Cambridge: At the University Press, 1903.

Moskowitz, Gordon B., *Social Cognition: Understanding the Self and Others*. New York: The Guilford Press, 2005.

Nagel, Thomas. "The Absurd*." The Journal of Philosophy,* Vol. 68, No. 20, Sixty-Eighth Annual Meeting of the American Philosophical Association Eastern Division (Oct. 21, 1971): 716-727.

Nisbett and Ross, *The Person and the Situation.* London: Pinter & Martin Ltd 2nd edition: 2011.

Nisbett, R.E. Caputo, C., Legant, P., and Maracek, J. (1973) Behavior as Seen by the Actor and as Seen by the Observer. *Journal of Personality and Social Psychology*, 27, (1973): 154-164.

Perfloff, L.S. and Fetzer, B.K. "Self-Other Judgments and Perceived Vulnerability to Victimization." *Journal of Personality and Social Psychology, 50,* (1986): 502-510.

Plato, *Plato: Five Dialogues: Euthyphro, Apology, Crito, Meno, Phaedo.* Edited by John M. Cooper. Indianapolis: Hackett Publishing Company, Inc.; 2 edition, 2002.

Railton, Peter. "Normative Force and Normative Freedom: Hume and Kant but not Hume Versus Kant," *Ratio,* vol.12, issue 4, (2002): 320-355.

Raz, J., *Engaging Reason: On the Theory of Value and Action*. Oxford: Oxford University Press, 1999.

Rawls, John. *A Theory of Justice.* Harvard University Press, 1971.

Sartre, Jean-Paul. *Being and Nothingness; An Essay on Phenomenological Ontology*. New York: Washington Square Press, 1966.

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

Smith, M.,"The Humean Theory of Motivation," *Mind*, 96, (1987): 36–61.

Sosa, Ernest. *A Virtue Epistemology*. Oxford: Oxford University Press, 2009

Skidmore, James."Skepticism about Practical Reason: Transcendental arguments and their Limits", *Philosophical Studies*, vol.109, no 2, (2002): 121-141.

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

Swan, W.B., Jr. "To be Adored or to be Known?: The Interplay of Self-Enhancement and Self-Verification." In E.T. Higgens and R. Sorrentino (Eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior* (Vol. 2 pp. 527-561). New York: Guilford Press, 1990.

Thomson, J., *Normativity.* Illinois: Carus Publishing, 2008.

Wallace, R.J. *Normativity and the Will: Selected Essays on Moral Psychology and Practical Reason.* New York*:* Oxford University Press, 2006.

Wedgwood, Ralph. "Practical Reasoning as Figuring Out What is Best: Against Constructivism," *Topoi*, vol. 21, (2002): 139-152