UNIVERSITY OF CALIFORNIA, SAN DIEGO

Breakdown of Morality

A dissertation submitted in partial satsifaction of the requirements for the degree Doctor of Philosophy

in

Philosophy

by

Eric Michael Campbell

Committee in Charge:

Professor David Brink, Chair
Professor Richard Arneson, Co-Chair
Professor Christine Harris
Professor Dana Nelkin
Professor Joel Robbins
Professor Donald Rutherford

2012

The Dissertation of Eric Michael Campbell is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2012

DEDICATION

This dissertation is dedicated to:

Doug Peters, for early philosophical conversations without which I would have been less sharp but thought myself sharper;

Tarun Menon, for later philosophical conversations in the course of which I was taught a lot of philosophy and even more humility;

Jamin Luoto, for exemplifying the virtues of friendship and proving that the Light of God can exist without God.

Keith Campbell, for putting up with many years of jackassery and for developing his talent for making fun of me into a subtle and refined art form;

My mother, for putting up with as much combative argumentativeness and stubbornness as can rightly be asked of any mother, for doing what she thought was the right thing even when it wasn't the easy thing—and even when she knew she didn't know if it was the right thing, for matching stubborn with stubborn until she didn't have to anymore, and then for supporting me in every way she can;

George Ainslie, for providing much of the empirical and conceptual foundation for my commitment model of moral judgment, and for helping to illustrate and explain the potential pathologies of (especially unwitting) commitment strategies;

Herman Melville and Friedrich Nietzsche, for helping teach an orphan how to love the open sea;

And to the things that didn't kill me.

EPIGRAPH

It goes without saying that I do not deny -- unless I am a fool -- that many actions called immoral ought to be avoided and resisted, or that many called moral ought to be done and encouraged -- but I think the one should be encouraged and the other avoided *for reasons other than hitherto*. We have to *learn to think differently* -- in order at last, perhaps very late on, to attain even more: *to feel differently*.

Friedrich Nietzsche

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge Professor David Brink for getting back to me with two pages of single-spaced general criticism and 82 targeted critiques of the nearly 100-page prospectus that I would have turned in too late to defend on schedule but for the fact that he responded so quickly, allowing me to revise in time. This example stands for many; he is a model of professionalism as well as intellectual and personal responsibility, and the standard he set as an advisor is one I hope to be able to approach in my own career. His patient support and thoughtful criticisms of the dissertation have made it much better than it otherwise would have been.

I would also like to acknowledge Richard Arneson for, when I was originally toying with the idea of desire-dependent reasons, continuing to pester me with the question of why 'me-now' should care about what 'me-later' wants or cares about. Out of the frustration of trying to answer that question came a central aspect of this work. Even more important, his sustained enthusiasm for the project lightened my spirits in heavy times.

I would also like to acknowledge John Jacobson for reading Ainslie with me and for saying early on that people are still looking for the Ten Commandments.

VITA

2001	Bachelor of Arts in Philosophy with Honors University of Texas at Austin
2003	Awarded UCSD Humanities Fellowship
2004 – 2010	Teaching Assistant, Department of Philosophy University of California, San Diego
2008	Master of Arts, University of California, San Diego
2012	Doctor of Philosophy, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Breakdown of Morality

by

Eric Michael Campbell

Doctor of Philosophy in Philosophy

University of California, San Diego, 2012

Professor David Brink, Chair
Professor Richard Arneson, Co-Chair

My dissertation has three main parts. In the first I develop a commitment model of moral judgment. I argue that moral judgments and the broader discourse in which they take place can be understood in terms of the operation of distinct but interacting commitment strategies. To a first approximation, these strategies operate at the levels of individual and social psychology, and biological and/or cultural evolution. All commitment strategies provide motivational stability by reducing one's (perceived) flexibility of action. Some such strategies are undertaken deliberately, such as when an

addict signs a contract forcing her to donate money to a despised organization if she is caught using. However, the commitment strategies associated with moral judgments are rarely if ever conscious or deliberate. In fact, a central aspect of my commitment model holds that the peculiar motivational power of moral judgments is importantly connected to their power to deflect attention away from our actual motivations and values, and that this motivational strategy is undermined by an awareness of its workings.

Introduction

Know thyself. Though I have not always realized it, this Socratic injunction is at the heart of the interrelated philosophical projects that have come together in this dissertation. The core ethical contribution I hope to make here lies in illustrating the value of spending much more time and effort inquiring into to our own motivations and values in the course of ethical inquiry than many of us (especially moral philosophers) tend to do. This recommendation can be called the positive face of my proposal. The central motivation and justification for it is that our *actual* values, concerns, attitudes and the like are the sole sources of our reasons for what to do and how to live. The negative face is my advocacy for a departure from moral discourse, a position I call moral abolitionism. That negative proposal is meant to be in the service of the positive one; I think it is a design feature of moral discourse to deflect attention from our motivations and values, and so my abolitionism is proffered as part of strategy to learn how to better attend to just those things when thinking ethically, i.e., thinking about what to do and how to live.

Though I would like to move beyond moral discourse, I would never have undertaken this project but for my having had very strong moral(ized) commitments nearly all my remembered life. The often intense emotional nature of those commitments, combined with the fact that others either did not share them at all or did so less intensely, motivated me to try to justify and make sense of them. I was particularly troubled by the seeming conflict between the (rational) requirements of self-interest and morality, and could see neither how to resolve this conflict nor how to give one kind of requirement priority over the other. Thinking of rationality in terms of desire-satisfaction rather than self-interest did not help to resolve this conflict.[1] My many years of struggling with these issues has resulted in this dissertation, which is divided into three main parts.

The first part is an account of the essential nature of what I will call peculiarly moral judgments (or discourse). I'll argue that peculiarly moral judgments have an essentially committing function. This is not to say that that's all moral judgments are or do, or that all moral judgments (conceptually or necessarily) have the function of committing someone to something. I don't think moral judgments are a natural kind or have necessary and sufficient conditions for their existence. There are sure to be cases of things we (at least some of us) are willing to call moral judgments that do not seem to have a committing function or effect.

However, I'll argue that *peculiarly* moral judgments can be understood as essentially commitment devices, or an essentially committing 'technology'. I take peculiarly moral judgments to be those that the maker of the judgment does not and cannot (consciously) conceive as relative to her own attitudes, values, drives, interests or (other) motivations[2] without an important loss. I do not regard the loss to be essentially one of meaning, but of motivation; or better, a motivational strategy. This strategy is rarely if ever conscious or deliberate. In fact, I will be arguing at some length that the peculiar motivational power of moral judgments is importantly connected to their power

[1] Though thinking of reasons as related to desires did point the way toward both my account of both practical reason and the nature of moral judgment.

[2] For simplicity's sake I will often say only 'motivations' to stand for all these things.

to deflect attention away from our motivations, and that this motivational strategy is undermined by an awareness of its workings.

The second part of the dissertation comprises a single long chapter and is a general account of practical reason. There I make as strong a case as I can for the claim that all practical reasons are made from what I call 'motivated perspectives'. My account of practical reason is neo-Humean in that it rejects the possibility of intrinsically irrational desires and holds that all reasons for action are in relation to some motivation(s), more specifically, in relation to the myriad things that we care about. The third and final part of the essay stretches over three chapters. The first two can be regarded as a kind of extended *practical* moral error theory, and the final chapter focuses on specific benefits of moving beyond moral discourse.

Unlike the most familiar and influential moral error theories, such as those of John Mackie (1977) and Richard Joyce (2001), mine does not identify some essential conceptual feature or presupposition at the heart of moral discourse and then argue that that presupposition is indefensible. The conclusions of such error theories are that all moral judgments are false, or perhaps nonsensical or otherwise 'untrue'. I will expend a considerable effort trying to show that whatever essential error there is in employing moral discourse is not essentially theoretical but practical, and the essential nature of that practical error lies in the tendency of moral discourse to deflect attention away from our own and others' motivations. My aim in these final chapters is to argue that (especially) once we are aware of the nature of both moral judgment and practical reason, peculiarly moral discourse loses much if not all of its point, holds out significant potential for pathology, and/or distracts us from the core normative project of discovering and evaluating our values. In the remainder of this introduction I'll give a more detailed account of my project and argumentative strategy.

I said above that when one makes a peculiarly moral judgment,[3] one does not (consciously) conceive of the moral values or moral reasons at stake as relational, especially not as related to one's own or others' motivations. The way I just put that claim is unusual—in fact, to my knowledge, unique in the relevant literature. First, I take the essential feature of moral judgments to be their apparent *nonrelationalism*, in the sense described. This is a relatively minor point, though I think it avoids the confusion often brought on by saying that moral judgments are meant to be *objective*. As Mackie pointed out in arguing against the existence of 'objective values', there are many ways for evaluations, and therefore values, to be objective. Specifically, evaluations can be *objectively* correct *relative to* ends or standards or institutional requirements. He even granted that relative to such requirements, one could have categorical reasons for action, i.e., reasons that are not relative to one's motivations. However, in denying that any such reasons were 'objectively valid', he was precisely denying that they were nonrelational 'all the way down'. So really, what Mackie was interested in denying is, I think, more perspicuously described as *nonrelational*, or as I will sometimes call them, *intrinsic* values.[4]

The second aspect of the way I characterized peculiarly moral judgments is perhaps less salient, but it is more important, and is the aspect of what I said that I think

[3] From now on I will often simply call them moral judgments, but I will always mean peculiarly moral judgments.

[4] Throughout this essay I will use intrinsic value to mean nonrelational value, unless otherwise noted. It is also often used to mean 'final value', i.e., a value which is not for the sake of any other value. I think the concept of final value is very important in ethics, whereas I think that of intrinsic value is not, and is the source (or sustainer) of many serious confusions.

is unique in the literature. I characterized moral judgments in negative terms, specifically in terms of what the makers of such judgment do *not* perceive about those very judgments. All the standard arguments for moral error theory proceed by identifying some presupposition supposedly essential to moral discourse, and then arguing that that presupposition, and therefore all moral judgments, are false. Mackie argues that they presuppose objectively valid categorical reasons, and Joyce argues that they presuppose categorical rational authority and applicability (which amounts to essentially the same thing as what Mackie denied). Bertrand Russell thought that 'our ethical judgments all claim objectivity' and that for this reason they are 'all false' (Russell 1922/1999, p. 123).[5] Those who reject the error theories then argue either that the various presuppositions are true, or that they are not in fact essential presuppositions of the discourse. In either case, moral claims can be true.

I think there is plenty of room for honest and informed disagreement about whether our moral judgments or concepts 'entail' or 'presuppose' or 'contain' these conceptual features, all of which seem to be essentially concerned to deny the relational nature of moral value and/or reasons. In chapter 5 we'll see that some philosophers deny that moral concepts are nonrelational. Stephen Finlay argues that they are in fact conceptually 'end-relational', very often related to the ends of the maker of the judgment. This might be correct, but if it is, the makers of peculiarly moral judgments are *not aware* of it. That is, even if their moral concepts and judgments are end-relational in the way Finlay suggests, they do not conceive of their own concepts in that way, and most would

[5] Others have claimed that our moral judgments presuppose a divine lawgiver (Anscombe, 1958), or libertarian free will, or an untenable view of human character, or a strong motivational internalism.

sincerely deny it if asked. In other words, whether or not conceptual antirationalists are right that our concepts of moral value do not presuppose the kind of 'objectively valid' categorical reasons that Mackie denied, *the makers of peculiarly moral judgments are not aware of this fact* (at least not while making the judgments). Therefore I regard what is at the heart of peculiarly moral judgments not to be some positive conceptual feature, but the *absence* of a (conscious) conception of moral reasons and obligations as relative to one's motivations or values.

Despite the fact that I conceive of my 'error theory' as essentially practical rather than theoretical, it has very important affinities with the error theories of Mackie and Joyce. What we (and some others) regard as most peculiar about moral judgments is connected with their putative reason-giving force, as well as their manifest ability to generate motivation. For us, the most troubling or problematic feature of moral concepts[6] are those involving the core notions of moral obligation, duty, permissibility and the like.[7] These notions seem, to many at least, to involve 'objective prescriptivity' as Mackie (1977) put it. Mackie argued that such properties were metaphysically 'queer', that if they existed they were of 'a very strange sort, utterly different from anything else in the universe' (1977, 38) and it was equally mysterious how we could come to be in any kind of reliable contact with them.

[6] As I've said, I'm not committed to whether these features are actually part of moral concepts or 'only' their pragmatics. But for simplicity's sake, esepecially when discussing authors who do assume that the problematic elements are part of moral concepts themselves, I will sometimes speak of moral concepts rather than moral judgments or discourse.

[7] Joyce and Mackie focus on these, but I think the notions of moral responsibility and desert are every bit if not more problematic, and for precisely the same reasons—they systematically deflect our attention from our own motivations in holding people responsible.

Mackie was objecting to the notion of a *categorical imperative*.[8] A categorical imperative differs from a hypothetical imperative in that the former is a kind of command that is supposed to apply and give reason for action to a person independently of any of that person's goals. The latter only gives a reason to those who have an end served by the imperative. 'Leave the house by ten o'clock' is a hypothetical imperative if the addressee is understood to have some goal that will be served by doing so, such as catching a flight. On the other hand, 'Don't steal' is meant as a categorical imperative if there is no such presuppostion (and yet it is implied that the person nevertheless has a reason not to steal). Mackie thought that such imperatives were too strange to be believed in. Since Mackie thought that moral commands were just these sorts of commands, he thought that they were committed to a sort of thing that does not exist and were therefore all false.

Following in Mackie's steps, Joyce (2001) maintains that moral discourse is essentially committed to categorical imperatives, and that there is 'no sense to be made' (p. 46, *passim*) of the reasons that they purport to provide. As I said above, taking such a position as Mackie and Joyce do is called being an 'error theorist' about morality. Error theorists in general are those who, for whatever discourse they are error theorists about, take that discourse to be committed essentially to some claim that is false (and/or impossible or incoherent). Atheists are error theorists about God-discourse and we are all (I surmise) error theorists about phlogiston[9] and witches. What I want to highlight here is

[8] Technically, he objected to a categorical imperative that is 'objectively valid'. I will get to this distinction soon.

[9] A substance that was thought to be stored in objects and released during combustion; it was abandoned upon the success of the replacing theory that oxygen was consumed in combustion, leaving no role for phlogiston.

that both are error theorists about morality because of what they take to be its commitment to a mystical and/or mythical kind of practical authority.

Let me take a moment to motivate the idea that morality is centrally committed to categorical imperatives. Let's suppose that moral judgments were not so committed. In that case, should it be discovered or believed that a person had no goals that required her to refrain from stealing, murdering, lying, torturing and/or whatever else you want to dream up, then she would have no reason not to do those things. Joyce reasonably concludes that moral obligations are not 'escapable' in this way. Therefore, he holds that *as a conceptual matter*, one cannot get out of the relevant moral obligation by sincerely claiming that one has no aims inconsistent with killing for money. The common way of thinking of moral requirements is that one simply *must not* do certain things, regardless of whether one wants to or not, or whether it is in one's interests or not.

The way I've just put Joyce's argument invites a distinction I've put off until now. One might admit that as a conceptual matter one cannot escape a moral obligation by (truthfully) claiming a lack of ends or interests suitably connected to the obligation. Joyce could be right that moral obligations are not escapable in this sense, but this would not show that such judgments need have rational authority over such agents. This is to distinguish *applicability* from (rational) *authority.*[10] A moral judgment might *apply* to you, but not have rational *authority* over you, no matter your ends. However, both Mackie and Joyce argued that moral judgments do assume not only categorical applicability but also authority.

Joyce follows Mackie in agreeing that there can be and are categorical imperatives,

[10] We will discuss this distinction in more detail in later chapters.

but none of them are ‘objectively valid’ in Mackie’s terms. They are all ‘institutional’, in the sense that many and various institutions have rules that both apply to and have a kind of authority over those governed by the institution. But Mackie and Joyce maintain that one’s reasons for following such institutional rules depend on one’s choice or desire to be a part of the relevant institution. For example, the rule against moving your bishop vertically along the chess board both applies to you and rationally constrains your behavior—so long as you have the goal of playing chess. If you do not have the goal or desire to play chess, you likewise do not have chess-related reasons not to move your bishop however you like. Joyce and Mackie think that all reasons are goal- or desire-dependent in this way, but that the core moral concepts presuppose that there are a class of reasons whose authority in no way depends on goals or desires that one might or might not have. Joyce calls this combination of categorical rational applicability and authority ‘practical clout’. I follow this usage.

This essay will be centrally concerned with the connections between the ostensible practical clout of morality and moral motivation. While people disagree about whether moral judgments presuppose practical clout, it is agreed on all sides that there is an intimate connection between moral judgments and motivation. Motivational internalists believe that moral judgments *necessarily* provide motivation (though some, e.g., Michael Smith (1994) provide a ceteris paribus clause here to allow for cases of severe depression and the like), while externalists deny this. Nevertheless, both recognize the widespread connection between the judgments and motivation.

A fruitful way to approach the issue of moral motivation is by means of a puzzle. The appearance that moral judgments are a kind of belief, which purport to state matters of fact about the world, combined with their intimate connection with motivation, can, in connection with widespread assumptions in folk psychology, give rise to a 'puzzle about moral motivation'. The following version (as well as the just-quoted name for it) is taken from David Brink (1997, p. 6):[11]

> 1. Moral judgments express beliefs.
> 2. Moral judgments entail motivation.
> 3. Motivation involves a desire or pro-attitude.
> 4. There is no necessary connection between any belief and any desire or proattitude.

This quartet is inconsistent. As Brink notes just below this puzzle, many metaethicists' views can be seen as rejections of one of these claims on the strength of the other 3. Following him, I summarize these positions below.

Noncognitivists (such as A. J. Ayer, C. L. Stevenson, Simon Blackburn and Allan Gibbard) reject (1), claiming that moral judgments are really in some way or other expressions or reflections of some noncognitive attitude. All who accept (1) are then cognitivists, and must pick from the other 3 a claim to reject on pain of inconsistency. Externalists (such as Brink and the early Phillippa Foot) deny that moral judgments always come with motivation. Rationalists reject (3) or (4). Some rationalists argue that the sheer belief that one has a moral duty can motivate without any desire or pro-attitude.

[11] Michael Smith's (1994) presents itself as a solution to a similar puzzle, which he (titularly) calls 'The Moral Problem.' Smith's puzzle combines Brink's (3) and (4) into one element, and Smith's (2) includes a ceteris paribus clause. I prefer Brink's version for reasons that need not detain us.

Others argue that the moral beliefs ‘entail’ or necessarily bring with them the required desire. Thomas Nagel (1970) seems to fit this view, as well as possibly Smith (1994).

A core aspect of this essay will be dedicated to providing an answer to the question how it is that moral judgments take the form of beliefs and yet are capable of often tremendous amounts of motivational power. In order to do so, I am going to make extensive use of the work of George Ainslie, whose work over the past (roughly) 40 years has culminated in a model of the will that I (but not only I) find extraordinarily fascinating and fecund, and potentially very valuable in ethical philosophy. One of the many potential contributions I believe Ainslie’s model of the will is capable of making is providing at least a partial answer to the above puzzle.

The solution I propose, in part derived from Ainslie’s model of the will, is to deny (4), but in an antirationalist fashion. While Nagel and Smith think that moral beliefs (at least ceteris paribus) ‘entail’ motivation due to the workings of rational principles that stand apart from and (somehow) govern desires, Ainslie’s model of the will suggests that moral beliefs can be understood as essentially in the business of commiting us to some of our desires, or what I will sometimes call ‘motivated perspectives’. Ainslie’s model shows us how beliefs in desire-independent reasons and values can be motivational in at least two ways, both of which rely on deflecting attention away from our motivations. While Ainslie’s model is not required to support this specific claim, it does give an experimentally-based, unifying model of the will that explains the motivational role(s) of this deflection of attention.

Because of the large amount of work Ainslie’s account is doing for me, I devote the entire first chapter to it, in part because it has to be elaborated enough to render plausible what is in some ways a counterintuitive model of the will. As I've already mentioned, we'll see how this model has the resources to explain how some kinds of beliefs, especially beliefs in attitude-independent moral requirements and intrinsic value, can be explained in terms of their motivational effects. The primary purpose of deploying Ainslie's model is to explain how we could have the belief, as well as the strong *feeling* that we have good, powerful reasons to do something even without having any corresponding aims. Contrary to the way it can often seem to us, all these beliefs can be explained as essentially in the business of generating, stabilizing and/or intensifying motivation. Their representation as beliefs about mind-independent facts is hypothesized to render the commitments they represent apparently free from being tampered with or hedged in ways that can undermine the motivation which it is their business to maintain. This explanation crucially depends on the motivational power of subjectively limiting our own options and the deflection of attention from our own motivations.

Though extremely valuable, Ainslie's model is far from adequate for my purposes. It gets us on the right track by showing us the enormous theoretical significance of commitment and its side-effects, but by itself does not give us close to as complete a story about moral judgments and beliefs as we would like. Most obviously, it leaves out an account of the moral emotions and how we come to have them, as well as any role of culture in fashioning moral beliefs, judgments and motivations. I want to emphasize that this is not a criticism of Ainslie, since it was no part of his task to explain moral judgments, beliefs or particularly moral motivations.[12]

[12] To the extent that his model does provide explanations for these things, it should be counted as an additional point in the theory's favor.

Chapter 2 will continue arguing for the essentially motivational, committing function of peculiarly moral concepts, with a focus on the role they play in keeping us unaware of the relational nature of our values and reasons.[13] I will draw from, improve on and extend arguments made by Richard Joyce and Robert Frank to the effect that moral judgments (in Frank's case, moral sentiments) are essentially in the commitment business. These authors (both of whom are motivated to *validate* the value of moral discourse and sentiments) both make crucial usage of Ainslean insights, but do not recognize the full importance of what his work can contribute to a commitment model of moral judgment/sentiment. Whereas Joyce and Frank are focused on biological evolutionary accounts of moral judgments and sentiments, I think cultural evolutionary accounts are at least as important as biological ones, and in fact the most compelling versions are of interactive processes at these two levels. Next I will discuss the work of Jonathan Haidt, who is in no way arguing against moral discourse or for morality as commitment-device. I will show that the most plausible aspects of his work support the hypothesis that moral judgment has a committing function.

The arguments in chapter 2 are meant to combine with those of chapter 1. My overall goal will be to show that we can make sense of moral judgments essentially in terms of their tendency to commit us to ways of feeling and acting. I think that together these arguments provide a good (if imperfect and incomplete), unifying explanation of a common conviction that some things have intrinsic value and that moral judgments carry a peculiar motivation-independent authority. Further, they imply that an important subsidiary function of such convictions is to deflect attention away from the workings of

[13] I will give positive arguments for the end-relational nature of reasons and values in Chapter 3.

our own motivations. The motivational role and (dis)value of such deflection will be a central theme of this essay.

The explanations of chapters 1 and 2 are in the ‘debunking’ style, as they attempt to explain beliefs in a way that makes no reference to those beliefs being true, and are meant to engender suspicion about the source(s) of one’s confidence in them. In other words, these explanations are grounds for suspicion to the extent that they give us reasons for thinking that the arguments some of us make in favor of our beliefs in nonrelational values or reasons are post-hoc rationalizations of the way things seem to us. In the context of my debunking explanations of the way these things seem, I hope the post-hoc flavor of those arguments will be more readily perceptible.

Though I hope they achieve a debunking effect, I recognize that these explanations cannot replace a direct engagement with philosophical arguments.[14] That direct engagement is the primary task I set myself in chapter 3. There I will argue directly for a neo-Humean (i.e., desiderative) conception of motivating and justifying reasons. Specifically, I will argue that all reasons and values are relative to, or ‘from the perspective of’ some desire(s). Therefore no reasons or values have any rational authority independently of our own (actual) desires, and nothing is intrinsically (dis)valuable or (ir)rational.

Chapter 3 is very important. It is a crucial aspect of my project to convince my readers that all reasons and values are (most plausibly understood as) relative to what I call ‘motivated perspectives’. Whenever we accept a reason for action or a value-claim,

[14] Except for ‘arguments’ to the effect that it just really seems like this kind of value exists. The arguments in the first two chapters are supposed to explain why it would seem that way without it actually being that way.

that acceptance proceeds from some aspect(s) of our motivational psychology (not that that's *all* it proceeds from). Therefore when someone makes a claim to the effect that something is valuable, there is something about their actual motivations (or motivational structure) playing a crucial role in that judgment, and likewise when someone accepts a reason for action.

This is a crucial aspect of my project for two reasons. The first is that I think that such a conception of reasons and value fundamentally threatens moral discourse. As I said above, normally this threat is couched in terms of providing a crucial premise in an error theory about morality. The standard error-theoretic strategy proceeds by attempting to show that the concepts employed in moral discourse are essentially committed to some features or propositions that are not true, and therefore the entire discourse is systematically flawed.

I doubt that any such error theory is correct, but that is not my primary concern. I think that the conception of values and reasons that I put forward in this chapter, if generally accepted, threaten moral discourse because of the motivational role that a lack of awareness of our motivations plays in peculiarly moral discourse. And I think that peculiarly moral discourse cannot survive (long) in the presence of a general awareness of the relational nature of reasons and values.[15] Of course I don't expect to demonstrate beyond the possibility of rational doubt that my view of practical reasons and values is correct, but I do expect to show that my view is at least as plausible, and I think more so, than any of the leading anti-Humean (or alternative Humean) proposals.

[15] These specific claims, and others related to them, will be elaborated and defended in chapters 4 and 5.

My general strategy in chapter 3 will be to show that a version of neo-Humeanism that is *end-relational* (as opposed to instrumental), and avails itself of some of the empirical and theoretical resources that I've developed in the first two chapters, can handle all the rationalist (anti-Humean) objections of which I am aware. It handles some better than others, to be sure, but whatever work remains to be done in fending off these objections pales in comparison to a very deep problem affecting all of the most prominent rationalist proposals. I argue that all these proposals are committed to a fundamental misconception of the relationships between practical rationality, evaluative belief, desire and the will. All, in their various ways, are committed to the in-principle rational priority of evaluative beliefs over desires in a way that I will argue is demonstrably untenable,[16] and indeed bizarre. Making matters worse, none of these rationalist authors have, to my knowledge, even attempted a response to this problem, though it is potentially fatal to all their proposals. My view, and the model of the will I employ to defend it, not only helps to point out this common rationalist flaw, but also has the resources to explain how such a large problem could have gone unrecognized for so long.

Subsequent chapters are primarily in the business of addressing the question of what to do about the conclusions of the first three chapters. Joyce and Mackie thought that if there is no categorical rational authority, then no moral claims can be true, since moral claims presuppose that there is such authority. I will consider 4 reactions to that conclusion. The first reaction is that if morality *does* presuppose such authority, then we should not believe the conclusion of Chapter 3—or any arguments to the effect that there is no such authority—because of 'our' extremely strong confidence in the truth of at least

[16] At least untenable without allowing for a potantially large gulf between rational and normative behavior.

some basic moral claims. One might reasonably think that one's confidence in the correctness of at least some moral claims is stronger than any conceivable philosophical argument about the nature of practical reason. On this view we should treat the fact that my arguments entail that there is no such rational authority as a reductio of my arguments, if in fact all moral claims presuppose such authority. Or less strongly, one might think the presumption in favor of the truth of at least some moral claims is so strong that one would need far more compelling arguments than I (or anyone else) has provided to thorougly undermine it.[17]

My very brief response to this reaction is to say that the nature of the strong presumption at work here is not best understood in terms of a confidence in the *truth* of moral claims per se, but is fundamentally a reflection of (often very deep and strong) *practical commitments*. I hope that Chapters 1 and 2 will have made this claim even more plausible than it already was, and a large part of the aim of chapters 4 and 5 is to further convince the reader that any strong confidence one has in moral claims is essentially a matter of having various practical commitments. I do not attack any of these practical commitments in this essay; rather, I advocate becoming more aware of their status as practical commitments rather than perceptions of truths that transcend such commitments.

The second reaction is that of the fictionalist, which I address in chapter 4. The fictionalist proposal is also largely a means of addressing the first reaction. If the fictionalist can show that we needn't give up either our deep commitments or even their peculiarly moral flavor, then that might strike at what is really motivating the first,

[17] This is Dworkin's (1996) strategy for resisting Mackie's (1977) error-theoretic arguments.

defensive reaction. That is, Joyce thinks that the error theory will be much easier to swallow so long as we can continue on in our moral discourse as closely as possible to the way we did prior to believing his error theory. Joyce and other fictionalists see the usefulness of moral concepts as so great, that despite the fact that no moral claims are true, it is advisable to retain the discourse in a modified way that does not commit us to speaking falsely (even perhaps lying) whenever we employ it.

The third reaction, which I will deal with in chapter 5, sees the strength of the arguments against practical clout and the strong presumption in favor of the truth and/or usefulness of moral claims generally, as suggesting that we are and/or should be *antirationalists* about morality. That is to deny that morality *essentially* presupposes practical clout. One way to be an antirationalist is to claim that as it stands, morality does not presuppose practical clout, and so moral claims can easily be true in virtue of agents' having the relevant ends. Such a position I'll call conceptual antirationalism (CAR). The other way is to suggest that we *reform* moral discourse such that even if it does presuppose practical clout now, that presupposition is not essential, and we should alter the concept(s) so as to get rid of the offending presupposition. Such a person could be described as a conceptual moral rationalist (CMR), but one who denies that the rationalist conceptual feature is an *essential* one, and so proposes that we change our concepts to make CAR true. Either of these moves would allow us to keep employing the discourse, though there might be a significant amount of psychological adjustment required.[18]

[18] One might think that the adjustment is only required if we are reformists, but I will argue that that is very far from the truth.

I will begin by addressing fictionalism, which might seem a counterintuitive strategy. There have been antirationalists about morality for at least hundreds of years, including most obviously Hume, while fictionalism about morality is relatively recent and without a comparable pedigree. In addition, I have no commitment to the claim that morality is essentially conceptually rationalist. One might think then that my first order of business should be to explain why we should not simply be antirationalists; surely at first glance antirationalism is preferable to fictionalism and therefore should, so it seems, be addressed first. It is for this reason that Joyce (2001) devoted considerable attention to arguing that moral discourse without practical clout would simply not be recognizable as moral discourse at all. Only then did it make sense to pursue a fictionalist alternative to moral belief.

If I were interested in an error theory as such (i.e., conceived in truth-theoretic rather than practical terms), then it would have also made sense for me to address antirationalism first. However, it is the dual threat and opportunity posed by the loss of a sense of what we've been calling practical clout that I am concerned with. I do think that peculiarly moral discourse loses its point without this sense, whether or not we would or should continue to consider some moral judgments true. The reason I address fictionalism first is that it is put forward as a normative proposal about what to do given that all moral claims are false. In addressing this overtly normative proposal, I get an opportunity to argue that 1) the important question isn't whether moral claims can be true, 2) the benefits of moral discourse are overrated and the costs almost entirely unrecognized, and 3) that Joyce's fictionalist proposal amounts to a project of deliberate self-deception and/or training ourselves to be unaware of our own motivations that is as psychologically unrealistic as it is undesirable.

Fortunately, having done so in that context will allow me to argue in chapter 5 that a *self-conscious*, substantive (as opposed to merely conceptual) antirationalism faces similar problems to fictionalism. That is, whether or not antirationalism is conceived as true of our current moral concepts or reforming them, being *aware* of the end-relational nature of reasons and values threatens whatever is peculiarly valuable about moral discourse. And to the extent that one is *unaware* of it, one is subject to the very serious downsides of moral thinking that I will have described in arguing directly against the (net) practical benefits of fictionalism. The problems of a self-conscious antirationalism are surprisingly similar to some of those we will have dealt with in addressing fictionalism.

Chapter 5 will not only argue against antirationalism as a practical proposal, but also show that arguments between error theorists and antirationalists that focus on the question whether moral concepts conceptually presuppose rationalism (or practical clout) are potentially interesting, but of quite secondary importance to the practical issues that come with a better understanding of the nature and value of moral discourse. An antirationalist response to the error theorist that is focused narrowly on semantics and does not concern itself with the *threat* that the lack of a sense of practical clout represents does not address what is most significant about an error theory predicated on undermining the basis of our feeling that moral considerations have some special authority.

Once we accept (and internalize) that our reasons are relative to our own motivations, moral discourse will no longer be able to serve its peculiar function(s), and further, even if it could, it is far from clear that we should want it to. Chapter 6 ends the essay by making the positive case for moving beyond moral discourse to 'straight talk'. What I mean by this includes talk about our cares, concerns, values, emotional and/or aesthetic responses, motivations, commitments, and the like, including which seem to us most central, important, deep, to have greatest priority, etc. Straight talk is not limited to descriptive claims, but can and should include normative ones about what we should do or value or be motivated by. But insofar as they remain forms of straight-talk, there will be no attempt to disguise the fact that these claims proceed from our own end-relational, or motivated perspectives.

A helpful way to think of the value of straight talk is as a means of internalizing descriptive and normative subjectivism, which theses follow from what I will have argued in Chapter 3. Descriptive subjectivism is the claim that there are no practical reasons or values metaphysically independent of our contingent (dispositional) motivational states. That view leads to a kind of practical nihilism unless one also rejects normative objectivism, which is the claim that the *authority* of our values *depends* on their having objective standing. To the extent that we accept descriptive subjectivism, unless we thoroughly reject normative objectivism in favor of normative subjectivism, we have a nihilistic cocktail. Normative objectivism and descriptive subjectivism together entail the rejection of the normative authority of not only our current values, but any values we might come to possess.

Therefore I think internalizing normative subjectivism is important to avoid a kind of practical nihilism, where we accept descriptive subjectivism but still hold on to (shadows of) normative objectivism.[19] Moral discourse is built for descriptive and normative objectivism. Part of what I aim to do in Chapter 6 is show how poorly suited it is to normative subjectivism. Therefore to internalize normative subjectivism, it's important to stop speaking in moral discourse, so as to start getting comfortable locating the normativity of the socio-moral domain in our actual concerns. The other part is to describe the kinds of benefits to be had in doing so. These benefits go well beyond defeating practical nihilism—though that is no small thing!—and mostly consist in, and flow from, getting to know ourselves better.

[19] That many people currently hold something like this combination I believe partly explains the apparent prevalence (among undergraduates at least) of what Williams (1985) called 'vulgar relativism'.

Chapter 1: The Mysterious Bane

1.0 Brief Overview of Two Approaches to the Problem

There have been two primary approaches to solving the puzzle of akrasia, or weakness of the will. The first can be called cognitivism or rationalism and begins with Socrates, who denied the possibility of akrasia.[20] For Socrates, what appeared as akrasia was the result of a lack of 'that particular knowledge called measuring' that made objects appear larger as they drew close to one's eyes and smaller as they moved farther away.[21] Contemporary philosophers and psychologists[22] in this tradition regard the faculty of judgment or reason to be the prime mover**,** and acts which appear akratic are the result of miscalculation or some other cognitive error. To the extent that their passions get the better of their reasons, it is because the latter have mistakenly 'give[n] priority'[23] to those feelings. Though they also employ the metaphor of 'strength,' they haven't been able to offer specific mechanisms by which sometimes the judgments win out over the passions and sometimes not (Ainslie, 2001, p. 15).

[20] Socrates had in mind what we call now 'synchronic' akrasia, which involves performing some action X while *at the same time* believing or even knowing that not-X was the right thing to do. Though there is an interesting question whether and in what circumstances synchronic akrasia does happen, it won't be my focus, and it isn't Ainslie's. What will interest me is the ramifications of the process(es) by which the self-defeating behavior described below (i.e., 'diachronic' akrasia) is overcome.

[21] Plato's *Protagorus*, sections 356-7 in Jowett's translation (1892/1937), quoted in Ainslie (2001, 4).

[22] Examples of philosophers: Bratman (1987, 1999); Davidson (1980, pp. 21-42); Parfit (1984). Psychologists: Baumeister and Heatherton (1996); Kuhl, (1994); Perris et. al. (1998); Polivy (1998); summaries in Karoly (1993) and Mischel et. al. (1996).

[23] Baumeister and Heatherton (1996), quoted in Ainslie (2001, 17).

The second approach employs the resources of utility theories.[24] Utility theories employ the notion of 'reward,' which can be understood as a single dimension along which all substitutable options are chosen. According to these theories, the values of all desires[25] compete against one another in a kind of 'comprehensive internal marketplace,' where the desires that 'win' promise to maximize expected reward. Prospective rewards are assumed to be 'discounted' with respect to the future, such that the absolute value of some reward will be valued more highly as the reward approaches in time. Crucially, the rate of discount is assumed to be a fixed percentage per unit of time. For example, if something is worth 100 utiles (a unit of reward) to you today, and you discount it at a rate of 10% per day, then *today* you would only give 90 utiles to have it *tomorrow*, 81 (= 90 - (90 x .9)) utiles to have it the day after that, and so on. When preferences for rewards are plotted against time using this assumption of a fixed percentage of discount per unit time, the resulting curve is exponential. The important thing to know about these curves, in the context of an investigation into akrasia, is that a preference for one reward over another mutually exclusive reward will not change as the delay until reward is available is increased or decreased.[26] That means that if you now prefer staying sober later tonight so as to feel sharp tomorrow, your preference will not be predicted to change as a sheer function of time.

[24] Description of utility theories and Ainslie's model itself taken from his (2005), unless otherwise noted.
[25] Since desire and not reason is thought to be the 'prime mover' here, such theories can be called 'noncognitive.'
[26] There is still debate over whether any or how much discounting of the future is rational, after taking uncertainty into account. That is an interesting question, but not one addressed here. It isn't obviously connected to what we think of as *akrasia*, since those who discount may and often do think, even on theoretical grounds, that the discounting is rationally justified.

Since utility theorists posit these exponential preference curves, they seem to have no way of predicting, much less solving, the problem of preference reversals. All of the proposals to explain akratic 'impulsiveness' generated from within utility theory have significant shortcomings. For example, lack of experience with the consequences was proposed by Herrnstein and Prelec (1992). This 'primrose path to addiction' however, fails to account for the common occurrence of former drug addicts, quite familiar with the experience of being addicted to drugs, becoming addicted again. The notion that people, or at least those with impulse problems, have 'short time horizons' (Becker & Murphy, 2008) falls short in the light of the fact that people often take precautions in advance to prevent themselves from having the option of succumbing to impulses. 'Conditioned cravings' (Loewenstein, 1996) and discoveries of the neurophysiological basis of reward (Ho, Al-Zahrani, Al-Ruwaitea, Bradshaw, & Szabadi, 1998) can't account for why some appetites are 'craved' and not others, since all appetites have conditioned aspects, nor why or how some people come to be able to avoid the pathologies in the reward-processes discovered by neurophysiology.

1.1 Ainslie's Model

1.1.1 Hyperbolic discounting

Ainslie's model of the will proceeds from one robust and apparently ubiquitous (in vertebrates) empirical fact; that rewards are discounted not exponentially, but hyperbolically.[27] This is to say that the 'value', or rather behavioral preference, for a reward is described by a curve that is hyperbolic, a more 'bowed' curve than an exponential curve. For our purposes, the essential feature of hyperbolic discounting is that, unlike exponential curves, it implies preference reversal over time. That is, if at a significant delay a subject is presented with a choice between a sooner, smaller (SS) reward and a larger, later (LL) reward, the subject will choose the larger reward, but switch preferences as the SS reward becomes imminent. Such preference-switching as a function of time has been widely found in humans and other animals, for a wide variety of kinds of rewards. Clearly, such a pattern of behavior is quintessentially 'self-defeating'.[28] Such a person, without some way of getting out of such a pattern, would consistently undermine all their own attempts to diet, save money, study, get enough sleep, or anything beyond the shortest-term rewards.

1.1.2 Multiple Interests, Multiple Selves

From this perspective, the interesting question switches from the traditional one of how is weakness of will (akrasia) possible, to how do we ever avoid it. Ainslie's answer begins by noting what, in some respect at least, *seems* a straightforward consequence of the fact of hyperbolic discounting; we are not unitary selves. Or at least insofar as our preferences are concerned, we are not the consistently valuing selves that would be implied by exponential discount curves. Our preferences are contradictory to one another

[27] I won't discuss this literature here. Many such studies over the past few decades are cited in Ainslie (2005, p. 636).

[28] Things are not quite as simple as this, as we'll see.

at different times, and therefore if one of them, an LL one, say, is to prevail against another, steps must be taken to ensure that an incompatible SS reward doesn't undermine it as the latter draws near. If at noon we have an intention not to drink at the party tonight (because of an LL reward involving getting a good grade on a test tomorrow, or because drinking tends to give us a headache), but when we get to the party, our preference changes to drinking, our 'not-drinking interest' will have to find a way in the future to keep itself from being undone in a similar manner.

The point of naming an 'interest' is to refer to a process selected for by a specific kind of reward. Further, in order for it to be useful to name an interest, it must be in opposition to some other interest which can dominate at a prior or subsequent time. They are analogous to interests within a society, such as the 'arts interest' or 'petroleum interest' (637). Thus one would not name a pizza interest vs. a lasagne interest, even though they might be mutually exclusive alternatives for dinner. The theoretical role of interests within the model is to simply to name opposing sources of reward, such as 'sobriety' vs. 'drinking'; or 'studying' vs. 'playing video games.' The interests are at *different time ranges*, and in order for your longer-term sobriety interest to be rewarded, it must forestall your shorter-term interest in drinking, and your drinking interest must subvert your sobriety interest to increase the reward in *its* time-range. It's important to be clear that these interests, which are essentially goals, are not considered to be transcended by an ego or independent self that can choose among them. The model is 'deterministic' in the sense that '[w]hichever faction promises the most discounted reward at a given moment gets to decide [your] move at that moment' (637). But, as suggested, strategizing is crucial. The postulated interests have access to the linguistic, confabulatory, ‘rationalizing’ apparatus. Before addressing the counter-intuitiveness of the idea that we are engaged in such intrapersonal bargaining, let’s address some basic features of the economics of an internal marketplace of strategizing interests.[29]

1.1.3 Bundling Decisions and Personal Rules

People who set out to constrain their own behavior, even being willing to pay to have their own freedom restricted, have been a puzzle for utility theorists. A unitary reward-maximizer should have no need for such seemingly irrational behavior. However, if preferences naturally shift as a function of time, then LL interests will have to find strategies to receive the reward on which they are based (639). Some familiar tactics include: 1) using external (to your own mind) means of restraining yourself, such as diet pills, disulfuram (a drug that makes alcohol nauseating), illiquid investments (Laibson, 1997), and other people; 2) directing attention away from temptations and 3) inhibiting or promoting emotions. Even animals will sometimes commit themselves in advance for rewards or punishments. A fascinating study showed that rats will often commit themselves to getting .5 seconds of shock in 40 seconds rather than 5 seconds of shock in 45 seconds. If they fail to commit themselves in this way, they almost never choose .5 seconds of shock *now* rather than 5 seconds of shock 5 seconds later (Deluty et al. 1983).

[29] In the interest of brevity, I regrettably bypass an interesting discussion of involuntary behaviors here, as well as the case of how pain can be mediated by reward. It is important however to note that ‘reward’ is to be understood as ‘that which increases the likelihood that the processes it follows will recur’ (639). Ainslie’s ambition to unite the determinants of action along a common dimension has neurophysiological support; see Shizgal and Conover (1996).

The fourth and most interesting tactic for self-commitment is our topic, willpower. What this consists in has always been difficult to explain or describe. We just make up our mind to do something, or make a resolution, or employ resolve, but in light of what we've learned about the fundamental, biological hyperbolic discount curves we share with the other vertebrates, how exactly do we avoid the predicted preference-switches when we *are* able to avoid them? Ainslie cites the philosophers Aristotle (*Nicomachean Ethics 1174a*; Aristotle 1984, pp. 24-28) and Kant (1793/1960), as well as psychologists Sully (1884, p. 663), Heyman (1996) and Rachlin (1995) as all contending that decisions will be less akratic if one acts according to *principles*. The important aspect of acting in accordance with principles is that decisions are thought of as 'overall' or 'molar' instead of 'local' or 'molecular' (Heyman and Rachlin); according to 'universals' instead of 'particulars' (Aristotle). In other words, one should conceive of one's individual actions as belonging to a class of actions of a similar sort, rather than as isolated from one another, to be decided upon case by case.

Ainslie combines this common wisdom with the empirical discovery of hyperbolic discounting and sees that such advice makes sense if we represent our individual decisions as 'test-cases'. If we see a particular action as a predictor of how we are likely to act in similar circumstances in the future, then that group of decisions becomes effectively *bundled*, and consequently worth more than the single action on its own. Crucially, if we take a *series* of several pairs of SS and LL rewards, such that each individual pair would represent a choice for the LL reward at a distance, but a switch to the SS reward as it becomes near, and combine (mathematically add) them, we can achieve a situation in which the LL reward is preferred to the SS, even at the moment the SS is available.

There is experimental evidence in animals that series of hyperbolically discounted rewards in animals add together in this way (Mazur, 1997), and research with both people and rats have demonstrated that bundling choices together has the predicted effect of inhibiting impulsive, or 'akratic' behavior (640). However, this ability is postulated to be quite rare and only evokable for nonhuman animals under artificial circumstances (such as when the experimenter makes the rewards bundled *in fact*). Humans are likely the only animal with enough perceptiveness and/or foresight to construe individual decisions as relevantly similar to others they have made in the past and will likely make in the future to imaginatively bundle them in the required way.

Bundling decisions in this way Ainslie calls making a 'personal rule' with respect to that sort of decision. A personal rule treats an individual decision as a precedent for future decisions of a relevantly similar sort. Michael Bratman's (1999) example of a pianist illustrates the basic phenomenon. At a distance he chooses not to drink wine before his performances on separate nights, but when the performances come near, switches his preference to wine. However, it might well be, and probably is the case that the pianist, even at the time the SS reward is due, would prefer not drinking to drinking if that decision were applied to *all* future nights. That is the sort of result found in the above-mentioned research on humans and other animals. So if Bratman's pianist sees his drinking as a precedent, then if he drinks anyway, he suffers a greater cost than the negative effects of his drinking on tonight's performance. His violation of his rule weakens the rule's power to regulate his behavior. We now turn to a bit more detailed and complicated investigation of how such personal rules are hypothesized to play out in life, and see how ‘bargaining’ between interests—for all its initial counter-intuitiveness—entails a considerable variety of familiar phenomena.

Acting according to a ‘personal rule’ is much like, or the same thing as, acting according to principle, what the pioneering developmental child psychologist Lawrence Kohlberg considered the highest form of moral reasoning (Kohlberg, 1963). Ainslie also compares it to Kant’s categorical imperative. With one’s decisions bundled together by means of such principles, the discount curves they form begin to look like the ‘rational’ exponential ones. Unfortunately, rationality is not so easily had as that. For there are indefinitely many potential personal rules, or principles, according to which one may choose.[30] Taking Ainslie’s example, eating ice cream may violate one diet but not another, and even if it violates them all, there are bound to be extenuating circumstances, such as holidays, special occasions, consideration for the person who has already bought you the cone, or its being especially hot today, etc. The crucial point is that if you can in fact eat the ice cream ‘just this once,’ and not interpret it as a violation of your rule (and therefore not weaken it), then you can have your rule and eat it (the ice cream) too.

And there’s the rub. For sticking to an extremely strict rule is not in fact the best way to maximize reward. A good eye for exceptions to the rule will enrich your life, for

[30] Though Ainslie cites Kant and compares bundling to the categorical imperative, he doesn’t mention that the indeterminacy of the principle upon which one is acting in his hypothesized bundling and the indeterminacy of the ‘maxim’ upon which one is acting when deciding on an action are in striking parallel. Critics of Kant have maintained that (among other problems) the categorical imperative can’t provide any substantive guidance, because there is no clear answer to the question what maxim underlies one’s action when, say, cheating on a test. Kant wants it to be something like, “I will always cheat on tests,’ but it is far from clear why it couldn’t be something like (assuming it had to be a principle at all of course), “I will always cheat on tests just in case I wasn’t feeling well last night” or any other qualifying addition. Our ability (or rather, our inability not) to qualify and interpret our maxims, or personal rules, in just this way gives rise to the complications discussed below.

you’ll get to enjoy the rewards that come from indulging yourself in those exceptions, and so long as you don’t perceive it as a violation of your rule, your willpower in that area doesn’t suffer. But of course, the more inclined you are to indulge in general, where self-control is difficult, the more you’ll be tempted to interpret indulgences as exceptions to the rule, in which case the rule will tend toward uselessness. And if you catch yourself in the act of rationalization, the willpower staked on the rule is weakened, for failure predicts future failure.

We can already see that on this model, there will often be reasons and motivations to interpret your (potential) behavior in conflicting ways. Neither interpretation would be the result of ‘dispassionate reason’, but the result of conflicting interests. Ainslie emphasizes that this model is not ‘top-down,’ but ‘bottom-up,’ meaning that rather than there being an ego or executive organ of some kind that can adjudicate between desires or interests and simply will whichever one of them it chooses, the agent’s decision is the result of competing interests in a common marketplace. Were she an exponential discounter, she would have no need of bundling, acting on personal rules (or principle), or predicting her future decisions, but only ‘strength’ of will to carry out intentions. But if Ainslie is right, stabilization occurs as it does in markets with participants who interact repeatedly.

1.1.4 Intertemporal Bargaining and Bright Lines[31]

The method of self-prediction postulated here relies on the interpretation ('taking the pulse') of past behaviors in terms of their conformity with or violation of a personal rule in order to predict what one will do in a relevantly similar situation in the future. It suggests that the will within an individual is similar to the will of nations not to use poison gas or nuclear weapons. Both are situations in which agents or interests have some goals in common and some incompatible ones.

This situation is usefully modeled as a repeated prisoner's dilemma (RPD). In the classic, one-shot PD, we are to imagine two criminals being interrogated separately by a police officer, such that the criminals cannot commnicate or influence one another. Briefly, each criminal gets off scot-free if she confesses but the other one doesn't, in which case the one who doesn't confess gets 5 years in prison. If both confess, both get 2 years, and if neither confesses, both get 90 days. The crucial fact is that no matter what one does, if taken in isolation from the other, the rational thing to do for each is to confess. For if (let's call her) A confesses, then B should confess, since not confessing would mean 5 years for B rather than 2 years. And if A doesn't confess, then B should still confess, since that way she'll get off entirely rather than do 90 days. And of course the situation is symmetrical for A.

However, in repeated prisoner's dilemmas (RPDs), the situation is different. Nations rarely meet on the field of battle only once, husbands and wives are repeatedly faced with conflicting desires, and an individual's interests (between ice cream and

[31] The following section is taken from Ainslie (2001), pp. 91 - 104.

dieting, say) will be repeatedly pit against one another. However, the situation within an individual is different in one seemingly important respect, namely, that a subsequent motivational state, or frame of mind, cannot literally retaliate against a prior one (as Bratman (1999, pp. 45-50) and Elster (1989b, pp. 201-2) point out) Nevertheless, let's say a current self who prefers ice cream *this once* to maintaining her diet *right now* still prefers being healthy in the long run to always eating ice cream and being unhealthy in the long run. Future selves are going to have to obtain this desired outcome by not eating ice cream in similar situations in the future. So by setting a precedent of ice-cream-eating, she threatens to weaken the power of the rule, since ice-cream-eating predicts more ice-cream-eating.

Since I think this is a conceptually difficult framework, especially to those of us used to thinking in terms of a governing self or ego, rather than an ungoverned marketplace of interests, I'll give what I take to be a helpful analogy. Suppose you are in a lecture hall of 100 people, and each person has the choice of getting one dime along with everyone else, or one dollar only for themselves.[32] People choose publicly one at a time until an unknown number of people, but no fewer than 20, have chosen. Each person is to attempt only to maximize earnings. This has the structure of successive motivational states within the person, as opposed to typical RPDs, because in the case of the latter, the rationality of cooperation is based on the fact that the same players will play again. When it's my turn to 'play,' I know that I will have another turn after you do,

[32] The usage of what appears to be the 3rd-person plural 'themselves' is deliberate. It is in this context singular, according to common usage, and certainly my usage. Please regard other instances likewise.

and that my choice now will likely affect your choice, and so on. So long as it is clear that we do better both cooperating than both defecting, it makes sense to cooperate.

In the game described above, each person only plays once, so they cannot have the concern that subsequent players will attempt to punish them according to the same logic as applied above. It would make no sense for B to defect after A defected in the hopes that that would teach A a lesson, for A won't get to go again (remember, we're only trying to maximize profits, not teach lessons *per se)*. However, there remains an incentive to cooperate nonetheless. So long as one thinks that there are a significant number of turns remaining (more than 10), one has an incentive to accept a dime in the hopes of establishing or continuing a pattern of behavior. If everybody accepts dimes, then everyone will make a minimum of 2 dollars, up to 10 dollars. If everybody defects, then they each make a dollar. Of course, the optimum situation is that everybody cooperate but you.

But how could that happen? Suppose you're a man and the whole room is otherwise full of women (or vice versa). Or you're of a different religion (and this is publicly known), or you're a child in a room of adults, and so on. If there's something that stands out about you that could plausibly allow you to defect without initiating a chain-reaction of defections, then that might work out to your advantage. But if not, then although you will not play again, you setting off such a chain acts just like retaliation as far as your prospects go. Your task is to *predict* what subsequent players are likely to do given what you do. Your continuation of a tacit rule makes it more likely that the rule will hold up and you will thereby benefit, while a defection has an even stronger chance of reversing that rule (causing defections), especially as there become fewer possible remaining plays.[33]

In response to a defection (or more than one), the room may continue to adhere to the rule ('trust itself') or not, presumably at least largely as a function of whether the individuals in the room can perceive the defections as exceptions. The more exceptions that seem to be permitted, the more defections will be likely to happen, in the hope that oneself will also be perceived as an exception. If perceived instead as a violation, subsequent players will defect because of a 'loss of will,' as opposed to the first case, in which the room ends up with less money without ever perceiving itself as having lost 'willpower', but only due to their being so many excusing or extenuating circumstances.

Of course, in the context of intertemporal RPDs, an individual person is analogous to the whole room just described. The total amount of money of all the people in the room is like the total reward in an individual person. That is why a strict strategy of adherence (cooperation) is unstable, because although that strategy is superior to always defecting, it is inferior to defecting just in those cases in which there is 'good reason' to violate the principle, or where the principle doesn't apply.[34] At stake in such

[33] I realize that it might still not be clear in what sense these different selves are supposed to persist so as to be effectively retaliated against. I will just note two things, rather than discuss this in detail. (1) So long as the current self can be meaningfully and accurately said to *have the preferences* mentioned (never ice cream and healthy rather than always ice cream and unhealthy) and is a 'player in the game' that current self will have the rough incentive structure modeled in the lecture hall example, whether that self, or the interest it represents, continues to exist in some sense. (2) We should at any rate not concern ourselves too much at this point with whether to interpret these selves and successive motivational states realistically, instrumentally, empirically or otherwise, prior to an evaluation of how well the model serves the explanatory, predictive, and perhaps even therapeutic ends to which it is put (and even those to which it was not put, for as we will see it explains phenomena that were not regarded as related to the will). For what it's worth, when Ainslie has done this exercise in his classes, students report using roughly the same reasoning process described above (2001, 93, footnote 5).

[34] Suddenly I think of Thomas Jefferson, who was one of the most outspoken and staunch adversaries of the use and expansion of federal power, especially involving the Treasury. However, the Louisiana Purchase was simply too 'rewarding' to pass up. One can imagine him, if he were a Kantian, successfully

interactions between interests is your own credibility; the total reward that you stand to gain by cooperating is gambled against whether an exception to the rule is interpreted as defection, and how much damage will be done as a result.

As one might be able to see at this point, there is great importance in one's *interpretation* of what counts as a violation and what doesn't. As might be expected then, *bright lines* are an important feature of willpower. Where there are powerful motivations to act counter to a personal rule (such as in dieting or alcoholism or smoking), and yet hope that the action will be treated as an exception rather than a breach, there is benefit in having lines which cannot be fudged in order to stabilize conduct. Rules such as smoking 'only a few' cigarettes a day or week, or only drinking a little, or until one has a buzz, are too open to interpretation, and therefore either becoming less useful as rules or risking being interpreted as violations that result in the weakening of the force of the rule.

Lines such as no drinking or smoking at all cannot be fudged. It is unavoidably apparent when one of these lines has been crossed.[35] On the other hand, having less-bright lines, in areas where threats by short-term interests are not so strong (such as someone for whom alcohol or smoking is not as rewarding, or whose will still has plenty of credibility), provides for greater flexibility, as one takes advantage of good exceptions to the not-so-brightly-dileneated rule.

interpreting the maxim on which he acted as something like, "Do not use the Treasury as a means of pursuing federal policy unless it is an impossibly good deal.' He could plausibly have remained committed to the general rule, given the *exceptional* nature of that purchase.

[35] Interestingly, Ainslie notes that there is not the easy availability of such bright lines in dieting, which may be why long-term recovery rates of alcoholism are much higher than 'medically significant' weight-loss (Campfield et. al., 1998).

It's worth pointing out that at this point, the entire model so far has been generated from nothing but the empirically well-founded assumption of hyperbolic discounting, in addition to the rough additiveness of summed curves (also with empirical support). With these assumptions, Ainslie has been able to generate a model that 'predicts credible weapons for each side in the closely-fought contests that occur as people decide about self-control. Long-range interests define principles, and short-range interests find exceptions' (2005, p. 642).

1.2 Some Implications of the Model in the Moral Domain

There is still quite a bit to be said before you can be expected to take this model as seriously as I'd like you to. First, there is the problem that such a model still seems counterintuitive, not the least reason for which is that it doesn't seem to jibe with our subjective impressions of exercising willpower. Since I need to convince you that this model is at least plausible, I'll need to address that fact. I'll be restricting myself, however, to only one of Ainslie's three responses to the question of why it doesn't seem as if we're bargaining with ourselves. That is because this answer is an important component of the model as concerns my wider project--the role of normative belief in stabilizing preferences and/or organizing motivation. That is the subject of section 1.2.1.

Section 1.2.2 will address the unrecognized downsides of willpower, as they are understood on this model. Section 1.2.3 employs the framework we've been describing to analyze and explain four 'puzzles' of the will that arise in the context of premature satiation of appetite. I intend for all three of these sections to contribute to the following things. First, I want them to make Ainslie's model of the will seem (more) plausible and fruitful. This is not to say that it should be thought of as the correct and complete model of the will, but rather that it provides us with some powerful tools to help us understand its workings. Second, I want to show that these tools, almost completely absent in moral psychology, have the potential to address questions there, most obviously those connected with moral motivation. Third, insofar as the model can address questions of, e.g., moral motivation, it will hold out the hope of dissolving (or reducing the sharpness of) the distinction between 'moral' and 'prudential' motivations by addressing them in a unifying framework.

1.2.1 Moral Belief as an Experience of Intertemporal Bargaining[36]

The most obvious and perhaps most important objection to the intertemporal bargaining model of the will is that it doesn't seem as if we are bargaining with ourselves in the way it suggests. First of all, most of us almost certainly feel as if it is a cognitive, reasoning process that at least sometimes, even often, determines our actions. We certainly don't make explicit bargains between interests that we perceive to be independent and partially conflicting. Only in cases of explicit resolutions does it seem that something akin to what Ainslie describes would be happening.

Explicitness, however, shouldn't be necessary. To analogize again to actual societies, there were bargaining and rule-making processes going on far before anyone had explicit awareness of them, much less a theory of how they worked. Before the

[36] Taken from pp. 105-112 in (2001).

Renaissance, laws were considered to be sacred mandates instead of pragmatic rules; one discovered ancient laws instead of making them, or so they represented it to themselves. Piaget reports that young children believe that rules must be unchangeable.[37] It shouldn't be hard to believe (or see) that people today often perceive the rules that they use to conduct their behavior as given by some authority outside themselves, a feature of the world that is independent of their own wills.

Crucially, we would expect from the nature of the intertemporal bargaining process described above that rules that are *conceived of* as our own 'expedient' solutions to practical problems are more vulnerable to being hedged when the heat gets turned up. In the presence of strong contrary motivations, if we feel that the rules are in our service, then we are free to change them if necessary, which in many situations will amount to interpreting them in ways favorable to SS rewards. If, on the other hand, they appear not as rules but as *beliefs* about the external world, i.e., as something that is true or false independently of our motivations, then the stability provided by such a belief would be staked on the belief itself. That is, to violate the rule that we perceive as a belief in the desire-independent goodness or rightness of some action is to threaten the belief itself.[38]

So, could projecting value onto objects or states of affairs be a tacit personal rule? The way a personal rule works is that the benefits of the rule are at stake each time you choose whether to follow it or not. There is no apparent need for you to conceive of the personal rule as such, and we saw why such a rule could provide added stability were it

[37] Kern (1948, p. 151); Palmer 1993, pp. 254-257; Piaget (1932/1965), cited on p. 217.

[38] By hypothesis, these beliefs will often be religious, mythical and/or moral. These beliefs, in addition to their hypothesized function here of organizing motivation, play important roles in establishing identity, membership in communities, conceptions of 'the good,' and often nothing short of entire world-views. Threatening such beliefs is no trifle.

not perceived as subject to your interference. Ainslie suggests that ‘praying to Saint X’ for the strength to do x is to in effect create a ‘dummy stakeholder’ who can be pleased or affronted. If you succeed in overcoming temptation, your belief in Saint X and her power and willingness to help you is strengthened, which will make her help more effective in the future. Contrariwise, if you fail despite Saint X’s help, you will offend X and/or discontinue believing in her help, which will manifest itself in the greater difficulty you have in overcoming temptation next time.

Ainslie provides other examples of beliefs-as-personal-rules, such as the belief that street drugs are always addictive. You learn this ‘fact’ as a grade-schooler, and when you encounter contradictory information, you might discount or ignore it, not because it seems objectively flawed information, but ‘out of a feeling that it’s seditious’ (2001, 109). Decriminalizing drugs in the United States has been a nonstarter politically, despite abundant evidence that marijuana is no more, and likely less harmful overall than alcohol or cigarettes, in addition to the colossal financial and human costs of the ‘drug war’. Despite the enormous potential benefits of reorganizing our policy toward drug-users, it is interesting to note that there can be almost no public discussion of the matter. Allowing liberally for human irrationality, it seems difficult to believe that if most people’s beliefs about drugs were in the business of attempting to ‘track the truth,’ that this would be the result. However, if those beliefs were (whether they started out that way or not) in the business of committing people to courses of action, then one would expect arguments in favor of decriminalization to be met not with counterarguments and evidence, but with accusations of being ‘soft on drugs’ (109).

Even those of us who are prone neither to believe in saints or gods,[39] nor to silly political rhetoric, are very much in the habit of perceiving situations or actions as in themselves good or bad. Certain states of affairs are 'to-be-promoted,' whether or not we have any desire or interest to do so. Perhaps not coincidentally, paradigmatic instances of actions which are believed (morally and/or intrinsically) bad are just the kinds of actions that 'we' (broadly construed) are frequently tempted by, such as lying, stealing, breaking promises, harming others, and the like. Like the laws prior to Renaissance times, we do not consider these moral rules (insofar as we consider them moral rules) to be solutions to bargaining problems, either within ourselves or society at large, ultimately dependent for their worth on their usefulness (broadly construed) to us.

Perhaps the single best example of this phenomenon is the issue of the morality of abortion. Debate continues over whether the fetus is or isn't a person, assuming not only that there is a fact of the matter but that that fact will settle the question of abortion's 'permissibility'. Even in Judith Jarvis Thomson's famous (1971) article in which she granted the fetus personhood for the sake of argument and then argued that abortion is yet 'permissible' in some cases, there is not the remotest hint that the legitimacy of the prohibition of abortion should in any way turn on that rule's pragmatic or other benefits

[39] Or any of the indefinitely many 'New Age' beliefs that seem to rival even the barest theistic notions of God for their indeterminacy. One of the most popular is the belief that 'the universe conspires to help you' when you pursue your dreams, from the internationally best-selling book *The Alchemist* by Paulo Coehlo. The readiness with which intelligent and educated people believe such things strongly suggests (at minimum), especially in light of the present model of the will, that such beliefs are purely in the service of helping people overcome, in this case, their fear of 'pursuing their dreams'. The motivation to pursue your dreams is staked on this belief. The more you succeed, the easier it is to believe this and the easier motivation becomes. Perceiving that you are not pursuing your dreams suggests that you don't really believe that the universe will conspire to help you if you do pursue them, which recursively makes it that much harder to pursue them.

to us.[40] Thomson 'merely' points out that our moral belief that killing innocent people is wrong does not apply absolutely; there are imaginable cases where it seems permissible, including some cases of abortion.

The belief that killing innocent people is wrong, however, was (and is) ostensibly so powerfully motivating that despite the apparent absurdity of attempting to nail down a fact of the matter as to when (or if) a fetus turns into a person within the womb, just this question was and still is debated. For, so *both camps* thought, and most still think, that (our belief in) the wrongness of killing innocent people is neither of our making nor subject to our will. On the contrary, our will is, or should be, *subject to it.* All Thomson did was try to convince us that this belief did not apply in every instance; there was no question that where the belief *does* apply that we are *required to act* in accordance with it. 'The less someone's belief seems accountable for by the objective facts, the more it's apt to be the representation of an underlying personal rule' (2001, p. 110).

Even for those who believe in objective moral facts, it appears very unlikely that *most* moral beliefs are best explained by reference to those objective facts. This appears to follow from the extent of moral disagreement both within and across times and cultures.[41]Insofar as it is a feature of paradigmatically moral beliefs that the requirement

[40] Glover (1985) is the only one I know of to do this. He concludes that abortion would not psychologically harm 'us' enough to outweigh the costs of criminalizing it, but infanticide would. However, it is crucial to note that Glover only presents this kind of consideration (psychologically contingent benefits to us, as opposed to categorical rules against killing babies) because he rejects the claim that either fetuses or infants are persons. They are therefore not owed the moral considerations we owe to persons, and so facts about our own psychological health can decide the matter.

[41] I realize that the issue is more complicated than I seem to suggest. For example, some maintain that the extent of fundamental moral disagreement across cultures is not nearly so great as often supposed. Of course this could be true and it still be the case that it is rare at best that moral beliefs are best explained by reference to moral facts. Since this is not the place to wade into these waters, I will be content to let the point be taken by those who do believe that *in general*, moral beliefs are not explained by reference to

to act in accordance with them does not depend on one's contingent motivations, then the subjective experience of firm belief in the truth of a moral principle has just the features we might expect from particularly important personal (and social) rules.[42] The especially important rules, for the reasons illustrated in the hypothesized workings of the will, should be ones that are conceived of as beyond our power to meddle.[43] By 'especially important' I mean (roughly) those rules that have the highest costs associated with their being violated, in the context of strong and/or frequent motivations to violate them. Thus we have an account of why we would be motivated to think that we *passively discern* value in objects and actions rather than that we *actively value* them. According to Ainslie's model, this is the result of a strategy of regarding rules as beliefs in order to make them more stable.[44]

I want to head off one objection, or perhaps only a misunderstanding, before discussing the downsides of the will. It's important to be clear here that I am not suggesting that people come to perceive these rules as beliefs (exclusively) by

objective moral facts, even if there are such facts, and even if some people's moral beliefs *are* best explained in terms of their contact (in one way or another) with those facts.

[42] Notice that I do not say 'insofar as it is an essential conceptual component (or commitment) of moral belief'. A conceptual antirationalist should be able to agree that paradigmatic moral beliefs are not typically *thought to be* 'legitimate requirements' only if those subject to them have the relevant contingent motivations. In other words, an antirationalist can recognize that most people are not antirationalists (even if they are wrong not to be).

[43] This is just the benefit of categorical imperatives that Joyce (2001) invokes in arguing that we should consider fictionalism. He cites the benefits of such imperatives in combatting weakness of will, but doesn't describe the means by which such rules are supposed to work, nor any of their downsides.

[44] It's important to be clear here that I am not suggesting that people come to perceive these rules as beliefs (exclusively) by individually strategizing in the ways described above. Rather, what I want to take away here is that this model of the will suggests that perceiving rules as beliefs of this kind has a stabilizing effect on the motivation to act in accordance with them. But rather than individuals each coming upon this strategy individually, this fact about our motivational systems helps to explain why cultures would come to represent rules as beliefs in this way, and teach children accordingly. See the sketch of the acculturation of children in Orissa in Chapter 2 for an illustration of how rules of behavior are perceived not (merely) as such, but as proceeding from beliefs about what is (im)pure. As I will argue in Chapter 2, rules that are 'important' in this sense needn't be understood in terms of maximizing reward or utility in an individual.

individually strategizing in the ways described above. Rather, what I want the reader to take away from the discussion of ‘beliefs as personal rules’ is the point that perceiving rules as beliefs about the world (external to our will) has a stabilizing effect on the motivation to act in accordance with them. But rather than individuals each coming upon this strategy individually, this feature of our motivational systems helps to explain why there would be cultural and/or biological evolutionary pressures for us to represent rules as beliefs in this way.[45]

Chapter 2 elaborates a commitment model of moral judgment, which is bolstered but not dependent on such evolutionary arguments. The authors whose arguments I draw upon in developing my commitment model rely on the core of Ainslie’s model, though without appreciating the full contribution that Ainsliean psychology has to make. The particular contribution I am focusing on now is why it would be the case that perceiving values (or ‘rules’) as being ‘in the world’, independent of our own contingent motivations, could have significant motivational consequences. Part of the commitment model will make use of arguments for projectivism. Other authors have relied on intuition to support the idea that ‘projecting our values’ onto the world would make for greater stability of motivation than representing them as ‘merely’ our values. I am here showing that not only is this intuitive (to whomever it is), but it is also a natural extension of a well-supported model of the will. And it is worth keeping in mind that the model was extended in this way so as to account for how it is that an intertemporal bargaining model could be correct, though we do not typically experience ourselves as bargaining

[45] See the sketch of the acculturation of children in Orissa in Chapter 2 for an illustration of how rules of behavior are perceived not (merely) as such, but as proceeding from beliefs about what is (im)pure.

with ourselves—and not with any metaethical axe to grind.

Finally, rules that are 'important' in the sense defined above needn't be understood in terms of maximizing reward or utility in an individual. Though there are serious downsides to such rules, as we are about to see, if all the beliefs about intrinsic value were geared toward such ends, I would be much less anxious to attempt to undermine belief in them. Rather, we have inherited a variety of commitments from myriad sources which are stabilized, at least in part, by their being perceived/conceived of as independent of our wills. Many of these commitments have little if anything to do with making our lives go better on the whole (as opposed, perhaps, to spreading our genes or 'memes'). I want us to reexamine some of our commitments, and in order to do so I think it is important to recognize that they are *our commitments*, the fundamentally practical/motivational nature of which is disguised by their being perceived of as independent of those very practices and motivations.[46]

I now want to briefly discuss the very important downsides of the will, which can be serious enough at the intrapersonal level, but in the socio-moral domain have the potential for immeasurable pathology and catastrophe.

[46] To be clear, the role of this part of the descriptive psychology is to provide an explanation of why it is that we would be motivated to have the kind of beliefs I am discussing, whether or not they are philosophically defensible. This does not not show that they are not philosophically defensible. However, I think the attempts heretofore to defend conceptions of value (or practical reason) that are independent of our motivations are unsuccessful. I argue against all such conceptions in Chapter 3.

1.2.2 Downsides of the Will

These downsides are important for at least two reasons. First, I hope that the downsides that follow naturally from Ainslie's model of the will help to make the model more convincing. They add significantly to the phenomena explained and unified by it, connecting them as phenomena related to the will.[47] Second, these downsides have even greater potential for pathology in the socio-moral domain. Once we've seen how they follow naturally from Ainslie's model of the (intrapersonal) will, we will be in a better position to analyze them, as well as potentially ameliorate them, when they arise in interpersonal and intergroup contexts. These will be presented more briefly than what has come before in the interest of moving on with the substantive arguments.

Rules Overshadow Goods-in-Themselves[48]

Above, I mentioned the importance of bright lines, as these are not subject to interpretation in favor of your short-term interest as its reward draws near. However, this process of drawing lines and becoming disciplined in the rules that they define can make the rewards one is attempting to maximize elusive in some cases. The essential problem is that there is a corollary to the 'fact' that every lapse from a personal rule reduces your ability to follow it; every adherence reduces your ability not to. The more you see your choices as precedents for future choices, the more legalistic your decision-making

[47] Specifically, the explanation of compulsions as 'overly-successful' strategies against impulses is a novel theoretical explanation and connection of these two phenomena (Cf. Elster 1986, 22).
[48] Downsides drawn largely from Ainslie (2005) , pp. 644-649.

becomes. You lose flexibility as your perception of the importance of following the rule overshadows the good(s) that it was intended to secure. It bears noting that this phenomenon, something like which is certainly familiar, makes no apparent sense on the assumption that we are essentially exponential discounters, always in principle free to maximize our utility at every choice-point.

Rules Magnify Lapses

When you perceive yourself as having had a lapse, you see your own will as being weak, which, due to its recursive nature, weakens it further. This fact motivates you to find interpretations such that what happened was not really a lapse, but an acceptable exception to the rule. In order for it to be an acceptable exception, there must be something about the situation that marks it off from where the rule does apply. Features of the situation that become identified as exception-making then can have the effect of salvaging the general rule (and the will it organizes), but at the expense of making similar situations in the future appear as those in which will is either weak or inapplicable. These 'lapse districts' (like the Victorian vice districts where sin that couldn't be suppressed was instead enclosed) see failure predicting failure, and within them one feels incapable of even attempting to muster an effort of will. Cognitively, these urges or addictions may be represented as irrational, immoral or otherwise bad, but nevertheless irresistible. To your short-term interests, this is of course a welcome interpretation (2001, 49). It's also worth noting that neither opponent-process nor models of will based on overall strength predict or explain how will can be very ‘weak’ in one area while strong in others.

Rules Motivate Misperception

By now it should be clear why personal rules would provide motivation to mis- or not perceive your own motivations on occasion. On Ainslie’s account, great amounts of willpower are often organized on the basis of personal rules. Violation of such a rule, *if perceived as such*, threatens the ability of that rule to organize motivation in the future. Therefore, not only is it in your short-term interest for you to fail to perceive a lapse, so that it won’t be stopped by your long-range interest, but it is also at least somewhat in your long-range interest insofar as perceiving the lapse threatens the ability of this rule to organize motivation in the future.

Ainslie’s apt analogy here is that your ‘long-range interest is in the awkward position of a country that has threatened to go to war in a particular circumstance that has then occurred’ (645). The interest needs to maintain its credibility without having to go to war, so if it can manage to appear not to have noticed the breach, this may be accomplished. If you catch yourself ignoring a lapse, of course, this will damage your credibility with yourself, so there is motivation not to catch yourself, and so on. These processes of ignoring are supposed to arise ‘by trial and error’ (645), allowing you to feel fine without your knowing why. Potential examples are ending up broke at the end of every month despite a 'strict budget' and gaining weight despite an equally strict diet.[49]

Rules Serve Compulsions

There are many possible ways of bundling choices and of (not necessarily consciously) perceiving prisoner's dilemmas. We may find ourselves settled into patterns of bundling that are not well-suited to our longest-range interests. First, it is important to note that personal rules work best when operating on salient, especially countable goals. Amounts of money are a perfect example, and this goes toward explaining why people are most capable of approaching the 'rational' exponential discount curves assumed by utility theory when dealing with money.[50] Therefore, when some personal rules are marked by such salient criteria, and other, longer-range, richer rewards are not so easily demarcated and identifiable, the shorter-term interest may rule the day due sheerly to its well-markedness. This is even more likely when this 'compulsion-range' personal rule is set against a significant impulse-range interest, like saving money against prodigality.

[49] I am sure I've spent entire days doing 'nothing but working' but somehow getting very little done. I've learned to catch myself ignoring myself putting around on the web, or thinking about other things, which has helped. But I must note the feeling of accuracy Ainslie's description had in these cases versus ones I interpreted as 'akratic.' I would feel much better having 'worked all day' and only gotten a small amount done than if I knew I had not worked as much as I ought.

[50] Subjects learn to make maximizing choices much more easily when amounts of money are varied, rather than amounts of time. In an experiment where subjects could choose between two rewards of different value, they quickly learned to pick the smaller reward when subsequent options were better as a result. However, in the same game of fixed duration, they have much more difficulty choosing the smaller reward when the delay between choices is increased, so that overall reward is lower when one picks the larger option. However, when the delay is pointed out to them, they quickly learn to switch strategies (Herrnstein et al., 1993).

The most well-known and serious of these 'disorders of over-control' are conditions such as anorexia, bulimia, or obsessive-compulsive personality disorder. However, misers, those dedicated to studying as much as possible, 'gym rats,' teetotalers and health-food nuts all increase efficiency in achieving the rewards defined by those well-marked criteria, but lose out on subtler, potentially richer rewards marked by not-so-well defined criteria. Making matters worse in these cases is that the people who indulge compulsion-range interests do not experience them as urges, but often endorse the normativity of their own strictures and consciously attempt to enforce them. Such regimentation at the cost of subtle rewards (such as emotional rewards coming from friends, family, art and the like) is surely a pervasive phenomenon, arguably exacerbated by modern society. Ainslie's model of the will diagnoses these phenomena as side-effects of willpower within an analytic framework, going beyond the metaphors of Freud's id, ego and superego.[51]

1.2.3 Puzzles of Premature Satiation

> 'The greatest limitation of the will comes from the same process as its greatest strength: its relentless systemization of experience through attention to precedent.' (2001, 164).

In a part of Ainslie's model I did not discuss, he argues that 'seduction of attention'[52] is how aversive emotions and even pain can be reward-driven processes, and

[51] Elster (1986).

[52] A name for a process in which rewards/nonreward cycles are on short enough time-scales to attract attention and sometimes motivate behaviors that are experienced as aversive.

not in need of explanation by some different principle, such as conditioning. Though interesting, that part of the model is not (as far as I can see) highly relevant to the purposes to which I intend to put Ainslie's model. However, there is an issue connected with seduction of attention that I do think is not only interesting but more likely to be philosophically relevant, especially in the normative realm. In Ainslie's model, behavior is reward-driven. This makes one wonder why it is that people will go far out of their way, are willing to give up significant resources, and often fundamentally structure their lives to achieve emotional rewards, do not provide themselves with these rewards *ad libitum*. The answer has to do with *premature satiation of appetite*, which can't be controlled by will, and can even be worsened by it. The concept of premature satiation and its consequences will provide us with some interesting (if unsettling) tools with which to address some fundamental philosophical questions.

The Limitation of Positive Emotion

The most intense emotions seem especially like automatic, necessary responses to perceived conditions in the environment. Contrarily, insofar as the occasion for them seems arbitrary, they become pale and daydream-like. But emotions, even intense ones, are not in fact dependent on external factors in the environment. Accomplished actors can work themselves into intense and positive emotional experiences. All of us can use our imaginations to generate mild or moderately pleasant emotional experiences.

Given that there is no (external) physical 'turnkey' required for emotional experinece, how do they become scarce enough to function like economic goods?[53] We can break this question into two parts: 1) Why would you want your own behavior to become scarce and 2) how can you make it scarce without making it physically unavailable? (2001, 166). Call this the puzzle of the limitation of positive emotion. The solution to it will also answer the question of why passions seem to be had passively.

The answer to the first question has to do with the *potential* for reward, or what we can call appetite. This potential is outside of our direct control. Though emotions do not depend on the external environment, they are dependent on having sufficient appetite for them, and the appetite that is used up in the reward process can't be replenished voluntarily. This is the case for all kinds of reward; they depend on a potential, or readiness, that is used up during the reward process and cannot be restored at will. We can call this readiness 'appetite'.

It is a familiar feature of appetite that quick consumption leads to an earlier peak but less overall pleasure than slower consumption. In other words, there are ways to make very poor usage of available appetite, whether in eating, having sex, reading literature or being entertained in other ways. One can expect reward-governed organisms that hyperbolically discount the future to consume freely available rewards faster than would be optimal if they were to maximize total amount of reward. Brief but (even a bit more) intense wins out over long and pleasant. Were we exponential discounters, we shouldn't have this problem, but rather be able to lie back, wait for appetite to build up,

[53] Something freely available cannot function as an economic good because one will not work for what is freely available, no matter how important it is to them to have. This is why gold is more economically valuable than air.

and then reward ourselves at what we feel to be the optimum times. However, we know from experience that doing so is difficult. We need to have or develop restricted access to our rewards in order that they be robust and deeply satisfying. Failure to restrict this access and the consequent harvesting of rewards relatively near the satiation point is what is meant by ‘premature satiation’.

Physical rewards (like food, drink, sex, drugs) are dependent on the environment, so we can limit our access to them via personal rules connected to sufficiently rare criteria if necessary. But emotional rewards need to be limited by other means. Willpower as understood here has no way to stop our minds from rushing ahead and harvesting reward as soon as some appetite is available. But such direct paths to rewards become quickly unproductive; the only way to limit your access to the reward is by making the pathway to it uncertain, variable, or if certain then such as to defy efficient acquisition. The occasions for (at least many kinds of) reward have to be *surprising* in some way, a feature that has been shown to be necessary to activate reward centers.[54]

So, activities that hope to provide emotional reward (which makes up what most of us seek most of the time) must compete for control of your behavior on the basis of their capacity to maintain your appetite, which requires that they not be subject to your control. ‘In modalities where you can mentally reward yourself, surprise is the only commodity that can be scarce’ (647). What was once a source of intense emotion becomes incapable of generating any as you become able to leap ahead imaginatively to what used to be the intense moments, and what used to occasion a sense of achievement when first accomplished no longer has any such power once one’s success is a foregone

[54] Berns et al. (2001) and Hollerman et al. (1998), cited on p. 647.

conclusion, along with the route to it. People have learned techniques to maintain surprise, and therefore potential for emotional reward, from the ‘rule’ not to read ahead in a book, to the elaborate steps taken by millions of Sopranos fans not to find out how the final episode ended after recording it while at work or away. Internet sites devoted to art or entertainment have also very quickly adopted the tool of a ‘spoiler alert’ to avoid revealing any information that would prevent ‘appetite’ from ripening as the artist had intended.

Now that we have seen the importance of the problem of premature satiation of appetite, we are in a position to use it to explain three more puzzling phenomena. All three have very significant philosophical implications if they are (roughly) correct, especially for normative issues.

Construction of Fact

There are a range of beliefs that we ‘can’t help but have,’ in that they are or would be regularly punished or corrected by observed features of the world. Our beliefs that objects fall when released, that our cars can drive on roads but not water and that we cannot fly can be called ‘instrumental’ beliefs in that they are shaped fairly directly by their practical effects.

There are other beliefs that are either not so shaped or the shaping is significantly delayed, so that we might expect them to be subject to shaping by emotional reward rather than external effects. There is plenty of scientific evidence, as well as common experience, to suggest to us that processes such as ‘wishful thinking’ occasion beliefs in ways that seem to be guided by things other than an attempt to get at the truth. I confess that I take this to be obvious to the point of truism. Taking for granted that many of our beliefs are goal-directed, motivated behaviors, the puzzle is what features distinguish these beliefs from straightforward pretense, for there is an apparently important experiential and motivational difference between the two.

In our discussion above, we were noting that our tendency toward premature satiation combined with the ability to generate emotion ad lib gave rise to the problem of how to restrict our access to emotions so as to avoid having our emotions become pale and lifeless daydreams as we rush to harvest appetite for them before it could become ripe. 'Constructed' beliefs provide a plausible answer. Beliefs have the essential representational character that their content either obtains (is true) or not, independent of the vicissitudes of our psychology. And whatever truth there might be in doxastic voluntarism, it is clear that we cannot voluntarily switch from belief to belief on a whim, or because it would make us feel good (I cannot choose to believe that I just won the NBA or Ultimate Fighting Championship, even though that would feel great).

For 'noninstrumental' beliefs to play the role hypothesized for them, i.e., to limit the occasions for emotion, there must be restrictions on what can be believed, and the most 'successful' beliefs will be those that obtain an optimal long-term balance between restriction and production of reward. The kinds of beliefs that can serve this purpose need to be: 1) Out of your direct control, 2) rare, and 3) surprising. They needn't be true, though verifiability is one of the best ways of achieving rarity (something that really happened is more emotionally impactful than a fiction, all else equal, and a fiction that someone else wrote is more so than one you invent with the same content).

A nice example provided in Ainslie's (2001) is that of watching a movie that has become too upsetting for us, at which point we disengage by saying to ourselves, 'It is only a movie,' as if we were learning a new fact. This announcement that we are *actively* disinvesting our emotions in the plot is reported as a reminder of *a fact*, though at no point did we ever fear for our lives when guns were pointed toward the camera. Rules are ubiquitously reported as facts external to one's desires or interests, whether they deal with the commandments of a deity, the requirements of morality, or the laws that were 'discovered' by the pre-Renaissance legislators.[55] Cultures communicate values by means of myths, which are reported as (something like) facts, although a community's active participation in keeping these values alive is apparent from the point of view of an outside observer. The reason for representing them as beliefs in facts independent of our wishes is that 'we protect the uniqueness in these commodities by not noticing our participation in assigning them value.'[56] One of the important aspects of my project, and one of the primary uses to which I hope to put this model of the will, is to both convince you that many of these 'facts' are in our power to alter as we see fit, and that in many ways we would be well-advised to become more aware of it.

[55] Of course, all these rules came associated with real and/or imaginary punishments, which made sure to connect up with your desires and interests in no small way. The point is that the rules were conceived as reflecting a reality that existed independently of those desires and interests. Religion, morality and (most conceptions of) the law suppose that there are reasons for conforming to the rules they prescribe that do not depend on one's desires or interests.

[56] (2001, 179). See 175 - 179 for a more thorough and convincing account of this important claim.

Vicarious Reward

Interacting with other people is a well-known source of emotional reward. They need not even be real people; fictional characters can occasion reward in us, which is seemingly their most valuable capacity. The problem of premature satiation suggests that other people should be good sources of reward, but only insofar as you cannot control them, for this reduces or eliminates the necessary surprise and uncertainty required for maintaining appetite for reward. According to the hypothesis, 'gambling' on others' behavior is a great source of potential reward so long as the gambles aren't 'rigged.' That is, to the extent that you can predict their behavior (either because they are highly predictable or that you can boss them around successfully) or that you're not really invested (you're ready to terminate the relationship if it becomes difficult), the potential for emotional reward is prone to become weak and veer in the direction of daydreams.

The other important aspect of the explanation of the power of vicarious reward is that trying to predict the other person's emotions is plausibly done by modeling those emotions in oneself, which makes the reward more intense than if we were dealing with some other sort of challenging game or puzzle. That we can and do model others' emotions is nothing new or surprising. The most important things to my mind are to note that 1) the model predicts that other people are likely to be (one of) the best potential source(s) of emotional reward and 2) that this reward is available to the extent that one is *invested* emotionally in the other people.

This requires no particular relationship, but only a 'willingness' to be affected by their perceived affective states. Like telling yourself 'It's only a movie' when you want to disinvest your emotional responses from it, you may equally tell yourself ‘It’s only a bum’ or ‘He’s just the servant,’ so as not to become invested in their emotional states. As Ainslie and Haslam argue (2003), ‘altruism[57] is a primary impulse.’ Vicarious reward is a basic source of reward in humans. Rather than having to learn altruism as a form of self-discipline, children start out as highly empathic (Harris 1987; Zahn-Waxler et al. 1992), and generally learn to *control* (to different extents of course) their ‘empathic impulses’ (Ainslie 2004, p. 730). This is another potential area where ‘self-control’ is not the unmixed blessing (to oneself or others) it is often thought to be.

1.2.3.4 The Indirection Puzzle

Let’s now recall how the will fundamentally operates on this model. It can be expected to motivate you in the future just in case it motivates you now. That is, you can expect yourself to follow personal rules just in case and/or to the extent that you have been doing so. But, as we saw in section 1.1.4, the will needs salient criteria of when a rule has been obeyed or not. This will to systemization makes your own behavior predictable, which is just the thing that, according to the premature satiation hypothesis, makes you incapable of achieving the appetite for powerful emotional rewards. Will is incapable of solving this problem due to the lack of well-defined guidelines for directing attention, even if something as quickly shifting as attention could be regulated in accordance with them in the first place.

[57] I think empathy would be a much better term than altruism.

One major obstacle to consciously attempting to slow down our reward-getting is that in order to render goals capable of generating appetite, we’ve placed value in those goods themselves, and often see our goal as maximizing them in some way. To deliberately move slowly toward them or restrict our access to them would be to contradict that very belief. Gambles are one of the principal ways to stimulate appetite for reward. Gamblers typically take themselves to be attempting to make money (which is of course true at one level), but most realize that gambling is rarely the best way to make it and in fact is one of the classic ways of losing it. Wealthy people gamble despite having no need for the money, but rather because there is the risk of losing it. What appears (and of course often is) highly irrational behavior is not always so obviously irrational when approached as a means to generating appetite.

Gambling, and other putative ‘impulses’ can thus be sometimes seen as ways of undoing some of the satisfactions that accrue to the affluent (‘everybody needs a vice’) in order to generate more appetite for reward. These impulses will be guarded against the more carefully as the goods that they threaten to undo are the well-marked ones of money, material possessions, and other salient forms of wealth, which are also the most likely to be associated with conscious beliefs as to their value. Therefore, activities which threaten these must evade detection by the will. It is thus that rationalizations and ‘lapse districts’ can sometimes be beneficial.

A strategy that is less costly to the will and one’s resources is to find indirect routes to emotional reward. This is done by perceiving activities or goals that seem and/or are agreed upon to be normal, rational or necessary in some way, but are primarily maintained by their ability to deflect attention away from the emotional reward that maintains belief in the value of the activity. Thus, though on one hand it is common to admit that ‘everybody wants to be loved,’ undertaking romance with that conscious aim seems inappropriate and if detected is a turn-off. The paradox of happiness is that if consciously sought for, it is elusive, but if ignored in the pursuit of other activities, it may ‘alight upon us,’ as Hawthorne put it, comparing it to an elusive butterfly.

Belief in the intrinsic worth of myriad activities as a means of indirection is hypothesized as ‘maybe the most elementary ... tactic’ (2001, 192) employed in generating appetite for reward. From belief in the intrinsic importance of winning in sports, to belief in the intrinsic value of friendship, knowledge, love, truth, and indefinitely many things and activities that bring us rewards in life, these beliefs are plausibly the result of important processes of indirection. This largely unconscious tactic prevents the rewards based on them from becoming predictable and thereby incapable of generating the kind of emotional reward that has maintained such beliefs.[58]

Ainslie notes that ‘piercing indirections [may be] the basic mechanism of wit,’ and when wit goes too far, it appears as cynicism (2001, 194). And it likely appears cynical to suggest that there is no intrinsic value to love or friendship, but ‘only’ the emotional (and other) rewards that they bring. In fact, providing such an explanation, even if perfectly successful, doesn’t show that there is no intrinsic value. It does however provide a reason for thinking that we would be powerfully motivated to think that there

[58] I think Ainslie overstates the extent to which belief in intrinsic value is based on avoiding predictability of rewards. I think their value is largely due to the fact that conscious awareness of the ‘goal’ of these activities makes it difficult to persevere in them when they are aversive at the time, and so they can also be understood as commitment devices. But the important point is that the beliefs are in the business of stabilizing or intensifying motivation and a central feature of their functionality is deflection of attention from those very motivations.

were, even if there is not. More importantly, the feeling of cynicism only arises from the perspective of someone who holds intrinsic value to be a higher, more noble sort of value than the emotion-centered relational values in terms of which the belief in intrinsic value is explained. I aim to undermine that perspective.

1.3 Summary and Look Ahead

In this chapter, we saw how Ainslie's model of the will can explain how some kinds of beliefs—e.g. in moral requirements, in constructed facts, and in intrinsic value--can be explained in terms of their motivational consequences. The primary purpose of providing these examples (in addition to displaying some of the various phenomena that Ainslie's model is equipped to handle) is to explain how we could have the belief, as well as the strong *feeling* that we have good, powerful reasons to do something even without having any corresponding aims. Contrary to the way it can often seem to us, all these beliefs can be explained as essentially in the business of generating, stabilizing and/or intensifying motivation. Their representation as beliefs akin to those about mind-independent facts, on this view, is to render them apparently free from the tampering that could otherwise undermine the motivation which it is their business to maintain. This then is meant as at least a partial explanation of why we would believe in moral requirements and intrinsic value even if there were none, and this explanation crucially depends on the power of commitment and the subjective limitation of options.

Ainslie's model, while fascinating and fecund, is incomplete for my purposes. It gets us on the right track by showing us the enormous theoretical significance of commitment, systemization and their side-effects, but by itself does not give us close to as complete a story about moral judgments and beliefs as we would like. Most obviously, it leaves out an account of the moral emotions and how we come to have them, as well as any role of culture in fashioning moral beliefs, judgments and motivations. I want to emphasize that this is not a criticism of Ainslie, since it was no part of his task to explain moral judgments, beliefs or particularly moral motivations.[59]

In Chapter 2, we'll see the explanatory role of commitment expanded into the realms we have so far neglected. Frank (1988) and Joyce (2006) have argued for evolutionary accounts of moral sentiments and moral judgments respectively which are squarely and explicitly commitment-based. I will in effect use their accounts as significant components of my commitment model of moral judgment. I'll begin with Joyce's account, since it includes Frank's as a component. When we get to Frank's 'theory' of the moral sentiments, I'll provide some important details that Joyce leaves out, as well as make my own contribution to Frank's model, which in turn improves Joyce's. Then I'll return to Joyce for an important criticism of Frank, namely that he leaves out the 'cognitive' component of moral sentiments.

This will put us in a position to expand our commitment model into the social and cultural realm. Joyce presents evidence to argue that moral projectivism is the most plausible interpretation of our best available science. For Joyce, this part of his project is meant to answer the question of how (by what means) natural selection forged moral judgments, while the personal and interpersonal commitment stories he provides are

[59] To the extent that his model does provide (partial) explanations for these things, it should be counted as an additional point in the theory's favor.

meant to answer why (what kind of selection pressures led to it). Joyce's drawing attention to the role of the cognitive component of moral judgment, though welcome, is too vague, overly linguistic and implausibly dependent on natural selection as a mechanism for its development.

I follow him in arguing for moral projectivism,[60] but not with the intention of making any claims about natural selection's toolkit, but as another explanation for our belief in intrinsic value. I then present a sketch of Jonathan Haidt's Social-Intuitionist Model of moral judgment, as well as the extensive evidence he recruits to support and expand it, in order to assimilate it to my own expanded commitment model. I argue that the conceptual and linguistic resources made possible by culture augment, if not create, the cognitive component(s) that Joyce insists upon. Including the role of culture and public moral reasoning allows me to argue for commitments at the intrapersonal, interpersonal and cultural levels. The role of moral reasoning on Haidt's account is best interpreted as a (complex of) commitment/stabilizing device(s), though he says nothing about any role for commitment in the model.

[60] Understood as a causal account of moral experience.

Chapter 2: A Commitment Model of Moral Judgment

2.0 Introduction

In this chapter I'll argue that we can understand quite a lot about moral judgments in terms of commitments. More strongly, I'll argue that peculiarly moral judgments have an essentially committing function.

In chapter 1 I employed Ainslie's work to provide an empirically-supported theoretical underpinning to the claim that awareness of our own motivations can weaken the power of those very motivations. One of my primary goals there was to show that both (but not only) moral beliefs and perceptions as of intrinsic value could be explained in motivational terms. These explanations were meant to imply that the core function of such beliefs is to deflect attention away from the workings of our own motivations.

This chapter will continue an argument for the essentially motivational, committing function of moral judgments, with a focus on the role they play in keeping us unaware of the relational nature of our values. But it might be thought that I am getting ahead of myself. The arguments in Chapters 1 and 2 provide an account of why we would conceive of (moral) values nonrelationally even if they are not, but that is not enough by itself to show that there are in fact no nonrelational values or reasons for action. While I do think that the arguments in the first two chapters provide grounds for being suspicious of our belief in nonrelational values or reasons, they cannot replace a direct engagement with philosophical arguments to the contrary.[61] That is the work of Chapter 3.

In this chapter I defend a commitment model of moral judgment. I'll draw from and improve on arguments made by Richard Joyce and Robert Frank to the effect that moral judgments (or in Frank's case, moral sentiments) are essentially in the commitment business. These authors (both of whom are motivated to *save* moral discourse and sentiments) both make crucial usage of Ainslean insights, but do not recognize the full importance of what his work can contribute to a commitment model of moral judgment/sentiment. Then I will discuss the work of Jonathan Haidt (who is in no way arguing against moral discourse or for morality as commitment-device). I will show that the most plausible aspects of his work make a very natural fit within my commitment model of moral judgment. Armed with an understanding of commitments and their motivational importance, an influential contemporary theory of moral judgment and an investigation into the origins thereof, we can make sense of moral judgments largely in terms of their tendency to commit us to ways of feeling and acting.

Peter Godfrey-Smith distinguished between two types of function using terminology that I will borrow. The *teleonomic* function of something is roughly the thing that it does that explains why it was selected for, if it was selected for at all. An adaptationist hypothesis with respect to some trait postulates some function that the trait served that increased the organism's (inclusive) fitness. I will be presenting Joyce's and Frank's hypotheses (with some modification of my own) that moral judgment has

[61] Except for 'arguments' to the effect that it just really seems like this kind of value exists. The arguments in the first two chapters are supposed to explain why it would seem that way without it actually being that way.

commitment as its teleonomic function. That is, its function is to ‘solve commitment problems’ in Frank’s words, i.e., those problems that cannot be solved without the capacity to commit to certain kinds of actions.

The other kind of function is an *instrumental* function. This is the function(s) that something has here and now, if it has a function at all. The former kind of function might be called what something is *for*, while the second what it is *good for*. The heater in a car is *for* warming the passengers, but it is also *good for* cooling the engine. So it has one teleonomic function and at least two instrumental ones. After presenting the evolutionary arguments, I will argue that at least one significant instrumental function of moral judgments is also commitment. If it is true that its teleonomic function is commitment, then it will not be very surprising that it has this instrumental function as well. However, though the argument for commitment as an instrumental function of moral judgment will not be as strong without the (cultural and biological) evolutionary, teleonomic argument, I hope to make it an independently plausible hypothesis that moral judgments do serve, here and now, a committing function. That is to say that moral judgments are ‘good for’ making commitments, and in ways that go well beyond whatever biologically or culturally adaptive functions they might (still) have.

Commitments are fundamentally about reducing one’s options, either objectively or subjectively. This is sometimes understood as removing or blocking the options entirely, but most commitments are not absolute in this way. It is enough to be committed to one course of action that there are obstacles of some sort to other options being chosen. When some such obstacles exist, I will say that options have been reduced, and therefore one is (to that extent) committed. In arguing for this committing function, I will be concentrating on the subjective reduction of options that moral judgments facilitate. Unsurprisingly, the primary means by which I hypothesize that we do this (in the moral domain) are: our experience of and belief in intrinsic value, the deflection of attention away from our own emotions and desires, and the linguistically-mediated moral concepts such as (moral) rights, duties and obligations, as well as (moral) responsibility, praiseworthiness and blameworthiness.

2.1 The Evolution of Helping

There are several evolutionary forces which can explain what I will follow Joyce (2006) in calling helping behavior.[62] All of these forces can operate without any trace of a moral judgment. More loosely, all of these processes could be effective to an arbitrary extent and the resulting behaviors and creatures could be regarded as pre- or otherwise non-moral. This is important because it helps us see better what we do and don't have in mind by moral judgments, which will be helpful when we are addressing the questions of the specific teleonomic, and later, the instrumental function(s) thereof. Also, in our all-too-brief discussion of the evolutionary forces that can lead to helping, we will lay a bit of the groundwork for those that we hypothesize to have led to moral judgments.

[62] As opposed to altruistic behavior. Though this term has been used in evolutionary theory, most notably by Trivers (1971), the term 'altruism' suggests a certain kind of motivation for one's actions, one that is concerned with the well-being of the object of one's altruism. However, no such motivations are suggested by the evolutionary accounts of 'altruism'. For this and related reasons I agree with Joyce that we should avoid the term 'altruism' in this discussion in favor of the motivationally neutral 'helping'. My discussion of the biologically-evolved helping behaviors follows his (19 – 43), until I get to cultural evolution, where I go beyond Joyce.

Kin Selection

Organisms that are prone to make some sacrifice of their individual fitness for their (close) relations can gain an inclusive fitness advantage. If we direct our attention to an individual's genes, we see that many of those genes are the same in related individuals, the more so the more closely they are related. William Hamilton (1964) originally developed the concept of inclusive fitness, and created what is now known as Hamilton's Rule, which is the following. The trait of reproductively benefiting others at some reproductive cost to oneself is expected to be favored by natural selection if

$rB > C$,

where r is the extent of genetic relatedness, B is the reproductive benefit to the beneficiary and C is the reproductive cost to the benefactor.

Based on this mechanism alone, we would expect there to be a significant tendency to sacrifice for closely related family members, less so for less closely related members and so on. This can but needn't involve the sacrifice of one's life, and more typically includes the common sort of helping associated with sharing food, offering protection, tending to one's children and nieces and nephews and so on. We now know that the social insects, which engage in very significant sacrificial behavior, are on average more closely related even than human siblings (which share on average 50% of their genetic material).

This is the most plausible and most highly confirmed theory by which helping behavior can arise by natural selection, and I don't want to spend more time on it beyond noting that the mechanisms that are developed by natural selection based on this force can result in helping behavior to nonrelated individuals (that is, not more related than average). The first reason for this is that the mechanisms may not, and likely do not, employ foolproof tactics for identifying kin. Hatchlings have presumably evolved their 'imprinting' behavior due to selective pressures to follow their mothers, but the mechanism that has evolved is one that has them follow whatever object they see first in life, no matter how little it looks like a duck (Lorenz 1937; Bateson 1966). There is evidence that people have similarly coarse means of (nonconsciously) 'identifying' kin, such as those with whom they have lived or been familiar from childhood (Shepher 1971, 1983; Lieberman et al. 2003).

The second potential ramification of kin selection beyond helping kin is that the mechanisms that were developed by kin selection can be modified and/or appropriated for novel forms of helping. As an example, there has been quite a bit of discussion of the role of oxytocin in 'attachment' or pair-bonding (Nelson & Panksepp, 1998; Strathearn, Fonagy, Amico, & Montague, 2009). Oxytocin is an extremely old hormone that seems to have been exapted from some other function(s) to regulate maternal nurturing behavior some 200 million years ago (Allman, 2000). Its role in bonding mothers to infants seems to have been extended to bonding those who are not genetically related, most notably mates.

Mutualism and Direct Reciprocity

Some tasks require, or are made more likely to succeed, by cooperative effort. A single wolf or lion in a hunt is much more likely to be unsuccessful than if others help. Up to a point, each additional helper makes the likelihood of success increase, as well as reduces the likelihood of injury to any of the members. However, since helping incurs risks, there will be selective pressures to let others do the work in circumstances where one's added value is less than the likelihood of incurring fitness-reducing costs. Nevertheless, in many circumstances, helping will be directly beneficial to one's fitness and so one would expect evolutionary pressure to favor this kind of helping behavior in at least some circumstances. This specific kind of cooperative behavior, restricted to behaviors that directly and immediately increase fitness is called *mutualism*.

Mutualism is different than *direct reciprocity* in that the former does not require any relationship between the cooperators beyond the specific, simultaneous cooperative venture at hand. The advantages of being willing to engage in directly reciprocal relationships are large and clear. Some tasks are extremely important and difficult for one individual, though much easier for another. Removing parasites is an excellent example. It is a small task for one monkey to clean the parasites from another, but very valuable and difficult for the one with the parasites. Of course there is no obvious reason for the other monkey to do this, unless we suppose that that monkey might also have parasites (now or later), and that a willingness to help now will be repaid by a willingness on the part of the other monkey to help later. But once the first monkey's parasites are removed, why would she help later? Because that might not be the last time she needs help, and if she does not reciprocate, she is unlikely to receive help the next time, at least not from the same monkey. The costs of failure to reciprocate do not have to be restricted to lack of help in the future. The original monkey could actively punish the nonreciprocator. Doing so entails a cost, but depending on the details, it might be a cost that is more than offset by its ability to deter nonreciprocation.

Indirect reciprocity

Direct reciprocity has limits. In a small group, direct reciprocity could lead to the evolution of helping behaviors because once a critter has exploited all the others, nobody will help it again. It is easy to model circumstances in which such an exploitative strategy is fitness-reducing. However, in a larger group there will be lots of critters to take advantage of, and in such situations an exploitative strategy could well be fitness-enhancing. This strategy runs into serious trouble once the possibility of developing a *reputation* comes into existence. If A sees B fail to reciprocate with C, then A will be more likely to engage in helping behavior with C than B, all equal. Further, if A can tell D – Z about B's unhelpful ways, then B could very quickly be at a great disadvantage relative to where he could have been by simply reciprocating. Creatures that live in groups where there are many opportunities for mutually beneficial reciprocal exchanges, and where those creatures prefer to engage in such exchanges with creatures with good reputations, will find their reputations of very great importance to their (reproductive) interests.

Provided that there are sufficiently many members of the population with good (enough) reputations, the price of punishing (reputed) defectors will be very low to the punishers and very high to the punished. And this is only considering punishment to include a lack of willingness to engage in cooperative ventures. It says nothing of forms of punishment that are of higher cost to both punisher and punished (though still higher for punished than punisher(s)). This is a very important feature of the power of indirect reciprocity to establish helping behavior in a population. A good reputation is of very great importance, and yet can be bestowed or withheld 'like a magical substance' (41) by the rest of the community.

Another important point is that where punishment is low-cost to the punishers and high-cost to the punished, there is the possibility of developing almost any trait, even ones which in other circumstances would be very fitness-reducing, such as indiscriminate food-sharing or peacock-tails.[63] It is in large part due to the power of indirect reciprocity to render almost any trait fitness-reducing or –enhancing that Joyce agrees with Alexander (1987) that this force is at the heart of the development of moral systems. I agree that it is important, but also that full-blooded moral cognition and moral judgments require cultural explanations, whether evolutionary or not.

Group Selection and Cultural Evolution

I want to acknowledge the possibility of group selection at the genetic level, but mostly to point out its limitations, and to use it to transition to a discussion of cultural

[63] Peacock tails are not considered the result of indirect reciprocity per se but sexual selection. However, in a context in which females can be choosy at low cost, and they, for whatever reason, take a liking to tails of a certain variety, they can 'punish' those without the desired tails by not mating with them.

evolution. My discussion of cultural evolution will be more extensive than the other mechanisms.

I am not going to describe them, but there are models which show that group selection at the genetic level is possible. Genetic group selection could occur in cases where some individuals in a population engage in what appears to be fitness-sacrificing behavior, even nondiscriminately. If by doing so they make the group stronger (including and especially by simply increasing their numbers), and then that group competes, as in warfare perhaps, with another group without such helpers, then the first group could dominate to the detriment of the latter group. In some cases, the group with such helpers could entirely wipe out groups without any, allowing for the spread of genes related to such helping behavior. However, all these models require that the groups have extremely low migration or 'intermarriage' rates in order to work. For example, even if one group kills all the men in another group, if they reproduce with even a fairly small percentage of the women, genetic group selection will not be likely.

However, in circumstances in which genetic group selection is rendered unlikely or impossible, cultural group selection can operate. Evolution needn't work only on genetic material. Wherever there are heritable variations in traits that differentially affect reproduction rates, there will be selection (Lewontin 1970). In order for group selection to work, there must be sufficient similiarity within groups and sufficient variability between groups. It is very difficult to achieve these conditions in genetically-based groups, but not in cultural groups.

As I just said, group selection requires heritable variation between groups, as well as a sufficient level of homogeneity within groups. When these two conditions are met, and there is intergroup competition, group selection will occur (Richerson & Boyd, 2005, p. 206) . First I want to point to the two primary mechanisms by which Boyd and Richerson argue that variation is maintained between groups and homegeneity stabilized within groups.

Henrich and Boyd (1998) have shown that a 'conformist bias' is likely to be adaptive in a variable environment because it allows for the very quick adoption of behaviors that are likely to be successful in that local environment. Since the majority are likely to be employing strategies that are at least as adaptive as the minority, imitating the majority is a good bet. Conformist bias plays a role in maintaining intragroup homogeneity and intergroup variability. Because it pays to imitate the common type, this selective force causes the cultural variants[64] that are already common to become more so, and the ones that are relatively rare to be yet rarer.

Another force that maintains the required intergroup variation and stabilizes intragroup behavioral norms is what Robert Trivers calls 'moralistic punishment'.[65] I will use 'moralistic punishment' to refer to third-party punishment that is costly in ways that go beyond the potential costs already implicit in indirect reciprocity. Boyd and Richerson's discussion of 'moralistic punishment' overlaps with the kind of punishment discussed above under indirect reciprocity, since the former includes the withholding of, e.g., friendship and mating opportunities by members who are aware of one's negative reputation. However, it goes beyond the mechanisms discussed under indirect reciprocity since it can include even more costly punishments such as being attacked and killed by

[64] Cultural variants are analogous to genes. They can range in complexity from individual phonological rules to entire grammars, from entire religious or moral systems to the specific morality of contraception (Boyd and Richerson 2005, 90- 91).

[65] Trivers 1971, cited on p. 204 of Boyd and Richerson 2005.

one's 'erstwhile compatriots' (200). Such severe punishments are part of why moralistic punishment is a more effective means of stabilizing large-scale cooperation than 'mere' reciprocity. For even if such moralistic punishers are relatively rare, the severity of the penalties they are disposed to mete out can yet make for sufficiently effective inducements to avoid punishable activities (200).[66]

Moralistic punishment is similar to what Gintis (2000) and Bowles and Gintis (2004) call 'strong reciprocity'. As they define it, a strong reciprocator (with respect to some norm(s)) is one who (typically) obeys the norm and who punishes norm-violators, even if such punishment carries fitness costs greater than that borne by those who violate the norm and/or those who don't punish violators. I prefer the term 'moralistic punishment' because 1) the kind of punishment modeled by Gintis and Bowles is ostracism, while I mean for moralistic punishment to include (but not require) 'moralistic aggression' (Trivers 1971) and not just ostracism, 2) the third-party nature of the punishment I have in mind makes the term 'reciprocity' misleading, and 3) I will be arguing that the motivational underpinnings of moralistic punishment centrally include what we call moral anger. Therefore I will be arguing and not just assuming that this evolutionary mechanism warrants the association with 'moralism'.

Following Axelrod, I'll call norms about punishing those who defect from norms 'metanorms'. Metanorms are one of the mechanisms by which Axelrod's evolutionary models showed that norms could get established and maintained. Metanorms are important because even in an environment where temptations to defect are low (perhaps

[66] Just to be clear, indirect reciprocity *can be* just as costly as moralistic punishment, since being shunned by one's entire group can be tantamount to being killed in many environments. However, killing can be done by as few as one individual, while being shunned by the entire group requires the entire group.

due to emotions such as guilt and shame) if nobody is motivated to punish the small amount of defectors, a norm can still collapse (Axelrod, 1986, p. 1104).

At the time of his original writing, metanorms were only theoretical, in the sense that there had not been experiments done to test for their existence. But now it is clear that we do engage in costly third-party punishment both in the laboratory (Carpente et al., 2004; Fehr & Fischbacher, 2004; Fehr & Gachter, 2002; Knutson, 2004; Ostrom et al., 1992; Yamagishi, 1986) and in the field (Barr, 2001; Cordell & McKean, 1992; Price, 2005) and that doing so significantly raises cooperation levels. There is also good evidence that this general willingness to undertake costly punishment is culturally universal, though there is significant variability between cultures with respect to how severely and at what cost third-parties are willing to punish what 'degree' of uncooperativeness (Henrich et al., 2005). This experiment showed that people in all of the 15 diverse cultures they studied are willing to incur a financial cost to punish 'defectors' in a cooperation game, even though they had never met the person and their punishments would have no chance of inducing the punished player to behave differently toward the punisher. Further, the degree of 'altruistic' behavior (sharing money) was positively correlated with the degree of costly punishing behavior across all societies. It is typical for subjects to say in post-experiment interviews said that their punishments had been motivated by moral emotions, especially anger.[67]

To my mind, these general results could not be any less surprising; it is obvious

[67] As we'll see below, in Kurzban et al.'s (2006) study, post-experiment interviews suggested that the extent to which moral emotions had been experienced was related to whether or not the subjects thought that their decisions to punish would be known to others.

that we often like to punish.[68] What is less obvious is that developing such a taste might have been crucial for the evolution of large-scale cooperation (Gintis, 2000; Bowles and Gintis, 2004). But punishment can stabilize *any* behavior, not just cooperative ones. However, the behaviors that are most likely to survive intergroup competition are 'cooperative' in the sense that they allow for larger group size, which puts those groups at a (great) advantage relative to (much) smaller groups.

For example, the Nuer and Dinka were large ethnic groups in the southern Sudan in the 19th century. Cultural differences between them allowed the Nuer to cooperate in larger groups than the Dinka. When the Nuer came to want grazing land on which the Dinka were living, the Nuer attacked, defeated and occupied the Dinka, eventually assimilating tens of thousands of them into Nuer culture. The fact that the vanquished Dinka could be assimilated, by conforming and/or punishment, shows that there is no need for group extinction in order for cultural variants (in this case, in the form of norms) to spread quickly. Even very high rates of physical migration do not necessarily wash out cultural variation (Boyd and Richerson, 2005, p. 207).

Intergroup competition isn't the only way that cultural variants related to cooperation can spread. Groups need not (directly) compete in order for group members to know about the norms in other groups. Since evidence and theory both indicate that people have a 'prestige bias', i.e., a tendency to imitate the successful, if there are norms that cause other people to be more 'successful', those norms can spread by imitation

[68] The term 'altruistic punishment' in this context is both ubiquitous in the literature and very misleading. Nowhere that I'm aware of in the literature is there evidence that third-party norm-enforcement is altruistically motivated. In post-experiment interviews, subjects tend to report anger at those they punished, not motivations to benefit anyone, including society at large. The motivations at issue are directed not toward benefiting others, but harming those who misbehave.

(210). Boyd and Richerson (2002) have modeled this process mathematically and the results indicate that this mechanism can operate in many different conditions (210).

I think this mechanism is important because it shows how cultural variants can spread due to their contribution not to group strength per se, but do to their effect on people's reward centers. 'Successful' here therefore needn't be meant in a way related to possession of goods associated with reproductive fitness, but rather related to subjective satisfactions. If some beliefs or practices for example make for more emotionally rewarding lives, those beliefs and practices can be acquired by others as a result of this (not necessarily conscious) perception or association. I will resist the temptation to elaborate or give candidate examples.

I'll close this section with a summary of the general picture that Boyd and Richerson provide of how cultural and biological evolution interactively shaped 'tribal social instincts'. I have left out some of the arguments that support the following picture, such as those supporting our tendency to identify with symbolically marked groups and those arguing for (moral) emotions as means by which norm-adherence is made more likely. I think the former is not much in need of argument, and the latter will be the subject of extensive argumentation below. At this point I just want to give the general view of some of the leading researchers in the field to summarize the arguments that have come so far, foreshadow those to come, and because it will be useful for me to refer back to this general framework later.[69]

Rapid cultural adaptation created the circumstances for the evolution of social instincts that are apparently unique to humans. Cultural evolution produced cooperative

[69] Summary to follow from pp. 214 – 5 of Boyd and Richerson 2005.

groups identified by a variety of kinds of symbolic markings. The environments created by such groups selectively favored 'instincts' geared to life in such groups, including psychological structures prepared to internalize the norms of one's group(s). Those without the instincts, which included emotions such as shame and guilt, were on average more likely to transgress against social norms and be subject to a variety of punishments as a result, reducing their fitness.

Cooperative groups in conflict with one another set the stage for an 'arms race' in which cultural variants (most obviously including different kinds of (meta)norms and punishment strategies) that could generate ever-greater in-group cooperation would tend to proliferate. Groups, or tribes, of up to a few thousand members became common by roughly 100,000 years ago, setting human groups apart from all other social animals whose groups are not based on kinship.

The emotions or other motivations constituting or underpinning these new tribal instincts evolved without erasing older instinctual motivations favoring self, friends and family. This simple but central fact of human existence makes for inevitable and inherent conflict in individual human lives. Emotions designed for tribal-scale cooperation are often in conflict with other emotions or motivations designed for directly reciprocal relations, relations involving kin and those involving oneself. Such a picture begins to suggest why humans are susceptible to kinds of psychological sickness and conflict that do not afflict other animals, who at most have self- and kin-directed motivations.

Still no moral judgments!

Boyd and Richerson wonder aloud why natural selection should favor new sorts of motives or instincts, rather than simply calculating the optimal mixture of cooperation and defection, given the risk of punishment (2005, 214). Their answer is a less-developed version of the one I will provide from Robert Frank (1988) below, which crucially employs Ainsliean resources.[70] But as we'll see, Frank's argument, as well as those of Boyd and Richerson and those upon whom they draw, do not provide us with an account of the evolution of *moral judgment*. Despite a lot of talk about moral norms, moralistic punishment and moral emotions, what we have a story for so far is 'simply' a suite of motivationally-loaded perceptions; we have so far nothing but desires and/or aversions in reponse to, or perhaps as part of, perceptions of happenings in the social world.

But moral judgments seem to be something more than desires and aversions. It is a commonplace that the sense of acting from a motive of fulfilling, say, a moral obligation, is not experienced as doing what one most strongly wants to do. Judging that another person is morally blameworthy for some action and deserves to be punished is not (typically) experienced as a felt desire to punish the person, which desire is stronger than whatever desire or aversion there might be that would motivate the person to not punish. For example, one can judge oneself or one's loved ones morally responsible and blameworthy where doing so appears to run very strongly counter to what one most strongly desires to judge.

[70] Though Boyd and Richerson do not cite Frank or Ainslie.

Therefore to the extent that I am arguing for the teleonomic function of moral judgment as a commitment device, I have at best only given part of such an argument. What comes next rejoins Joyce in his attempt to argue for the committing function of specifically moral judgments.

2.2 The Nature of Moral Judgment

Before we launch into a story about how moral judgments might have evolved, we need to get an idea of what we're talking about when we're talking about moral judgments. As it happens, I am in almost perfect agreement with Joyce about what we should consider a moral judgment.[71] As it also happens, I don't think there is space here to argue for these claims in a way adequate to alleviate significant disagreement. I will restrict myself to arguing in any detail for only one one them, and the others I will do little more than describe and motivate.

One of the claims is that moral judgments are fundamentally concerned with interpersonal relations, especially restricting individuals' pursuits of their (especially 'narrow') interests. Many of these restrictions will involve rules or norms involving the circumstances under which one is expected (or required) to help and/or cooperate. So the story about the evolution of moral judgment will in part be a story about the evolution of helping and cooperation. But the evolution of helping and cooperation have gotten on quite well in other species without anything like moral judgments. As we've seen, there

[71] Which is one of the reasons that I center my discussion of the evolution of moral judgment on his account. If we thought moral judgments were quite different things, his account would not be so useful. The extent of our agreement on these matters also allows for a particularly focused disagreement on the issue of normative moral fictionalism, to be taken up in Chapter 4.

are fairly well-understood and –established mechanisms of helping and cooperative behavior that operate both prior to and independently of moral judgment. I have included a sketch of these mechanisms prior to the hypothesis(es) about moral judgment both because they will help us to get clearer on what is and isn't being attributed to the power of moral judgment, and because the account of the evolution of moral judgment builds on the prior reproductive[72] benefits of helping and cooperating.

The Characteristics of Moral Judgments

There is very much disagreement about whether moral judgments express beliefs or something like desires. Those who think they express or represent beliefs are called cognitivists and those who think not are called (pure) noncognitivists. Joyce considers them to be expressions of both beliefs and desires, as do I. And I also follow him and Gibbard (1990) in calling the non-belief component a conative state. So moral judgments characteristically[73] involve some belief and some motivation to act in accordance with that belief. This motivation to act doesn't mean that one will actually act, or is even motivated to act in a particular case in a particular way (e.g., to punish someone). What it means is that the judgments (characteristically) express attitudes like approval or disapproval, and/or emotions such as (moral) anger or contempt, or a commitment (Joyce says subscription) to a standard of behavior.

[72] Whether in terms of genes or 'memes' (cultural variants).

[73] Which is not to say necessarily. None of these claims about moral judgments are meant to be claims about what moral judgments necessarily involve. I think there are rarely if ever necessary and sufficient conditions for folk concepts, and a concept with a history as long and varied a history as morality is one of the least likely candidates, or so it seems to me.

Moral judgments have a distinctive subject matter, which is the regulation of interpersonal interactions or relations. That is not to say that all moral judgments are necessarily about interpersonal interactions. It is to say that this is their central concern.

Moral judgments involve concepts such as desert, justice, rights, obligations and duties. These concepts figure into a system of punishments and rewards for behavior that, conceptually and experientially, are not merely prudential or consequentialist. This is no criticism of consequentialism as a normative theory. It is to say that when people feel and think and say that someone deserves to be punished or rewarded, that judgment can easily persist even in the absence of any apparent beneficial or other consequences of the punishment or reward. That's not to say that the judgment that all things considered they should be punished or rewarded will stand indefinitely in the face of consequentialist considerations, but rather that the judgment that someone does or does not *deserve* something will not change depending on what the consequences would be of their getting what they deserve.[74]

Self-directed moral judgments require a moral conscience involving emotions such as guilt. These emotions play a critical role in a person's regulation of her own moral conduct. Joyce thinks that the emotion of guilt involves the thought that 'one deserves some kind of penalty for one's actions' (67 - 68). I am not convinced of this, but I do think that there is an important truth in the neighborhood, which I will discuss below.

Finally, moral judgments have 'practical clout' (62, passim).[75] Practical clout is

[74] This is not exactly the way Joyce puts the point, but I think the gist is the same.

[75] Joyce also thinks they require language. I am sympathetic, but don't have a strong sense either way. I also don't think it matters for my purposes.

the combination of moral inescapability and moral authority, concepts distinguished and clarified by David Brink (1997). The former is the property of being applicable to a person regardless of that person's ends. Rules of etiquette and institutional rules have this feature. The rule against talking with your mouth full *applies* to you whether or not you have any concern for etiquette or anybody who cares about etiquette. Likewise, if you are a member of a dues-paying organization, the rule that says you have to pay your dues applies to you whether or not you care about the organization or want to pay your dues or anything else you might want or not want.

To say that these rules apply to you is not to say that you should follow them. Not only might you have countervailing reasons not to follow them, you might have no reason at all. If you have no reason to comply, then the rules have no (moral or other) *authority* over you. To say that a rule or norm has moral authority over someone is just to say that that person has at least some reason, if not overriding reason, to act in accordance with the rule or norm. Putting it in non-rationalistic terms, it is to say that the norm cannot *legitimately* or *appropriately* or *rightly* be ignored.

I agree that moral judgments are characterized by[76] both inescapability and authority, which is to say clout. Which is to say that if someone judges that some action is morally wrong for someone to do, then they will agree that it is wrong whaver ends the person might or might not have, and that the person cannot legitimately, appropriately or rationally, ignore the content of that judgment just because, e.g., they don't care about morality. I don't mean to imply that people have this in mind when they make moral judgments or even that they would assent to it just as I stated it. But I do mean to imply

[76] Notice I don't say conceptually entail or presuppose.

that in the context of someone making a (paradigmatically) moral judgment, they would not retract that judgment if they were informed, and believed the information, that the person had no desire to conform to the moral prohibition. They would not thereby think that the judgment did not apply to him nor that it failed to give him any reason for acting.[77]

Now I want to reproduce my own version of a scenario Joyce provides (58) as an illustration of what it is for a judgment to have moral authority. I do this both because it will help demonstrate that moral judgments are characterized by this kind of authority, and more important, because it will serve as a case in point for a theme that was introduced in the last chapter and will remain a theme throughout this essay.

Moral judgments involve emotions and attitudes of approval or disapproval. But when we make public moral judgments, we do not merely report or express those emotions or attitudes. To do so would be immediately recognized as different from what we actually do. To see this, imagine that Jimmy is an impassioned anti-abortion activist. He waits outside abortion clinics until people come to the clinic, and says to them, ‘The idea of you getting an abortion causes a feeling of disapproval in me.’ But this is silly. Jimmy is really passionate about the cause, so he is more likely to say, ‘The idea of abortions is disgusting to me!’ Or perhaps rather than report his feelings, he can express them thusly: ‘Boo abortions!’

These are not moral judgments as we recognize them. And the most salient characteristic they seem to be lacking is any ostensible reason for someone to care whether he approves or not. He doesn’t seem to be attempting to provide a consideration

[77] And if they were to do so, I think it is fair to say that their judgment is no longer peculiarly moral.

that his audience should take seriously, unless they care about his feelings. This is not the way moral judgments are. They do not (purport to) *refer* to one's own feelings, however much they might be *caused* by them. They say, 'Abortion is immoral!' or 'Abortion is murder!' or 'Abortion is evil!'. They all purport to describe (in this case) an action by attributing it a property, roughly speaking (intrinsic) wrongness, or claiming it as a member of a set of actions which are (intrinsically) wrong. *If* we think that something really does have the property of being morally wrong (or evil), then we do not feel anything like the same liberty to simply say 'So what?' as we could in response to another person (especially a stranger) expressing or reporting their feelings.

So the first thing to note is that making one's emotions known to another does not seem to provide the kind of authority over their deliberations or motivations that making a moral judgment does. I want to suggest that an equally important thing to notice is that the judgments seem to *direct attention away from the emotions*, for both speaker and audience. Notice that I didn't give as an example of reporting feelings, 'Abortion is disgusting!'. Because even that expression still directs attention away from the felt emotion to an ostensible property in the object that in some sense (appropriately) causes disgust. We do not mean to say only that it as a matter of fact elicits, or tends to elicit disgust in us (whoever is saying that it does). Saying that something is (morally) disgusting is different than saying that we are in fact (morally) disgusted by it. The first directs attention away from our particular emotional reaction and additionally has a normative component that implies that disgust is in some sense an appropriate response.[78] The second does not.[79]

What we're calling the belief component seems to direct attention away from the conative component, though the conative component is clearly doing a lot of motivational work. What's interesting, and for some paradoxical, is that it looks like the belief component is somehow *adding or bolstering* motivation. It seems that we are more capable of prolonged and intense emotional engagement if we consciously represent some cause as morally right or some action as morally wrong than if we consciously represent ourselves as 'merely' having emotions and/or desires in connection with some cause or action.

Ainslie gave us a way of understanding the relationship between moral belief, intrinsic value and motivation in the last chapter, one that centered on their roles in deflecting attention away from the emotional rewards they are based on. What for some is the paradox of beliefs adding motivational force to desires is not paradoxical for those who were persuaded by the arguments in Chapter 1 for (moral) beliefs as personal rules. Recall that experiencing these rules as beliefs was hypothesized to stengthen the motivational force of the rules by representing and experiencing the rules as not subject to tampering. If the rules were perceived of as the interpersonal bargaining strategies that Ainslie hypothesizes them to be, that would make them more vulnerable to rationalization by short-term rewards.

[78] Almost all sensibility theories make some kind of appeal to appropriateness or merit of a moral emotion. Cf. McDowell, 1985; Wiggins, 1987; Darwall et al. 1992; D'arms and Jacobson, 2000.

[79] Though saying that one is *morally* disgusted blurs this distinction because of the features of morality we've been talking about. To experience *moral disgust* might just be to experience something as being in some (moral) sense rightly or appropriately disgusting.

Section 2.3 will argue for why we can understand the evolution of moral sentiments and judgments in terms of their committing function(s). We will revisit the role of belief in deflecting attention from motivation when we move from moral sentiments to moral judgments proper. Section 2.4 will argue that our propensity to project our (moral) values also plays an important role in maintaining our lack of awareness of (the nature of) our motivations, and thereby the (mysterious) moral authority that seems to give (overriding) reasons for action to people independently of those motivations.

2.3 The Evolution of Moral Judgment

We should expect moral sentiments and judgments to play a commiting role

Section 2.1 laid out several forces which did or plausibly could have played a role in the evolution of helping, or cooperative behavior. One of those forces I called 'moralistic punishment' in part because I promised to argue that the kinds of motivational or emotional connotations suggested by this phrase are warranted. But the arguments of that section did not presuppose any particular mechanisms by which those forces achieved their results. However, the summary I provided from Boyd and Richerson at the end mentioned that new 'tribal social instincts' had (by hypothesis) evolved alongside instincts directed toward self, friends and kin. This section will argue 1) that those instincts take roughly the form of what are often called moral sentiments, 2) that those sentiments have a committing function, and 3) that moral judgment is something over and above those sentiments and also is in the commitment business.

None of the sentiments for which evolutionary arguments have been presented require the kind of linguistically-mediated representational abilities that are uniquely human. But I think moral judgments do. I will argue that moral concepts and judgments are plausibly in the business of stabilizing norm-adherence in the context of humans' ability to employ the resources of language and deliberation to 'rationalize' non-adherence to (onerous) social norms. The inherent conflict between different and powerful sources of motivation threatens the stability of adherence to social norms, especially those that cut most strongly against other aspects of our motivational psychologies. The sentiments that evolved to stabilize adherence to social norms have to compete with other evolved sentiments and motivations for control of our behavior. Being evolutionarily much newer than some of these other motivational forces (such as for food, water, sex, loyalty to family and allies), the pro-social sentiments will presumably often lose.

Where language makes rationalization possible, and social norms are onerous, a cultural variant that can stabilize motivation in the direction of norm-adherence gives groups a competitive advantage against others without it, as well as individuals a fitness advantage in a society with strong norm-enforcement. Remember, to the extent that we evolved in the context of an arms-race toward ever-greater ingroup cooperation, we should expect to find ourselves heavily equipped to behave in ways that were (and may still be) geared toward the survival and growth of large groups of unrelated members. And for the last several thousand years or more, these members have both been hyperbolic discounters and had linguistic resources to employ in the service of older and often more powerful rewards. We should expect successful groups to have developed cultural variants which employ these very conceptual/linguistic resources to aid in the commitment of their members to the social norms. I think the peculiarly moral character of our moral concepts and judgments is largely explained by their playing such a functional role.

Let's begin the argument that the sentiments and judgments have a committing function by asking about the different ways that moral judgments could be adaptive.

2.3.1 Moral Conscience as Personal Commitment

Moral judgments made by individuals could be adaptive at the level of the group and/or the individual. Also, it's worth distinguishing between moral judgments made toward others and those made toward oneself. These two aspects of the question generate four distinct questions regarding the possible adaptiveness of moral judgments. The first two questions ask how making moral judgments about *others* could benefit one's group and/or oneself, and the second two ask how making moral judgments about *oneself* might benefit one's group and/or oneself. Joyce thinks that the last one is intuitively the most challenging, that is, the question of how making moral judgments toward oneself could provide adaptive benefits for oneself (108). If this is in fact the most intuitively challenging for most people, it is because most people are unaware of the extraordinary importance of commitment.

Joyce identifies 'judging oneself in moral terms' with 'having a conscience' (108) and asks how having this trait could help its possessor outcompete those without it. Joyce begins with the reasonable assumption that judging an action to be morally good or right increases the likelihood that one will perform the action (and the reverse for judging an action morally bad or wrong). So, if there are classes of actions or omissions that tend to increase or decrease reproductive fitness, then any psychological mechanism that makes those actions or omissions more or less likely (respectively), especially as compared to other potentially competing mechanisms, then that mechanism will tend to be favored by natural selection. So, ‘self-directed moral judgment may enhance reproductive fitness so long as it is attached to the appropriate actions’ (109). Earlier, I gave evidence that helping behavior will often be fitness-enhancing**.** So it might well serve one’s own reproductive interests to judge one’s own ‘pro-social’ behaviors in moral terms.

The premise that self-directed moral judgment generally makes behavior in line with those judgments more likely seems very intuitive. It’s important to note that this claim does not commit one to the controversial thesis of motivational internalism, which maintains that moral judgments logically or otherwise necessarily entail some motivation on the part of the judger. Joyce briefly argues for a qualified non-cognitivism according to which ‘canonical moral judgments’ do entail corresponding motivation. But this needn’t detain us. The important point, with which I think very few will disagree, is that in general and on average, moral judgments provide (at least) a contingent psychological connection with motivation.

As Joyce rightly notes, this ‘connection can be brought out by considering the phenomenon of weakness of will’ (110). There is tremendous empirical support[80] for the claim that even when we believe and even perhaps know (*pace* Socrates) that the pursuit

[80] Joyce cites Schelling (1980), Elster (1984) and Ainslie (1992).

of short-term gain has long-term harms that outweigh those gains, we often fall prey to those short-term desires, to our long-term chagrin.

Having already discussed the issue at length in Chapter 1, I will not belabor the examples. Such an ubiquitous problem prompts Joyce to wonder why ‘natural selection would have left us with a design flaw that so often handicaps the pursuit of our own best interests’ (110), and concludes that it is likely that it is the result of competing psychological faculties, each pursuing its own agenda and trying to control behavior. Then Joyce goes on to puzzle over why natural selection wouldn’t have rectified such a ‘glaring problem’ and concludes that it is the ‘inevitable price to pay for some other end [which might be] the ability to calculate subjective preferences in a flexible way’ (110).

Fortunately, the personal commitment hypothesis does not depend on the controversial adaptationist assumption that ‘design flaws’ are to be understood as inevitable tradeoffs, the necessary (i.e., adaptive) price to pay for yet more adaptive features. The crucial point is that humans (and seemingly all vertebrates) show hyperbolic discounting in a wide range of scenarios; short-term desires often outcompete long-term desires. With this in mind, ‘[s]uppose there was a realm of action of such recurrent importance that nature did not want practical success to depend on the frail caprice of ordinary human practical intelligence’ (110). And there are plausibly such realms. Behaviors related to (conditional) cooperation, helping, and reciprocity (for good or ill) are excellent candidates. In these realms of action, those with effective commitments to cooperate, help and reciprocate plausibly had a reproductive advantage over those who were more flexible.

In the relatively small groups in which we evolved, the potential benefits of a good reputation and the harms of a bad one were almost certainly greater than they are today, though they are still significant. But these benefits are often significantly delayed and probabilistic. Long-term benefits are hard enough for our motivational systems to secure, but make them probabilistic and you severely exacerbate the problem. If it is (very) difficult for us to secure long-term over short-term goods when both are fairly certain, then it is even more difficult to do so when the short-term goods are fairly certain and the long-term not so certain. And this is under the highly questionable assumption that we could calculate the probabilities accurately. The temptations we face in the social realm are not dispassionate. Where short-term rewards have access to the rationalizing apparatus, those subjectively estimated probabilities, not strong or certain to begin with, are easily nudged (or shoved) in the direction favored by the short-term rewards.[81]

So Joyce's hypothesis is that 'natural selection opted for a special motivational mechanism for this realm: moral conscience' (110). The way he sees it, if you think about an outcome that you 'merely' desire, you can always rationalize to yourself that forgoing it wouldn't be so bad after all. On the other hand, if you see the desired 'object' as not something merely desired, but '*demanding* desire—then your scope for rationalizing a spur-of-the-moment devaluation narrows' (111). That is, if we believe that we *must* do something, then we no longer have the same latitude or freedom to negotiate with ourselves. Believing that certain actions have 'practical clout' that is independent of our desires and interests has the power to quiet or even silence our calculations, to limit our freedom in an important and beneficial way. For if left with

[81] The preceding paragraph is my own elaboration of the problem, not Joyce's.

maximum freedom, we will often choose the lesser goods. He agrees with Dennett that moral considerations function as beneficial ‘conversation-stoppers’ that act as a welcome kind of “‘unquestioning dogmatism that will render agents impervious to the subtle invasions of hyper-rationality’” (Dennett 1995, p. 508, quoted on p. 112).[82]

Joyce asks us to imagine someone without moralized thinking who wishes to have a reciprocal relationship either because of the sympathy he feels for the other or because of long-term profits. And then suppose he violates the terms of the relationship, perhaps due to having been won over by some incompatible short-term reward. Joyce insists that the person might feel surprised, disappointed in himself, distressed perhaps, but he could not feel guilty. For Joyce, ‘guilt requires the thought that one has transgressed against a norm’ (112). And since the person has no moral concepts, he cannot think that he deserves to be punished. Though he might wish to compensate the person out of sympathy, this feeling can quickly fade on its own, and without moralized thinking about the matter, there will be nothing to force his mind back to the violation. There is nothing to make him think that not only has he *risked* punishment but that he *deserves* it. That these additional, more intense and arguably more stable forms of self-recrimination can and do occur in the moralized thinker when he does cheat gives us reason to think that those same psychological attributes play an important role in generating and maintaining the motivation not to cheat.

In sum, the hypothesis is that *adding* moralized thinking, and the moral sentiments that come with it, to the psychological profile of an otherwise normal,

[82] An important difference in Dennett’s conception of the function of moral considerations as conversation-stoppers is that the hyper-rationality they avoid is not that of rationalizing oneself into short-term goals at the expense of long-term ones, but rather that of attempting to find *rationally conclusive* reasons for actions.

sympathetic person gives that person additional motivational resources to perform/avoid adapative/maladaptive social behaviors. It is important to see that *while Joyce relies only on thought experiments and intuitions about rationalizing desire, we now have Ainslie's resources* to (much) better understand why leaving all options open and relying on introspective desire to guide actions would put these values in serious peril.

At this point we turn from conscience as personal commitment to that of interpersonal commitment. In doing so, we review Robert Frank's (1988) interpersonal commitment model, which makes much more significant use of Ainsliean resources.

2.3.2 Frank's Interpersonal Commitment Model

2.3.2.1 The Importance and Ubiquity of Commitment Problems

Robert Frank (1988) offers a compelling argument that at least some of our 'moral' emotions evolved as a means to solving very important problems that 'cannot be solved by rational action' (4). Before we continue, a couple of things need to be said about Frank's usage of 'rational action'. First, by rational he means self-interested. As will become clear later (especially in chapter 3), I favor a desiderative view of rationality.[83] Fortunately, I don't think Frank's usage obstructs or obscures the specific points I wish to make.[84] Frank argues for the normativity of having, cultivating and acting on moral sentiments in terms of their overall contribution to material payoffs, which are assumed to be in one's interest and so rational to pursue. In presenting Frank's

[83] I thereby (unlike Frank) avoid the conclusion that someone who doesn't cheat because they strongly and stably desire not to, even when there's no chance of their getting caught, is thereby acting irrationally.

[84] However, I do think that the larger argument Frank makes is negatively affected by this choice.

view, I will often speak of ‘payoffs’ in terms of self-interest, but these arguments can easily be extended to ‘payoffs’ that are understood in terms of our desires, or things we care about, which need not be understood in self-interested terms.

Second, I realize that Frank’s conception of a ‘purely rational agent’ might raise philosophical questions and/or hackles. Frank uses it to refer to an agent capable of evaluating (it needn’t be flawless evaluation) which available actions are in her interest and then reliably acting accordingly. It is someone who ‘keeps her options open’ and is reliably capable of doing what seems best (for her) at any given choice-point. Again, we can eliminate the self-interested component and focus on the issue of perfect flexibility.

Perfect flexibility is to be contrasted with the perfect amount of flexibility. One might think that a ‘purely rational agent’ would have the perfect, or rationally appropriate, amount of flexibility. To conceive of a rational agent as one which is perfectly (maximally) flexible, is to conceive of rationality as a decision procedure rather than a normative standard. Specifically, it conceives of rationality as a decision procedure which seeks to maximize at every opportunity. But one could rationally judge that attempting to maximize at every opportunity is less likely to get results that one would want, and so rationally decide not to attempt to maximize in certain situations.

The best thing I can think to do here is to acknowledge this issue and to explicitly use Frank’s ‘purely rational agent’ as a term of art, referring to an (imaginary) agent who keeps her options maximally open, such that she is free to (re)calculate and act in whichever manner seems best to her at any time. In defense of this usage, it is fairly common to think of rationality (and freedom) in this way, and to view some paradigmatic examples of commitment-based reasoning and action as paradigmatic examples of irrationality. Also, to the extent that our rational faculties are tied up with our faculties of deliberation and executive control in line with those deliberations, then commitment requires the occasional, perhaps frequent, inhibition or avoidance of our rational faculties. Nevertheless, we can set these issues to the side by considering 'purely rational action' (what I will often just call 'rational action' in the context of discussing Frank) as including perfect, i.e. maximal flexibility.

There are two related but distinct fundamental reasons that rational action cannot solve commitment problems. The first is that because of the nature of our motivational systems, we often fail to do what is (perceived to be) in our long-term interest. As we've seen, this is problem enough in one's purely personal life. In social life, severe penalties of many kinds can result from being seduced (figuratively or literally) by short-term rewards.

Though the above limitation on rational action plays an important role in Frank's account, it is almost entirely in the service of supporting his argument for the second fundamental reason that purely rational agents cannot solve this range of problems. That reason is that being *perceived* as a purely rational agent puts one at distinct and often severe disadvantages in a host of interpersonal situations that are and have presumably been common for millions of years. Frank provides a plausible theory of moral sentiments to show how these sentiments could have evolved as a response to a variety of 'commitment problems.'

Frank defines a commitment problem as one that 'arises when it is in a person's interest to make a binding commitment to behave in a way that will later seem contrary to self-interest' (47).[85] What are some of these problems that require commitment to solve? Frank gives the following examples, involving cheating, deterrence, bargaining and marriage.[86]

Suppose Jones and Smith could have a profitable business venture together due to their each having different skills that when combined make for a great team. Unfortunately, being in business together as partners will provide each of them with opportunites to cheat the other with little risk of getting caught. Both of them know this, and though they would both be better off working together and neither of them cheating, they would each be better off going it alone than working together and getting cheated (whether only one or both cheated). So what they need is some way to make a binding commitment not to cheat.

Now suppose Smith grows wheat and Jones raises cattle on adjacent land. Jones is legally responsible to keep his cows from eating Smith's wheat, but fencing his cows in would cost $200. If he doesn't fence them in, he can expect his cows to eat about $1000 worth of wheat. However, he also knows that it would cost Smith $2000 to recover his money by going to court. Of course Smith threatens to sue Jones if his cattle eat Smith's wheat, but if Smith is purely rational, this threat isn't credible, so Jones has no material incentive to pay for the fence. What Smith needs is a way to make a visible binding

[85] I would not define them this way for two reasons. First, commitments don't have to be binding in the sense that alternative options are rendered totally unavailable; they just have to involve a reduction in one's flexibility to choose from alternative options. Second, as Frank's second example presented below shows, it is not always (perhaps not even most of the time) the case that one *will* later deem the commitment to be contrary to self-interest. What is crucial is not that they *will* later seem (or be) so, but that they *might*.

[86] Examples and part of ensuing discussion from pp. 47 – 50.

commitment to sue, in which case Jones will have clear material incentive to fence his land.[87]

Now we find Smith and Jones in a bargaining situation. Again, there is some mutually profitable venture they could join in together, but Smith needs the money more than Jones. This puts Jones in a stronger bargaining position when it comes to deciding how they are to split the expected profit, call it $1000. If Jones knows Smith is desperate and that Smith knows that Jones is not, Jones can believably threaten to walk away if he doesn't get, say, $800. Since $200 is better than nothing and Smith is needy, it would be in his interest to accept the deal. But again, if he had some way to commit himself, in a way that Jones could perceive, not to accept less than a fair split, Jones's strategy couldn't work. It would then be rational for Jones to offer a fair split.

Now let us suppose that Smith and Jones are considering marriage and adopting children. Such an arrangement will require tremendous investment from both partners, an investment that would not be worth it for either partner if she thought there was a significant chance that the other one would leave if she saw a more appealing opportunity in the future. In order to make the marriage a viable possibility, it is now in their interest to reduce their own options with respect to the future, i.e., to make a commitment to each other.

2.3.2.2 The Inadequacy of Secured Commitment

[87] This is a case of a conditional commitment, that is, one that only triggers the committed action contingent on another's action. So long as the commitment works, Smith will never have occasion to think it contrary to his self-interest, and will never forego any of his interests. But it is a clear case of commitment because at the time of making it, he realizes that he might have to act contrary to his self-interest since he can't be sure Jones's cattle won't manage to eat his wheat or that Jones will be rational and build the fence.

In all four of the above cases, it might be possible to commit oneself by altering one's *material* incentives. This could be done by means of contracts or by otherwise changing the situation so that fulfilling the commitment is straightforwardly in the person's interest. Following Nesse (2002), I will call such cases of commitment 'secured'. Commitments that are based on concern for one's reputation and/or on emotions are 'subjective'.[88]

Frank's examples bring out two important points. The first is that these kinds of situations are not at all uncommon or unimportant, leaving aside the simplifying details. Social interaction is replete with situations in which people have significantly but not fully overlapping interests, such as in the joint venture and marriage examples. In the former kinds of cases, monitoring the others' behavior is often difficult or impossible, and temptations for individuals to gain at the others' expense are common. The marriage problem often also contains instances of the cheating problem, in numerous kinds of ways, only one of which involves desertion. Social life is also full of cases where the interactants' interests are much less if at all overlapping, such as in the bargaining and deterrence examples. Being a 'social player' incapable of solving the kinds of commitment problems presented in these simple cases will plausibly entail significantly reduced fitness.

[88] Hirschleifer (1999) calls them 'rational' and 'emotional' respectively. I follow Nesse's terminology because I think that despite its misleading suggestion that the former are in some way stronger or more reliable than the latter, it more than makes up for this by 1) not implicitly opposing rationality and emotion, 2) avoiding the suggestion that following through on unsecured commitments is (necessarily) irrational and 3) because the unsecured commitments (arguably) needn't involve any experienced emotion. Of course nothing substantive hangs on the choice of terms.

Of course, these examples might seem too modern to be relevant to any evolutionary account. But their point is not to suggest that it was these specific kinds of problems (involving money and land-ownership for example) that had to be overcome. They are useful illustrations and reminders of the following truth. Where opportunities for exploitation, predation and the like are widespread, one can be guaranteed that there will be plenty of people ready to take advantage of those opportunities. As Frank says, 'being able to solve the deterrence problem woud be an asset of the first magnitude' (50). Likewise, 'to live is to haggle' (50); coming to agreements about how to divide the fruits of collective labor, or what can be exchanged for what, are ubiquitious aspects of human life. Failure to effectively engage in these social interactions surely would have tended to have a negative impact on one's fitness.

The other important point is that in very many cases, secured commitments are highly impractical or impossible. In the cheating cases, since an inability to monitor is crucial to the problem, it's hard to see what good a contract or other 3rd-party enforcement would do. Similarly, it's not obvious how to go about making contracts that require one to retaliate for injuries or to ensure that one does not accept a fair deal. More importantly, a point not mentioned by Frank, is that these options, as unlikely and unwieldy as they are now, were surely even less available prehistorically. Lastly, we might think that the importance of the commitment problem in marriage is reflected in the elaborate, various and often severe modes of securing marriage commitments. These include formal marriage contracts with (often severe) legal penalties and public pronouncements of vows which also generate significant incentives to avoid less formal, but not necessarily less severe, penalties. Nevertheless, in an environment where one can be choosy about mates, few would choose a mate bound only by such securements of commitment if there were others who displayed signs of being committed in an additional way.

2.3.2.3 The Power of Subjective Commitment

Attempting to solve these commitment problems by rearranging material incentives is only one of the strategies available to reward-governed creatures like us. In fact, as Frank notes (51), material incentives are never the *direct* motivators of action. Even food and drink, the material incentives *par excellence*, are pursued as a result of complex reward mechanisms in the brain. Though the material incentives clearly play a crucial role in explaining why we are motivated to eat and drink (we wouldn't be around if we weren't), the proximate cause of these behaviors is the reward-based motivational system, which generates feelings of intense hunger and thirst that are very powerfully motivational.

We know that there are other kinds of feelings that motivate us[89] as well, in which cases associated material incentives are not readily discovered and often simply don't exist. Emotions such as anger, guilt, greed, shame, envy and disgust can have powerful motivating effects independent of any discernible material incentives. Someone prone to strong feelings of guilt, for example, will be motivated not to cheat in many circumstances even when the material advantages favor cheating. Frank says of such a

[89] I will follow Frank in talking of the experienced emotions and felt aversions as being motivationally powerful. But I take no stand on whether the 'feelings themselves' (if we must reify them) play a causal, motivational role or whether they are epiphenomenal.

person that she ‘simply does not *want* to cheat’ (53). I think this is not quite right, or at least it gets us ahead of ourselves a bit. The feelings of guilt or aversion by themselves don’t *remove* the desires associated with material rewards, but rather provide *countervailing* desires.[90] At this stage, it is better to say that at least in many cases, though she might want (be motivated) to cheat, she is also motivated not to.

It’s important to note that her felt motivation not to cheat (her aversion to it) can be present *at the same time* as the opportunity to reap the rewards of cheating. This is important because of our old friend (or rather, foe) hyperbolic discounting. Were she to only anticipate feeling bad in the future, or losing out on future material benefits, the motivation to cheat would be much more likely to win out, since the rewards of cheating come now, or soon, and the punishments thereof only come later. To be sure, the anticipation of future bad feelings and material punishments can still affect this person, but two points should be emphasized. First, the aversion remains in the absence of anticipated future material losses or future aversive feelings.[91] Second, the immediately felt aversion to the action plausibly plays a crucial role in avoiding succumbing to cheating in many cases, given what we know about motivational discounting.

The same points apply to the emotions involved in solving the bargaining and deterrence problems. One rarely requires material incentive to pursue revenge; there are powerful motivations to seek vengeance that have nothing to do with material gain, and

[90] I think there is a plausible story to be told about how these aversions can play a role in removing or diminishing the countervailing desires, but Frank doesn’t tell one (though I will below). I do think that the effective removal of countervailing desires is an important component of commitment, but the sheer presence of aversive feelings of guilt only implies contrary motivation, not the removal of motivation. We’ll see below that Joyce’s and my account has the resources to improve Frank’s account on this score.

[91] In normal life, the present aversions and likelihood of aversive feelings in the future are not readily separated. But I think thought experiments involving the experience machine or memory-erasing drugs show that present aversions not only persist in the absence of anticipations of future aversions, but lose little if any of their motivational force. It seems to me that sometimes they might even be augmented.

are persistent in the face of significant material cost. The same can be said about our motivations to be treated in a way that we consider fair, or just. Plausibly, the fact that the emotional rewards of rejecting ‘unfair’ offers and taking vengeance are available immediately upon performing the required actions explains a large part of their motivational power. This seems especially true of vengeance. Frank opens his book with the history of the McCoys and Hatfields, who engaged in decades of very bloody reciprocal bouts of vengeance. Importantly, as with the infamous cycles of retaliation in the Middle East for example, the rewards of getting vengeance come (sometimes long) before the pain of retaliation. This fact is clearly not necessary for vengeance to occur, since people are often willing to die and suffer injury in the act of vengeance. However, one should expect that if whatever predictable pains associated with acts of vengeance were to come just prior to or coincident with the rewards of vengeance, the incidents of revenge-taking would drop dramatically.[92]

I think the point that we are at least sometimes powerfully motivated by such moral sentiments that have no discernible material payoff should be obvious. The question is how such sentiments could have come about, since as Frank notes, ‘we can’t eat moral sentiments’ (54). For such sentiments to have come about in a competitive world, there would have to be some material payoff to them. In Frank’s account, these payoffs all require that others be able to perceive that one has them, i.e., that one is actually committed in the relevant ways. In order for people to act on their preference to cooperate with honest people (people who are emotionally averse to dishonesty as such),

[92] A similar point is made in Ainslie 1992, that if the hangovers came *before* the drinking instead of the next morning, there would be many fewer of them.

they must have some way to distinguish them from those who are not. Only if they can do so will the honest be at a competitive advantage. Similarly, having a powerful (emotional) commitment to take revenge at great cost to oneself is likely to be counterproductive if the relevant people don't believe that you have it.

2.3.2.4 The Reputation Pathway

Frank discusses two ways in which one can effectively communicate to others that one has the relevant sort of emotional commitments, and therefore two ways in which these commitments could have evolved.[93] They are 1) hard-to-fake physical signals and 2) reputation. I'll discuss the second one first. We value our own reputations very much and are in general significantly influenced in our opinions of others due to their reputations. Universities, employers and consumers all rely heavily on reputation when considering whom to admit or hire (recommendations and references) or which companies to patronize (customer reviews, consumer reports, etc.). In the social sphere, reputation is no less important. The pool of potential friends, partners and mates from which one may choose will be greatly reduced if one has a reputation for dishonesty, thievery, unprovoked violence, of any number of other undesirable traits. Though this is much less the case in modern liberal societies than elsewhere and elsewhen, a reputation for 'weakness' (where that means inability or unwillingness to do violence if threatened or harmed) can have equally if not more dire consequences.

[93] I realize that putting it this way seems to put the cart before the horse. For they must have already evolved at least to some extent in order for them to be effectively communicated. That is correct, and is taken into account. However, once they can be communicated, additional and potentially much more powerful selection pressures can come into play.

How does reputation relate to commitment, and especially the emotional commitments we've been talking about? Recall, commitment problems are ones that 'can't be solved by rational action' where rational action means prudence.[94] Given that reputations are important, it seems that sheer prudence will dictate that we avoid cheating even when the likelihood of getting caught is low, and that we enact revenge even at significant cost to ourselves when there are reputational advantages to be had by doing so. And so on. So it seems that having a good reputation can be easily explained entirely in terms of prudence, with no role necessary for commitment.

The above reasoning only has force if one is ignorant of, or forgets, the crucial feature of our motivational systems that has been and will continue to occupy us. Future rewards are heavily discounted relative to the present. The benefits of establishing a reputation all come in the future, perhaps the distant future, and are probabilistic to boot. The costs come much sooner. Pursuing revenge typically entails costs and risks before, while and/or soon after doing so. The benefits of the reputation one contributes to generating come later (if at all, since one might die or be severely wounded by seeking revenge). Likewise, the rewards of cheating, of accepting unfair deals when desperate, etc. all come in the much shorter-term than whatever reputational benefits one accrues by doing the opposite.

[94] Factoring in the importance of a reputation might change some of the payoffs in the Smith and Jones cases above such that they are no longer commitment problems as we are understanding them. If Smith and Jones both know (and know that each other knows) that a reputation for being a weak bargainer will have quite negative long-term consequences, then Jones will know that Smith has material incentive not to accept an unfair deal, assuming that there is a significant chance of the deal becoming known. In such circumstances, it might be wise for Smith to commit himself to publicizing the deal, so that Jones would know that he 'couldn't' accept an unfair deal.

Due to the discounting of future rewards, purely prudential agents can be expected to get caught cheating, lying and/or stealing; be seen backing down from fights and/or failing to return injuries; be uncooperative and/or unhelpful when doing so would be onerous. And such people would be quite likely to develop corresponding reputations as a result of the ubiquitous human proclivity for gossip.

Therefore, the value we do in fact place on reputation has a very good rational warrant.[95] It provides us with good (though not sufficient) reason to believe that the person in question has emotional commitments of the relevant kind. The emotional commitments can overcome the problem of discounting because the aversive feelings of guilt and anger (just for example) are present at the same time, and usually before[96] the material rewards. We can infer (not necessarily consciously) from the fact that someone has the relevant reputation (which can include the mere lack of having a bad reputation, in many contexts) that they *really care* about being honest or faithful or about 'justice' (whether this is thought of as limited to concern about oneself being treated fairly, or extending the concern to others). Because of this, if Billy has a reputation for violence whenever his honor is challenged, one can be reasonably confident that he will retain this behavioral disposition even when nobody else is looking.[97]

[95] The same argument won't necessarily hold for the reputations of powerful institutions, for the reason that they often have the power to manipulate their own reputations by means of propaganda unrelated to their actual behavior.

[96] I think this is an important point, though Frank doesn't mention it. Frank rightly notes that the emotions are (can be) present at the moment of choice, and so are not discounted relative to the competing material rewards. But the fact is that the emotions are often present quite a bit *before* the material rewards are available. So they are not only not discounted relative to the material rewards, but the latter are discounted relative to them (though anticipation of those rewards is available sooner than the material rewards themselves, which complicates matters further).

[97] Though some of the third-party punishment literature below suggests that punishment goes up when potential punishers are observed.

2.3.2.5 The Sincere Manner Pathway

The other pathway by which moral sentiments could have evolved is that of the 'sincere manner'.[98] Emotional dispositions are detectable via physiological and behavioral signs. For these to be reliable signals of emotional dispositions, they must be difficult or costly to fake. This 'costly-to-fake' principle is a general principle of communication between potential adversaries, i.e., people whose interests do not overlap entirely. One head coach telling an opposing head coach what her team's next play is going to be is valueless under normal circumstances. Coaches do not tell each other lies about what their next plays will be because such declarations will soon if not immediately be rightly ignored as containting no relevant information. This is because a simple lie is an extremely easy way to 'signal' that one will run play X even if one will not. Likewise, if emotional dispositions were easy to mimic, then people could adopt them whenever it seemed prudent, in which case their value as signals of the relevant dispositions would be nullified.[99] Frank provides a considerable amount of evidence that a wide range of these behavioral and physiological cues are indeed difficult to mimic convincingly. These include but are not limited to a large range of facial expressions (including the timing thereof), tone of voice, posture, blushing, laughter, sweating, crying, pupil dilation, breathing rate, direction of gaze and eye movements. I won't review the evidence Frank provides that (at least many of) these signals are indeed difficult (for at least most people) to fake.

[98] These pathways are not mutually exclusive; in fact, Frank thinks the latter is unlikely to have been the exclusive pathway, for reasons we'll see shortly.

[99] One's reputation can act as a reliable signal because it too is costly (difficult) to fake, given the nature of our motivational discounting of the future.

I would like to discuss another principle that governs the kinds of passive signals that can arise by natural selection, the derivation principle.[100] This principle holds that any feature that acts as a signal will almost never have arisen just for that purpose. Frank's apt example of this principle is the dung beetle, which ostensibly benefits from 'communicating' to potential predators that it is a piece of poop. However, as Gould noted with characteristic humor, "'…can there be any edge in looking 5% like a turd?'" (quoted on p. 103). For small, incremental poopward changes in the beetle's appearance to have improved its survival chances by making it look more like poop, it must have already looked somewhat like a turd to at least one of its predators under some circumstances. Whatever the reasons for *those* particular morphological features, they must have been incidental to its making the beetle look like poop.

The reason I call attention to this principle is because the most likely answer to how the relevant behavioral and physiological cues came to be such cues brings us back, once again, to akrasia and hyperbolic discounting.[101] A small mutation toward a moral sentiment and/or an associated symptom thereof could not have been a signal to others that the bearer of the symptom also bore the sentiment, since no prior correlation existed between sentiment and symptom that could provide the basis for an 'inference' from symptom to sentiment. However, a small mutation toward a moral sentiment *can* increase fitness, even in the absence of any symptom thereof, if it makes a small contribution toward one's reputation. And the way it does that is by generating short-

[100] Due to the Nobel laureate ethologist Niko Tinbergen (cited by Frank, p. 102).

[101] Frank credits Jon Elster with noting that the main points of the reputation argument would hold without hyperbolic discounting (specifically, without the 'matching law', which is a product of hyperbolic discounting) if there were very high exponential discounting (82). I will continue to refer to the problem of akrasia as a product of hyperbolic discounting, since only that implies preference reversals, which are at the core of our experience and conception of weakness of will.

(and possibly long-) term rewards that can compete with the short-term rewards associated with cheating, cowering, etc.

2.3.2.6 Adding Personal Commitment

Since the reputation pathway depends on moral sentiments playing an anti-akratic role, and the sincere-manner pathway depends on the reputation pathway, Frank's *entire argument* for how we might have evolved moral sentiments depends on those sentiments having played a role in combatting the effects of hyperbolic discounting! And yet despite the crucial importance that anti-akratic sentiments play in Frank's account, they are yet understated. Because of this, we are in a position to improve on Frank's account **(**as well as Joyce's). In his preface (xii), Frank makes a claim the gist of which is repeated several times in the book: 'In order for ['irrational' emotional dispositions] to be advantageous, others must be able to discern that we have them.' Also, 'if altruistic behavior is to have a material payoff in the manner set forth by the commitment model, at least some sort of biological symptom of emotion … appears necessary' (64).[102] For Frank, the adaptive value of the moral (anti-akratic) sentiments is entirely dependent on and exhausted by their role in solving prisoner's-dilemma-like scenarios with others. It is strictly an interpersonal commitment model.

But it is implausible that the *only* adaptive benefits of such sentiments come as a result of signalling to others, whether by means of reputation or visible signs of emotion.

[102] Of course if he means 'interpersonal commitment model' then this is true. But adding personal commitment to the commitment model is still a commitment model. And a better one.

While a good/bad reputation can make for (reproductive) benefits/harms, these benefits and harms can come independently of reputation. That is, one's community can punish and reward (especially punish) in other ways than communicating one's traits to others. The history of human societies does not leave one with the impression that violators of norms need only fear for their reputation.

In a highly relevant discussion on the creation of conscience in man, Nietzsche lists just some of the punishments Germany has employed in its history: 'stoning … breaking on the wheel … piercing with stakes, tearing apart or trampling by horses … boiling of the criminal in oil or wine … the popular flaying alive …, cutting flesh from the chest, and also the practice of smearing the wrongdoer with honey and leaving him in the blazing sun for the flies'.[103] Nietzsche, with characteristic insight, describes these measures as a way to 'impose a few primitive demands of social existence as *present realities* upon these slaves of momentary affect and desire' (GM II 3, emphasis in original). That is, just as we've been suggesting, social life presents myriad opportunities and temptations that can seduce us even if we 'know better,' due (in part) to our 'slavishness to momentary affect and desire'. Nietzsche's point in the first few sections of the second book of his *Geneology of Morality* is that such punishments were part of a process of 'breeding a conscience' into us.

Nietzsche's examples and analysis help us to recognize that small mutations toward sentiments (and/or the capacity to develop them) that could (alone or in combination with cultural norms) mitigate our attraction to short-term rewards could very

[103] Geneology of Morality (1887); essay 2, section 3 (henceforth cited as 'GM II 3').

plausibly have been (and be) adaptive for reasons independent of their ability to be detected, either via reputation or visible signs.[104]

At this point it's worth saying something about why Frank, so astute and careful elsewhere, would have failed to see that these emotional commitments need not be, and likely wouldn't be, useless or counterproductive in the absence of others being able to detect them. Frank is focused on the commitment model as a solution to the 'problem of altruism,' which is the problem of how it is that humans, who have undergone a Darwinian process of natural selection, could have evolved such that very many of us (even in America in the '80s, the time of Frank's writing) do not take opportunities to make ourselves better off (in material terms) even when there is little or no chance of being caught, and risk our lives and wellbeing to help or save strangers who have little chance of repaying us in kind.

Frank notes (42) that cultures can inculcate values in a number of ways, but that there are some people, such as sociopaths, who do not take up even intense cultural conditioning. If such people are more adaptive, then they should eventually dominate. But this has not happened. The explanation Frank explores is that such people may be 'observably different' (though the reputation pathway doesn't seem to count as involving an 'observable difference') and thereby do worse. The adaptive benefits that 'sociopaths' get are theorized to be more than offset, on the average, by their inferior reputations and/or observable lack of (emotional) commitments. And the potential harms that come with being committed to act in ways that could be (significantly) harmful to one's

[104] There is strong evidence that prison inmates are more impulsive than average (Caspi et al. 1994; Del Raine 1993).

(genetic) interests are more than offset by their possessors' superior reputations and/or observable commitments. These commitments not only help them gain valuable partners, mates, group-memberships, etc., but many of the commitments that *would be* harmful if acted upon *do not need to be acted upon* just because others can detect that the relevant people are in fact so committed.

There are so many potential cases in which acting on one's commitments would seem to reduce fitness *unless* that commitment were detectable that, though Frank developed the resources to argue for it, he ignores (or at least severely downplays) the benefits that very plausibly remain even if they were *not* detectable.[105] The benefits that 'sociopaths'[106] gain by their opportunism are not only plausibly offset by their relative difficulty in finding good mates, partners, etc. because of reputation or (lack of) other observable features, but also by the severely fitness-reducing nature of much human punishment. Also note that Frank doesn't suppose that the biological component fixes the specific contexts in which the sentiments are had. On p. 65, he says that 'nature's role is to have endowed us with a capacity that is much like a gyroscope at rest; and that culture's role is to spin it and establish its orientation.'

So the capacity to develop emotional commitments is (let's suppose) a product of evolution, but culture tells us (at least in large part) what to feel guilty, ashamed, angry,

[105] It might be that since Frank was attempting to vindicate morality, he wasn't motivated to see 'how much blood and cruelty lie at the bottom of all "good things"!' (GM II 3). Nietzsche's motivations were, of course, different.

[106] I put 'sociopaths' in quotes in part because there is presumably something of a spectrum of susceptibility to these 'moral sentiments', as opposed to a dichotomy of 'cooperators' and 'opportunists' (as Frank recognizes). The point is that given the nature of our motivational systems to discount the future, and the many dangerous temptations that exist in social life, there are plausibly considerable benefits to having mechanisms that can counter those temptations. 'Sociopaths' can here stand for those with relatively little capacity for emotional commitments, especially in the directions encouraged by one's culture.

disgusted, etc. at and about. This rough understanding of the role of the sentiments makes the 'personal commitment' part of the commitment model even more relevant as I see it. Cultures are so variable in their rules and norms that what in one society might be a relatively harmless temptation, in another results in severe reprisal. Without the susceptibility to be guided by a culture's norms, with the emotional/affective rewards that can compete on the same time-scale as the material rewards, one will be more likely than average to find oneself in prison or worse.**[107]**

I emphasize this point because I think it makes Frank's (and Joyce's) overall account even more plausible. *This is a third and independent pathway in addition to the reputation and sincere manner pathways.* Call it the 'personal commitment pathway'.[108] It is hypothesized to operate in parallel with the reputation pathway and helps provide the preconditions for the sincere manner pathway. Adding this personal commitment mechanism to the interpersonal commitment mechanism(s) just makes it all the more plausible that the 'moral sentiments' are profitably analyzed and understood as, at least in large part, emotion-backed commitments of a personal and/or interpersonal nature.[109]

2.3.2.7 A Moral Conscience Includes Motivations to Punish Moralistically

I just argued that the personal commitment pathway is an important part of an

[107] All this having been said, Frank's model does entail that 'sociopaths' will be present in limited numbers, since if there were none, people wouldn't cheat and so nobody would go to any effort to detect cheaters, which would create a very favorable environment for cheaters whenever one happened along.

[108] I'll say below why this is different from Joyce's version of conscience as personal commitment.

[109] Though I admittedly don't develop an argument for it, given the strong positive correletation between criminal behavior and impulsiveness, and the presumably very strong negative correlation between subjection to prison and other forms of human punishment and reproduction rate, I find the personal commitment pathway at least as plausible as the other two.

account of the evolution of moral sentiments, in large part due to the severely fitness-reducing nature of human punishment. But the sentiments targeted by this explanation are those of guilt, shame, empathy and the like, i.e., those that plausibly play a role in altering or constituting one's motivations so as not to violate (first-order) norms of cooperation. But what about the sentiments that motivate punishment? How and why is it that human punishment can be so severe in the first place?

The kinds of punishments Nietzsche was talking about look like quintessentially moralistic punishments, in the sense that they go beyond the mechanisms associated with indirect reciprocity, and at least often involve risks and costs, sometimes significant ones, to the punishers. As I mentioned in the discussion of cultural evolution, if such punishing behaviors benefit the group but at cost to the punishers, then we would expect those who reap the rewards of punishment without paying the costs to have a selective advantage, undermining the ability of such punishing behavior to proliferate.

This is where metanorms came in. Recall that metanorms are one of Axelrod's original (1986) mechanisms by which norms which are costly to uphold can be stabilized in large groups. They are norms calling for the punishment of those who do violate the norms. Now, so far we have an argument for how sentiments could have evolved that made us less likely to get exploited or punished for norm-violations, but is there a similar argument to be made for how we could have evolved sentiments to undertake the kind of costly punishment that characterizes moralistic punishment? There is, and it is an aspect of (the evolution of) moral sentiments that is glaringly absent in Frank's and Joyce's accounts.

In the discussion of moralistic punishment above, I noted that I found it entirely unsurprising that people are willing to engage in such punishment at (significant) material cost to themselves, and that moral emotions such as anger toward the 'defectors' seemed to be motivating punishment. In other words, *of course* (some) people 'take sadistic pleasure in punishing … transgressions' (Boyd and Richerson 2005, p. 220). This further claim isn't directly supported by the experiments, but anyone who didn't already know this just hasn't been paying attention. If we haven't been paying the relevant kind of attention, perhaps part of the reason is that there is a strong norm in our society against taking such 'sadistic pleasure' in punishment. And when there is a strong norm against having a certain kind of motivation, one can be sure that such motivations will be less likely to be perceived or attended to, at least in oneself and one's friends and allies. They can even, over time, be greatly attenuated. Nevertheless, the uncomfortable, even shocked reaction many of us have to Nietzsche's claim that 'in punishment there is so much that is festive!'[110] is not best explained by these 'sadistic' motivations having disappeared, but rather by their being the objects of strong moral disapproval.[111]

Of course we need to be careful here. If we understand 'sadistic pleausure' to simply mean enjoyment at another's suffering, then we will miss an absolutely crucial distinction. For people who just like to hurt others for no reason other than that they like it, are, for us, quite a different animal than those who like to hurt people who have committed heinous crimes. Heroes (if often of the comic-book variety) are often modeled on the latter sort, whereas the former is our picture (or cartoon) of evil. In other words, taking pleasure in harming others, for the vast majority of people, depends on

[110] GM II 6

[111] In fact, it might be quite rewarding, even pleasant to hurt someone who took sadistic pleasure in hurting others. Surely some (most?) of my readers have fantasized about getting their hands on a really bad guy, whether in real life, literature or movies.

perceiving them as having done something wrong. More accurately, it tends to depend on perceiving them as having a certain kind of motivational structure, or character. The people we often really want to punish are not those who want to do good but fail, but those who don't care (much) about doing good or even actively desire to do bad (or harm).

Frank argued that sentiments evolved as *reliable signals of behavioral dispositions*. That's a good explanation of why we are so concerned with 'character', and it also provides for a fairly straightforward style of explanation for the evolution of the moral sentiments associated with moralistic punishment. Frank's argument for the evolution of moral sentiments relied on there being observable differences between those with the relevant sentiments and those without, as well as the importance of reputation.

These same factors are at the heart of several models of the evolution of moralistic punishment which focus on how such punishment affects others' perceptions of the punisher. Generally, punishing can act as a reliable signal that one is committed to cooperative interaction as embodied in the local norms (Barclay, 2006; Fessler & Haley, 2003). But how could punishment come to be such a reliable signal? Perhaps unsurprisingly, I think the answer can once again be traced to hyperbolic discounting.

Recall that Frank's entire argument for the evolution of moral sentiments depends on their having originally arisen as a means of overcoming our 'slavishness to momentary affect and desire' (in Nietzsche's words). Because of the derivation principle, the observable differences between those with and without the relevant sentiments *could not* have originally evolved due to their ability to signal a propensity to behave in certain ways (cooperate, retaliate). However, the sentiments *could* have evolved due to their contribution to one's reputation as well as their tendency to commit one to courses of action that avoided punishment, and then when they became associated with observable symptoms, that provided another lever for selective forces to get hold of to further their development.

A parallel argument works for moralistic punishment. We start with the internalization of norms. Internalization is one of Axelrod's (1986) original mechanisms by which norms can be maintained, and is essentially another word for moral sentiments tailored to social norms. He defines a norm being internalized when it is 'psychologically painful' for someone to violate it, independent of direct material payoffs. Now, beginning roughly where Frank's story leaves off, we have groups of people some of whom have observable symptoms of having internalized norms. That is, there are some people who reliably signal that they would experience psychological pain at violating a norm and are therefore *relatively* unlikely to do so.

Now all we need to suppose is that in at least some variants, whatever mechanism(s) by which norms are internalized is such that some symptom is *also* observable when the person perceives *another person* violating a norm. But because of the derivation principle, we might think that this cannot act as a signal that one has internalized the norm and is therefore a good candidate for cooperation. Rightly so, but just as in Frank's argument, that's where the importance of reputation comes in.

If someone who has internalized a norm has an aversive response to another person violating a norm, it makes sense to think that the internalizer is less likely to interact with the violator than if they do not have it. Notice that I am supposing something beyond the mechanisms already discussed, or presupposed, in indirect reciprocity. There, we imagine that if A sees B defect on C, then A will be less likely to interact with B in the future. On average and in the long run, A's reluctance hurts B and helps A, as well as helping the group since it punishes defectors. It helps A because it lowers the odds of getting exploited.

But we have said nothing about what kind of sentiments might be involved in motivating A to avoid B. We have also, in this highly simplified description, left out that there are many circumstances where A might wisely, despite what he's seen, cooperate with B. First of all, they could engage in mutualistic interaction, in which case B could not exploit A and A would gain. Or B could make a secured commitment of some kind to A, giving A good cause not to fear exploitation. Or A could think the circumstances in which B defected were relevantly unusual such that, especially if the potential payoff were high enough, it warrants trusting B. Or they could agree to a reciprocal relationship in which B contributes first. Or lots of other stuff. In other words, the information one acquires in seeing someone defect is almost never good enough such that it would be rational, or fitness-enhancing, to reject cooperative ventures of all kinds with them.

But now suppose that in addition to whatever motivational mechanism is in play that makes A less likely to engage in cooperative interactions (perhaps only reciprocal interactions in which A 'goes first') with B, A has an aversive response to B arising from the same sentimental commitment that underpins his internalization of the norm that B has violated. Notice that it would seem adaptive for the mechanism in play that does *not* have to do with internalization of norms to allow for cooperative exchanges that do *not* present any risk of exploitation. In such cases there is nothing to lose and something to gain. And in fact it is quite clear, for example, that we don't hesitate to sell something to someone who gives us their money first, even if we don't trust them enough to reverse the order of exchange, or trust them at all.

But the kind of aversive reaction I'm hypothesizing in connection with norm-internalization we would not expect to disappear in the absence of a fear of exploitation. After all, by hypothesis, the sentiments that are at work in the internalization process have as their very business to retain their motivational force in the face of perceived opportunity for gain. An aversive reaction arising from this very internalization process is likely to have similar properties. And now the groundwork for establishing a certain kind of reputation is established. For someone who has this sort of internalization of norms will tend to acquire a reputation for not engaging with norm-violators, even when it seems to be to their advantage to do so. *We should not expect any such reputation to accrue to people who avoid reciprocal exchanges sheerly to avoid exploitation*.

And so now, just as in Frank's account, we have a story for why others would be more inclined to cooperate with people with such a reputation than those without one--namely that this reputation is a reliable (since costly) signal that they have internalized the local norms. Such people then can be at a selective advantage in highly social environments, despite passing up otherwise good opportunites for cooperation with norm-violators. And now the observable symptoms associated with their aversion to norm-violators can serve as a signal to others that they have internalized the norms*, providing for far more leverage to drive this process further*. Displays of being upset at others' norm-violations is now a signal that one is to be trusted not to violate them oneself.

Such a set of circumstances sets the stage for the kind of arms-race that indirect reciprocity has the power to foster. Just as indiscriminate food-sharing and peacock-tails can be the result of 'run-away' selective forces based on signaling, ever-greater displays of (let's just call it) moralistic anger or outrage can result from pressures to signal that one is a good partner. For once this signal is established, *not* to show anger at norm-violations is a signal that one does *not* have the relevant kind of internalization. And the more valuable this signal becomes, the more costly it will have to be to fake.

Enter costly punishment as a signal that one does in fact have the relevant sentiments. While one might be able to learn to act outraged, it is harder to engage in costly punishment without the relevant sentiments, again, *because of hyperbolic discounting*. The benefits of signaling are highly probabilistic and almost always some distance in the future. The costs of many kinds of punishment are more certain and nearer in time.

In line with the hypothesis I am developing here, there is evidence that people undertake costly punishment *as a means of displaying their emotions* (Xiao & Houser, 2005). In this study, costly punishment decreases when people are allowed to express emotions compared to when they are not allowed to. Therefore the authors hypothesize that costly punishment is a means of expressing these emotions. This supports the hypothesis I've been describing, since the adaptive point of punishment is to signal the presence of certain sentiments.

The account that I just gave is entirely dependent on reputation effects. In this it is unlike the account I gave of the sentiments geared toward avoiding first-order norm-violations (that is, after including my personal commitment pathway). It hypothesizes that there have been adaptations of sentiments with corresponding motivation to engage in costly punishment so as to signal to others that one is a trustworthy partner. An account predicting a willingness to engage in costly behavior to achieve reputation effects seems to predict that there will also be adaptations designed to be sensitive to whether or not there are people around to see one's costly behavior. We should therefore expect that moralistic punishment is sensitive to whether there is an audience.

This is just what Kurzban et al. (2006) find. When punishment in a cooperation game is anonymous, punishment levels are lower than when even just one person (the experimenter) is watching, and when the audience reached a dozen or so, punishment levels tripled. The subjects did not expect to encounter any of these people again, so it seems likely that the sheer presence of an audience triggers the motivation to engage in costly punishment, as opposed to calculations of expected future interactions with audience members. Also, subjects reported feeling more anger in the nonanonymous conditions compared with the anonymous condition (but these effects were small).

The story I just told has (moral) guilt and anger co-evolving due to the latter signaling the dispositional presence of the former, and the former signaling a likelihood of nondefection. There are also models of the evolution of moralistic punishment that advert to purely group selection, since those groups with moralistic punishers will tend to outcompete those without (Fehr & Gachter, 2002; Gintis, 2000; 2005). Of course these kinds of models are not mutually exclusive, and it seems to me that the most likely scenario is that both individual and group selective forces were at work. But even if it were somehow entirely group selection at work, the motivational dispositions for people to engage in moralistic punishment in the first place had to come from somewhere. I think that the strong relationship between moral guilt and anger argues for their coevolution (or codevelopment).

Still, it is important that groups with moralistic punishers are likely to be capable of maintaining cooperation among larger numbers than those without. This will make moralistic punishers more common and metanorms more stable. And where there is a stable metanorm prescribing that one undertake costly punishment of norm-violators, that will further drive the evolution of sentiments that reward doing so.

I've just described a 'how-possible' account of interactive biological and cultural evolution. Sentiments evolve biologically due their signaling properties, these sentiments make groups stronger which then allows for the spread and strengthening of norms associated with punishment, which creates an environment that even more strongly favors the biological evolution of sentiments motivating punishment of norm-violators. Then the cycle continues.

The specifics of how this all worked, or might have worked, are not important for my main point. My main point is that it is a plausible hypothesis that the teleonomic function of the sentiments that make moralistic punishment rewarding (and underlie moral judgments involving moral blame) is to commit us to certain ways of feeling and acting. But the intrinsically rewarding sentiments associated with cooperation and punishment do not exhaust what is involved in moral judgments.

2.3.3 The Cognitive Component of Moral Judgments

Joyce criticizes Frank for having an ‘impoverished view of conscience—[he sees] it as merely a set of aversions’ (121). Though Joyce acknowledges both that emotions are ‘central to our moral lives’ and are important in communicating interpersonal commitments, he also insists that ‘there is much more to morality than emotions, more to conscience than mere aversion’ (122). For Joyce, moral judgments, as well as moral emotions, have an essential cognitive component. The latter involve thoughts such as ‘I would deserve punishment if I did that’ and the concept of transgression against a norm (for the emotion of guilt). Moral judgments involve concepts and thoughts about ‘*transgressions* or *prohibitions* or *deserved punishments* or [some] other moral concept’ (80, italics his).

I won’t review Joyce’s arguments for this claim, but (for the time being) only say that I agree with the main point, in that without our concepts of desert, blameworthiness, praiseworthiness, (moral) permissibility, duties, rights, obligation and the like, we would not be left with something we would recognize as a peculiarly moral mode of normative life. Since Joyce considers such thoughts crucial for morality, and provides arguments for moral conscience as both personal and interpersonal commitments (using Frank as a big part of the latter), he considers his account superior to Frank’s since Frank fails to explain ‘why conceiving of actions as ‘morally forbidden’ is a more effective sort of commitment device than just having strong emotional inhibitions against these actions’ (121). Joyce contends that his account of self-directed moral judgments as personal commitments shows why they would be more effective than mere inhibitions. Further, if one can communicate to someone else that they are so committed, then they can also be effective interpersonal commitments.

Reducing Options

I think there is a lot of room for improvement in Joyce's account of morality as personal and interpersonal commitment, but I do think it gets one thing importantly correct that Frank's account leaves out. That is the aspect of moral thinking that seeks to reduce the scope for rationalization, that attempts to 'foreclose future possibilities' (122). As Joyce correctly notes, this 'foreclosing' is the most important thing about commitment (or at least moral commitment). In leaving out any conceptual or cognitive component, Frank leaves out the concepts that plausibly function, at least in part, to remove certain kinds of actions from consideration. On Frank's account, there is simply a competition between rewards or desires. We noted above that it was misleading for Frank to say that someone disposed to feel guilt didn't want to cheat. Nothing in Frank's account gives him a right to conclude this. The bearer of the moral sentiment has a competing motivation, just as the dieter might well want to eat the cake, but not eat because of (successfully) competing feelings of (anticipated) guilt or shame were she to eat it.

Joyce points to, but does not develop, the basic idea of foreclosing options when he says that when someone thinks of something as 'demanding desire' or that something is '*morally* required—that it *must* be performed whether he likes it or not—then the possibilities for internal negotiation on the matter diminish' (111). I want to postpone my argument that moral thinking really does serve to narrow the range of (perceived) options by, among other things, reducing the scope for rationalization. For now, I want to do two things. First, I want to highlight how much more effective a commitment would be under these circumstances. Second, I want to make a point about moral phenomenology in Joyce's favor. I'll try to show that these two things give us good reason to think that peculiarly moral thinking does reduce our perceived options. Since the point about phenenomenology contributes to the point about reduced options, I'll address the former first.

I take it that at least some of the time when we do something that we take ourselves to be morally obligated to do, we do it with the feeling that it it runs counter to our (strongest) inclinations. In fact, I take it that *to the extent* that we think that we did something *because* we were morally obligated to, we feel that we did it against at least some inclination. That is, it might be our moral obligation not to kill our own children, but we don't generally refrain from killing them from a sense or belief that we are obligated not to. We generally don't want to. On the contrary, if a poor but morally upright person finds a large amount of money, she is likely to have a (strong) desire for the money, and if she returns the money *from a sense of moral obligation*, she will not in general think, say or feel that returning the money is her strongest desire. This is just a generic example of Kant's apt contention that moral obligation is intimately (and necessarily) connected to a sense of duty, not (merely) inclination.

But on Frank's account, all we get is competing desires, and there is nothing in the story to suggest why we would ever feel that when we act in accordance with the 'moral' ones, we do what runs counter to our strongest inclination at the time. But I take it that such a feeling is commonplace when acting on self-consciously moral motives.[112] Now, does moral thinking really reduce our (perceived) options, at least in part by reducing rationalizing? Yes. And the fact that we don't conceive of moral decisions as being a matter of introspecting our strongest inclination at the time is an important part of why moral thinking can be so motivationally effective.

First, as Joyce notes, peculiarly moral language itself indicates that options are conceptually being closed off. 'I can't keep the money; it's wrong' is a likely a thing to say when making a self-consciously moral judgment. Again, words and concepts like 'must (not)', 'impermissible,' '(morally) required' or even 'unthinkable', suggest not a weighing of options but rather taking some of them 'off the table.' Second, we can remind ourselves of the importance of 'principles' in moral thinking. When we say that we are against something on moral principle, this does not generally indicate that one option is simply better than another option, but that the other is in a way 'out of bounds'. To be clear, these remarks are meant to apply mainly to regular folks, not necessarily philosophers, who of course are in the business of wondering whether and why what is better than what, and who generate and read complicated and subtle theories about the nature of moral principles and duties and rights. The above remarks are admittedly

[112] Two things. First, I don't want to say that we have this sense whenever we act on moral motives, only that it is a quite significant and common phenomenon, and Frank's account doesn't give us a story for why that would be. Second, it could rightly be said that Frank was 'only' giving a theory of moral sentiments, not moral judgments or obligations. There are a few things to say about this. The first is that such a comment begs the question against Joyce's conception of moral sentiments, which require moral thoughts (I am ambivalent on this point). Second, my point is not to say that Frank's theory is wrong, but importantly incomplete. He leaves out the conceptual/cognitive component of morality, even if morality is 'based on sentiment' (Frank 1988, 146). Third, as Joyce says, Frank doesn't say why making moral judgments would be a more effective commitment device than merely pitting aversion vs. desire. Joyce has a number of answers for why this would be, but I think the only one that really gets traction against Frank (others seem to be either restatements of points Frank made or run afoul of the principles of signaling between potential adversaries we reviewed above) is that moral judgments have the effect of subjectively removing or limiting options, as opposed to simply opposing them.

rough, but for now I hope they will suffice to make it at least plausible that peculiarly moral thinking at least often has the effect of reducing the apparently live options (specifically the 'immoral' ones).

Now I want to say something more about why such an effect would result in a more effective commitment. At one level this claim seems so trivial that it's superfluous to argue for it. Suppose out of options A and B you want to commit to A. It stands to reason that you will be (much) more likely to choose A if you can convince yourself that B is not really an option. Though it is sure that one will not choose B if one is literally unaware of it as an option, the psychology of moral obligation is of course more subtle and complex than a total blindness to certain options.

Making decisions is hard when inclinations pull in multiple directions and/or when there aren't clear criteria for choosing. In the context of temptations to cheat for example, pitting desire vs. desire can generate anxiety. There is motivation to relieve the anxiety. So long as the options are 'cheat' vs. 'don't cheat', the anxiety is most easily relieved by cheating, since once one cheats the anxiety about whether one will cheat is lifted (perhaps to be replaced by guilt, regret, etc.).

The other way to relieve the anxiety is to remove the source of temptation. I will have no strong anxiety about whether to cheat on my wife if I have no tempting options. But we can make a useful distinction between two senses in which one does not have options, i.e., objectively and subjectively, where the first doesn't depend on whether you conceive of yourself as having the option or not, and the second does. As Chris Rock pointed out in the context of the Clinton/Lewinsky scandal, many of the Republicans criticizing Clinton didn't have the objective options that Clinton did.[113] On the other hand, had Clinton thought that such actions were simply unthinkable, out of the question, categorically prohibited, then he would be subjectively removing them as options. Again, intuitively, thinking of the matter in such a way is less likely to allow for rationalization, and the anxiety that goes with it, which is most easily relieved by rationalizing in favor of the temptation.

The point here is not *only* that conceiving of options as not really options, taking them off the table so to speak, has a more stabilizing effect than relying on relatively fluid and fickle sentiment *alone*, though that surely is true. I am also making the stronger point that if an option can be effectively removed from consideration, that option is less likely to be taken than if the option is left open to consideration, even in the context of judgments against it that go beyond introspection of (or guidance by) our felt desires or sentiments. Taking options off the table is not to be utterly blind to their existence, but it is to somehow preclude or limit the ability of one's rational(izing) apparatus to (continue to) provide reasons in their favor. So the point about the increased effectiveness of commitments by removing options is more than trivially true. I think we can recognize both the effectiveness of this 'strategy' in maintaining our commitments in general, and that peculiarly moral language at least often lends itself to representing some candidate actions (and/or ways of living, feeling, etc.) as not merely inferior *all things considered*, but rather as *not to be considered*.

[113] 'You see all these fat Republican guys going: "l would never do such a thing. This is a travesty." I'm like, "Nobody's trying to blow you."' HBO Special "Bigger and Blacker" (1999).

These observations put us in a position to help answer Bernard Williams' (1985) 'interesting question' about the force of a conclusion in terms of 'practical necessity,' i.e., a '"must" that is unconditional and *goes all the way down*' (188). He wonders 'how a conclusion in terms of what we must do, or equally of what we cannot do, differs from a conclusion expressed merely in terms of what we have most reason to do; in particular, how it can be stronger, as it seems to be' (188). In addressing this same question in his (1981), he sees what I take to be the crucial point, namely that 'in the face of "I must," the other alternatives are no longer alternatives' (127). I won't dwell on the many things Williams has to say about practical necessity, but only note the following. His discussion is about arriving at conclusions of practical necessity as the result of deliberation and how they differ from other conclusions of practical reason. He is fundamentally concerned with the question of '[h]ow, in deliberation, can anything stronger be concluded in favor of a course of action that that we have most reason to take it?' (1985, 188).

I don't mean to deny that such conclusions can and sometimes do come at the end of a process of deliberation. But I think Williams' focus on deliberation, and that such conclusions can come at the end, keeps him from seeing that the point of making alternatives no longer alternatives is to *stop or preclude deliberation*. So I think it can be importantly misleading to say that the judgments of necessity are the *conclusions* of deliberation; they are often rather the *concluders*, or likely even more often *precluders* of deliberation. The most common and effective judgments of practical necessity preempt deliberation altogether.[114]

[114] Probably the vast majority of moral judgments do not involve any deliberation whatever, as will be discussed when we get to Haidt's research.

The more that we come to see morality as involving *commitments* to ways of acting, living, feeling, and so on, the more such a role for judgments of practical necessity makes sense. We have seen some reasons, and we will see more, to think that reason is at least often the 'slave of passion'. What we now conclude we have most reason to do might change if the deliberative (or rationalizing) process can continue indefinitely, especially in the presence of temptations and the anxiety they can generate. Given what seems to be the central role of commitment in moral thought and sentiment, and the seeming effect of peculiarly moral thought to remove (apparent) options from consideration, we have a plausible story for why '[p]ractical necessity, and the experience of reaching a conclusion with that force, is one element that has gone into the idea of moral obligation' (1985, 188).

2.3.4 Summary of Joyce/Frank/Campbell Synthesis

At this point, we've combined the strengths of Frank's and Joyce's commitment-based hypotheses of the evolution of moral judgments/sentiments. To Frank's story of interpersonal commitment we have added Joyce's personal commitment and cognitive component as well as my personal commitment pathway and reputational account of the evolution of moralistic punishment. The essential contribution of the cognitive component is that it actively reduces options and represents moral properties as independent of desires or interests. Joyce's story benefited from the fuller fleshing out of Frank's interpersonal commitment model. Joyce presented a very thin version of Frank's model (mine also leaves out important details, but space is limited), omitting the importance of the reputation pathway and the crucial role that hyperbolic (or extreme) discounting of the future plays in establishing the predictive power of reputation.

The personal commitment part of Joyce's account also benefits from the personal commitment pathway I added to Frank. It's important to see that Joyce's account of moral conscience as personal commitment and my personal commitment *pathway* to moral sentiments are different. That difference lies in the absence of any cognitive component in my hypothesized pathway. My personal commitment pathway is like Frank's pathways in that it only posits the development of particular kinds of desires and aversions in certain kinds of circumstances. There is excellent evidence that these sorts of desires and aversions, such as sympathy and an aversion to 'unfair' rewards for similar work, exist in primates (Brosnan & de Waal, 2003; Brosnan & Waal, 2002).[115] In fact, the biggest problem with Joyce's account of the evolution of morality-as-commitment is what I have also called its greatest strength, i.e., its insistence on the conceptual/cognitive component of moral judgment. What's good about it is that this component is very plausibly essential to what we take to be full-blooded moral judgments, in addition to its seeming role in strengthening and maintaining (perhaps even creating) our peculiarly moral commitments. What's not so good about it is that the concepts and beliefs involved in moral judgments are presented as direct products of genetic natural selection.

Joyce hypothesizes that in our case 'natural selection opted for' thinking of things as being 'desirable', of '*demanding* desire'. The thoughts that Joyce takes to be crucial to

[115] What exactly the refusal of capuchin monkeys to work for less than others are getting for equal work has to do with 'fairness' is an open question, and not relevant here. What's clear is that they *will* refuse to work for cucumber if another monkey is seen to be getting banana, though they will work for the cucumber if alone and if other monkeys in their sight are also getting cucumber. This is the kind of behavior that seems helpfully analyzed as an aversion- (or emotion-) backed commitment. But it is presumably quite a ways from a moral judgment.

moral judgments and emotions, such as ‘I would deserve punishment if I did that’ (80) and the thought of having ‘transgressed against a norm’ (112) are, at minimum, less likely to have resulted from purely biological evolution than are nonlinguistic committing *sentiments*, which plausibly exist not only in other primates, but dogs and other social animals. In my judgment, Joyce is right to focus on the problem of weakness of will in motivating a ‘personal commitment’ pathway to moral judgment. But rather than hypothesizing that biological natural selection directly produced the evolution of conceptually rich moral judgments (with or without emotional backing), I find an explanation that begins with the plausibly adaptive benefits of committing sentiments much more illuminating, as well as a much more likely product of natural selection. I think culture is likely to have provided a much more significant contribution to the cognitive/conceptual component.

I also noted that Joyce and Frank left out any account of the moral sentiments connected with moralistic punishment. Joyce did argue that having thoughts like ‘I deserve punishment’ would be more motivational than merely feeling badly about a violation. And he argued that claiming ‘it would be immoral to do X’ serves an interpersonal commitment function, since it reduces one’s effective options to X in the future. But neither author provided an account for why we would have such powerfully rewarding emotions and motivations geared toward punishing transgressors. I provided one that, whether depending on reputational and/or group-selection effects, and whether cultural and/or biological evolution was in play, shows how and why we could have evolved powerful sentiments motivating costly punishment of norm-transgressors.

There is one final observation that Joyce makes about his commitment hypothesis that he does not make much of, but that I consider a fundamental and even potentially unifying insight.[116] He notes that both kinds of commitment he has hypothesized 'have a paradoxical air, in that the benefit of the commitment[s] [are] attained only by *not* aiming deliberately at [them]' (123, italics his). The crucial point is that if you are consciously aiming at the benefits of the commitments, then you are thinking prudentially, but it is just the pitfalls of prudential thinking that the benefits of the commitments are designed to avoid. I made the case above that the subjective removal of options is an important cognitive component of human commitments. This commitment strategy is threatened if one is aware of it. Therefore we should not be surprised to find that *those (individuals or groups) in whom the strategy is most successful are those that have effective methods of remaining relevantly unaware.*[117]

The issue of lack of awareness of (the nature of) one's motivations will continue to play a central role in the remainder of this essay, so before I more to the next section, I want to say more about why I think it is an important part of the function of moral concepts and judgments.

Joyce thinks guilt involves the thought of deserving punishment. I doubt this, but I doubt much less that that thought is important in *sustaining* the guilt. Here's how I think it works. We can experience guilt, or something much like it, at having violated a

[116] Not that he is the first to make it. Ainslie and Frank both saw that the commitments in question are less effective if one consciously aims at them. Also, recall from section 1.2.3 that the solution to Ainslie's four motivational 'puzzles' all require that the attention be deflected away from the object of pursuit. By 'unifying' I mean that it might be able to unify what I have called (after Nesse) 'subjective commitments'.

[117] For reasons that I will discuss in Chapter 4 (that have to do with the downsides of will presented in Chapter 1), it is also likely to be the case that the commitment strategy will be *weakest* in people who are the most unaware of their motivations.

social norm[118] without the thought of deserving to be punished. But we do not experience the guilt as resulting from our perception of having violated a norm per se. The process of internalizing is, in part at least, the process of seeing the norms as not 'just' the enforced norms of one's society, failure to heed which *predicts* punishment. That is why Joyce thinks that (moral) guilt includes the thought that one *deserves* punishment.

Now the reason why I think this thought is important is that if attention is focused on punishment-avoidance, norm-adherence is likely to be less reliable, for reasons with which we are now familiar. As culture makes possible ever greater possibilities for reasoning, rationalization and deliberation, motives contrary to the social norms can exploit this reasoning to their ends. To the extent that one conceives of norms or rules as to be followed in order to avoid punishment, the more one will look for ways to escape punishment, while violating them when they are onerous. Where social norms strongly conflict with other aspects of our motivational psychologies (whether 'evolved' or not) is where this rationalizing process is the most likely to happen, and the risks associated with it increase with the presence of (especially moralistic) punishers.

Therefore if norms are in place which run strongly counter to other powerful motivations, where punishment of norm-violations is sufficiently reliable and severe, those who can internalize the norms are at a significant advantage, even if those norms have other fitness-reducing aspects. At the group level, a society with norms that run strongly counter to other motivations will have a much better chance of maintaining those

[118] Or, in fact, at having done things against which there is no social norm—we seem to have evolved so as to feel guilt (some of us at least) in response to 'defections', or failures to care for offspring, that can plausible survive the absence of norms in their favor.

norms if it also acquires cultural variants which aid the process of internalization. It is now a commonplace among researchers into social norms that the internalization of norms is accomplished in large part by 'moral emotions' like shame and guilt.[119]

But, as I've said, such internalization is threatened by deliberation. Suppose, as is plausible, that emotions like shame, guilt and anger can serve as perceptions of norm-violations. These perceptions are motivationally loaded. By hypothesis, the ability to perceive norm-violations in such a motivationally-laden way is an evolved capacity that allows us to avoid norm-violations, despite those norms often running directly contrary to other powerful, and often evolutionarily much older aspects of our motivational psychology. So, where these contrary motivations are present, there is (the potential for) motivational instability and therefore norm-violation. Therefore we would expect both individuals and groups to do better on average if there were a way to achieve stability in the direction of the motivations geared to norm-adherence.[120]

How might such stability be achieved? It is well-known that attention is motivationally important. One of the primary ways of not succumbing to temptation is to direct attention elsewhere. Therefore a mechanism that deflects attention away from contrary motivations will tend to preserve the stability of such motivations as guilt, shame and (moral) anger.

Now, what would we expect such an attention-deflecting mechanism to look like? A bad way to not think about a white elephant is to try directly not to think about it. A better way is to think of something else. Effective forms of meditation do not work by

[119] Less often cited, but equally important, is (moral) anger.

[120] Of course we would only expect indidivuals to do better if norm-violations are sufficiently punished and groups to do better if their norms were 'pro-social' in the sense of contributing to stronger groups.

thinking of nothing, focusing the mind on nothing at all, but on specific objects, such as (subtle) bodily sensations or mental images. With practice, all other things can be crowded out of attention, and out of attention do not generate the anxiety that it is in large part the purpose of meditation to ameliorate. Therefore we should expect a cultural variant that had the function of keeping attention away from our motivations[121] to not simply suppress that attention, but deflect it elsewhere.

In my view, *this is the primary function of (peculiarly) moral concepts and the judgments they issue in*. They play a committing role by deflecting attention away from our motivations and onto a normative realm which is apparently independent of those motivations. Divine commands still play this role for billions of people, but the central peculiarly moral concepts that do this kind of work for most of my readers are those of moral obligations/requirements, duties, rights, (im)permissibility, blameworthiness, responsibility, desert and the like.[122]

It is these concepts, and others suitably related to them, which supply the kind of 'objective normative force' which any naturalistic account of morality must explain and justify if it is to retain the peculiarly moral character of moral concepts. And yet all naturalistic accounts struggle mightily with just this task. My view solves (or dissolves)

[121] Notice it is not only *contrary* motivations, perhaps not even primarily contrary motivations that have to be not attended to. Rather it is the *nature* of one's motivations, specifically the relational, emotional and typically alterable nature of our values. Even commitments that on reflection we want to keep can be threatened by an awareness of their nature. But commitments we would not want to keep on reflection (aside from fear of punishment) are in bigger trouble.

[122] I think that the central and/or traditional connection between free will and moral responsibility lies in their shared commitment to our being able to act independently of our contingent motivations. The idea that we 'cannot do otherwise' undermines traditional conceptions of freedom and moral responsibility, and can undermine our motivation to punish transgressors. This is not a coincidence. Absent the idea that people deserve punishment because they are morally responsible or blameworthy, one is left with the often thin motivational gruel of straightforward consequentialist considerations. Although these are also given in moral terms, i.e., not in terms of what we want or care about, it takes attention away from the transgression itself, the perception of which is what generates the motivation to punish.

this problem by holding that moral concepts retain this (felt) normative force due precisely to their function of deflecting attention (and therefore investigation) from the very motivations that they are in the business of supporting.

According to D'arms and Jacobsen, 'The central naturalist thought in ethics is that the normative force of values must ultimately be located in motivational force, if it's not to be convicted of systematic reification error. The puzzle, of course, is how then to make sense of the seeming lack of contingency on moral motive that moral judgments possess' (D'Arms & Jacobson, 1994). Accomplishing this apparent lack of contingency on motivation is, I believe, *the essential function* of moral concepts and judgments. That is, the felt lack of contingency (or better, lack of felt contingency) on motivation has the essential function of stabilizing motivation toward adherence to moralized norms (which very much includes norms about how to feel). The moral concepts involved in such moralizing are essentially commitment devices that require for their operation a lack of awareness of the relational, perspectival nature of (specifically moral) value judgments.[123]

That, at least, is what I will continue to argue throughout this essay. Now I move to a fuller discussion of how this lack of awareness seems to be accomplished. Although I think moral concepts help achieve the relevant lack of awareness, the nature of our moral experience suggests that these concepts are contributing to a function that existed prior to their development.

[123] I haven't directly argued yet for the relational nature of values. That happens in the next chapter.

2.4 Moral Projectivism

2.4.1 What Projectivism Is and Why I Am Arguing For It

What it is

Having provided a (partial) answer to the question of *why* natural selection 'would come up with the trait of moral thinking' (123), Joyce moves to the question of *how* it might have done so. His answer is that it 'manipulated emotional centers' (125) in such a way as to '[gild] and [stain] all natural objects with … [internal sentiment]'.[124] That is, the emotional responses that we feel in response to particular objects, actions and events are 'projected' onto those things themselves, as if (for example) disgustingness were an intrinsic property of an object.

The brief characterization of projectivism just given implies that projectivism is to be understood as a causal account of moral experience. Simplistically, we have emotional responses to real or imagined actions or events, and those responses are presented phenomenologically such that the actions or events seem to have intrinsic properties (say, rightness or wrongness) that we perceive and at least sometimes respond to accurately or appropriately. Notice that nothing said so far implies antirealism about values, moral or otherwise. While projectivism is closely associated with anti- or quasi-realist views about (moral) value, the sense in which I am arguing for projectivism does not imply antirealism about value, moral or otherwise. It is a purely psychological, descriptive thesis about our moral experience.

[124] Hume ((1740) 1978: 167), quoted on 125).

Projectivism in my sense is just about moral experience being caused by emotions. But emotions and intuitions can be affected by principles and norms. If there are true moral principles or objectively valid moral norms, then our intuitive judgments could become sensitive to them, even if most people's are not sensitive to them and/or are sensitive to false or invalid ones. Some authors, especially psychologists, seem to assume that if the psychological thesis of moral projectivism is true, then that directly supports antirealism, or antiobjectivism about moral values. But that is false, since moral realism is entirely compatible with this, or presumably any other, purely psychological thesis.

However, what projectivism can do is *provide an explanation of why we would believe* there were objective or intrinsic (moral) values *even if there are not*. Therefore while projectivism cannot *directly* refute philosophical arguments about the nature of value, it can still be of great value in such philosophical discussions. Arguments to the effect that it just really seems like there are intrinsic values can be met with the rejoinder that those appearances are explained by psychological facts that make no reference to intrinsic value. I want to use projectivism precisely to this purpose, in addition to showing how projectivism, if true, fits nicely into my commitment model of moral judgment, focused as it is on the crucial role of our not being aware of (the nature of) our moral values and motivations.

The more cogently I can argue that moral judgment is in the commitment-business, and that lack of awareness of our own motivations is an important part of the commitment strategy that we've inherited (biologically and/or culturally), the better I will be able to meet the challenge posed above by D'arms and Jacobsen (1994) of how to make sense of the ‘seeming lack of contingency on moral motive that moral judgments possess’ (762). The other part of their challenge was to make sense of the discipline of moral discourse on the assumption that moral judgments depend on moral motive. I do regard this as a legitimate challenge, but I also think that it is important not to meet it too successfully.

I already said that I think that a lack of awareness of our motivations helps explain why it doesn’t seem that our values are contingent on our motivations. I also think that the tendency to misperceive our motives in the moral realm can help explain some of the discipline of moral discourse.[125] But just as importantly, I also think this tendency can explain its *lack* of discipline, specifically its staggering proneness to generate hypocrisy and self-delusion, which I think is much greater than is generally recognized. It also explains why the very serious downsides of moral thinking itself, like those of willpower, are very seldom recognized. That is because *in order to (fully) recognize their downsides, you have to recognize the business they are in—but the success of that business is threatened by such a recognition*. So the downsides tend to be invisible, much as we would expect the downsides of God-discourse to be invisible to those steeped in it, no matter how much they disagreed on the substantive questions of what God commands.

[125] Due to the ‘civilizing force of hypocrisy’ (Elster 2007, 406). This is the process by which one comes to believe that they have public interests at heart because in making public proposals they will have to advert to the public interest, not their own. But since people are pretty good at spotting liars, the best way to convince them that one has the public interest at heart is to convince oneself. It is ‘civilizing’ because in order to convince the public that one’s proposal is directed at the public interest, one will typically have to propose something that is actually more in the public’s interest than the course or policy that one would choose absent any need to convince the public.

Before I enter into this argument, an important distinction needs to be made. In questions of the metaphysics of value, perceptivism and projectivism are often opposed, where the latter is understood to entail antirealism, whereas the former, while not realist in the traditional sense involving mind-independence, is 'anti-antirealist'. The evidential strategy I will describe and pursue below supports projectivism understood as a causal account of moral experience, but it is perfectly consistent with perceptivism understood as a claim about the nature of (moral) value. Sentimentalist theories (theories which relate emotions and moral/evaluative concepts) are thought of as perceptivist, in that they consider moral properties to be secondary qualities, like colors (D'Arms and Jacobson 2006). To be able to see objects as colored may depend on specifics about our visual apparatus, but it still seems to make sense to say that we can perceive colors.[126] That is, many people think colors can be both mind-dependent and perceived. Likewise, our emotions play a crucial role in our experiencing the world in an evaluative way (McDowell 1987; McNaughton 1988), but this doesn't mean that we don't perceive these values.

I do think we often can perceive via our emotions what it would be right or wrong to do, if only because they help us perceive what we are likely to regret and how much.[127] But I think that the sense in which this is true is a sense which makes it misleading at best to call these perceptions *moral* perceptions. I agree with Prinz (2007) that there is something importantly right about both metaphors, but disagree that what is right about

[126] Though some disagree, e.g. Hardin (1988).

[127] The reader might be wondering here whether I mean to say that the fact that we will regret something is (at least often) what makes it wrong. In a word, yes, but elaboration and defense of this view will have to wait until Chapter 3. For now let me just say that we often have the goal of acting in such a way as to minimize regret. Relative to this goal, acting in a way that fails to accomplish it is wrong. And all reasons are relative to ends. Or so I argue in Chapter 3.

the perceptivist metaphor is profitably cashed out in moral terms. For now let me say that perceptivism is consistent with 1) emotions being the primary cause and/or (partially) constitutive of our moral judgments, and 2) those judgments having the phenomenology claimed by projectivism. It is these two properties that I am concerned with. For simplicity's sake I will call the view I'm arguing for moral projectivism. But it is 'only' a psychological thesis about how we make and experience moral judgments.

Why I Am Arguing For It

While I favor perceptivism over projectivism as a view about the nature of values, projectivism gets the phenomenology right, and that is important for me. Because while I think that we perceive values by means of our sentiments (and also that our values are largely constituted by our sentiments), it is a crucial part of the committing function of peculiarly moral discourse that we not recognize what it is that we are perceiving, but rather regard it as something independent of our own values and motivations.

We have just seen an argument that moral judgments and sentiments are commitment devices, though it wasn't airtight to be sure. But it's not over yet. The arguments for projectivism I will give below can contribute to the claim that they are commitment devices. That the projectivist phenomenology is so conducive to effective commitment, combined with independent and extensive arguments for the committing function of moral sentiments and judgment, suggests that that is what the phenomenology is in the business of.[128]

Joyce wants projectivism to show how natural selection managed to give us moral judgments. I'm not primarily concerned with whether or to what extent natural selection caused us to project. I want projectivism to help explain belief in intrinsic value and categorical imperatives in terms of our emotions (and/or desires or values). I also want to show how it fits into a commitment model by subjectively reducing options, specifically by obscuring the relational nature of (moral) value. Whatever the contributions of biology or culture and/or individual psychology to our projections, they can be profitably understood as (parts of) commitment devices, and contribute to a plausible explanation of our belief in 'practical clout' and intrinsic value.

2.4.2 Evidence and Argument for Projectivism

Phenomenology

The thesis that moral judgments are the result of a projection of our emotions has two main parts.[129] The first is that emotions are the (primary) causes of moral judgments,

[128] And even if it is not the teleological function of the phenomenology, it is very plausibly an instrumental one; an exaptation.

[129] For simplicity's sake, I've described 'moral projectivism' as the thesis above, which suggests that *all* moral judgments are the result of a projection of our emotions. I don't think that the best statement of moral projectivism should say this, because it strikes me as very implausible and unnecessary for the main point. Since I don't think that 'moral judgments' or 'moral' anything form a natural kind or a category with necessary and suffificient conditions for membership, I doubt very much that any simple theory or hypothesis as to their nature will capture only and all (what we call) moral judgments. I think the case for moral projectivism is (at least) strong enough for my purposes if the projections of our emotions (values, desires) onto the world accounts for a large range of paradigmatic moral judgments. Further, the significance of the thesis will be strengthened if it can be shown that the motivation to act in accordance

and the second is that the emotions have the phenomenology associated with projectivism, i.e., that they seem to be responses to properties which are themselves independent of those emotions (or emotional dispositions). Therefore in order to provide evidence for moral projectivism, we must provide evidence for both parts of the thesis.[130]

I'm not going to argue that we have the phenomenology. Joyce gives evidence (Nichols, 2004; Nichols & Folds-Bennett, 2003)that children 4 – 6 years old distinguish between properties that depend on human preferences (like *yummy, icky and boring)* and those that do not (like 'good' associated with helping and harming). There is also robust cross-cultural evidence that people, including young children, distinguish between norms that depend for their validity on an authority and those that do not (evidence provided on 136 – 7). This lack of dependence includes even that of God's authority in Larry Nucci's (1986, 2001) study of Mennonite and Amish children, 100% of whom said working on Sunday would be fine if God said it was, while 80% of them said stealing would not be, whatever God had to say about it.

The fact that children as well as people in other cultures have this phenomenology, if they do, is important for Joyce since his goal is to provide an evolutionary story. But while this kind of evidence is suggestive of a projectivist phenomenology, it is indirect evidence at best. Children making the moral/conventional

with at least many of the non-projected moral judgments is parasitic on the class of projected judgments. I also hasten to add that the thesis as I see it does not entail the absence of effective rational thinking in arriving at or justifying moral judgments.

[130] This evidential strategy is from pp. 128 – 9 of Joyce (2006). However, Joyce goes on to say, 'Of course the thesis is strengthened if we can show that there is no good reason (independent of the phenomenology or its associated language, which is given a debunking explanation) to think that we are in fact detecting independent moral properties'. I don't see this *strengthening* the thesis, since the thesis as Joyce (and I) described it doesn't entail that one does not detect such independent properties in the first place. In other words, I don't think projectivism as described entails antirealism, or even response-dependence. It is rather a causal account of moral experience, consistent with a variety of other metaethical views.

distinction can be explained by their having a different phenomenology, but that is not the only possible explanation. Further, though explanations involving acculturation (without projectivist phenomenology) are less likely for young children than adults, children are not the canonical makers of moral judgments. An adult who denies that she has the relevant phenomenology, yet believes that she makes (paradigmatic) moral judgments, will not likely be convinced that (paradigmatic) moral judgments are nevertheless the result of projected emotions based on evidence from children, who have simply yet to mature out of their projecting ways. And of course someone who recognizes the phenomenology is in no need of convincing.

The best I think we can do is to describe the phenomenology (and the associated concept(s)) we're interested in and let people judge whether they know what we're talking about. Also, by employing the terminology 'denies' vs. 'admits' above, I do not mean to suggest that everyone really has the phenomenology but only some admit it. On the contrary, I think the phenomenology, whatever its cause, is alterable.[131]

In the meantime, let's describe briefly what we're talking about. The ubiquitous analogy is to color perception. When most people see something as red, they conceive of and experience that object as having the intrinsic property of redness, as opposed to 'redness-to-me'. Likewise, when most people (at least most people you're likely to have a philosophical conversation with) see or hear about someone getting tortured, they typically experience the action and the torturer as horrible, despicable, revolting, grotesque, and so on without the 'to-me' or even 'to-people-relevantly-like-me' component. Though the moral example is charged with emotion and potential motivation

[131] In fact I'm going to recommend altering it as an important part of moving beyond morality.

in a way that the color one is not, they both seem to have in common the feeling that something in the world, something external to one's own feelings or dispositions, is being perceived (and then in some sense 'appropriately' responded to, at least in the emotional case). The crucial aspect of the phenomenology (and associated conception) is that someone who is not, say, horrified, would be missing something important and/or have something wrong with her, even if this someone were the subject herself. At this point, you either know (well enough for my purposes) what I'm talking about, and acknowledge that at least many such judgments have this phenomenal character (whatever the reason or etiology), or not, and I don't think that further resources expended on this issue will be sufficiently repaid.[132]

As I indicated, having the phenomenology is a far cry from endorsing projectivism, since one could have the phenomenology as a result of perceiving (even mind-independent) moral properties by means other than emotional projection. In order to argue for projectivism, I'll first run through some of the fast-accumulating empirical evidence linking emotion and moral judgment. This will be useful evidence that our commitment model was not on the wrong track at least insofar as it suggested that subjective emotional commitments are crucial to the evolution of moral judgments. Then

[132] Perhaps the reader will wonder what the contrasting phenomenology is supposed to be. This will be discussed later in more detail. But for the time being, it is the sense one has that the property is relational, which directs part of one's attention to an aspect of oneself as constituting one part of the relata. This way of thinking/feeling is most clearly brought out in the context of stark contrasts between our own and others' felt responses to properties that we have little to no difficulty regarding as relational. A primatologist might describe a female baboon presenting her backside to a male baboon as very sexy to him (the baboon). We can both believe this and feel not the slightest hint of the response that we imagine (or at least believe) him to have. We can believe, and I think feel, that 'sexy' is relational. In contexts of such stark contrasts, it's easiest to recognize (and feel) the contribution our dispositions are making when we judge someone sexy. At other times, it might seem that sexiness is an intrinsic property of the perceived object.

I will discuss the work of Jonathan Haidt, which provides some of the best evidence available that the *direction of the link* is from emotion to moral judgment.[133]

Evidence for a Strong Link Between Emotion and Moral Judgment

Surely the most cited evidence for the importance of emotion in moral judgment is the case of Phineas Gage, who in 1848 received an iron tamping rod through his ventromedial frontal lobe in a dynamite-induced explosion. The 1 ¼ -inch rod went in underneath his left eye and came out of the top of his skull. Astonishingly, he not only survived the accident, but appeared to have completely recovered soon after the accident, save for blindness in his left eye.[134] However, the once-industrious, respected and virtuous Gage was now ‘anti-social’; he was impulsive, given to immoderate profanity (which he did not use before), disrespectful, confrontational, inconsiderate and irresponsible. This damage, as saddening and upsetting as it was to those who knew him (as well as Dr. Harlow, who treated his wounds and recorded his recovery and ‘progress’), might not have been very surprising or noteworthy if not for the fact that much of his other reasoning and linguistic faculties were intact. He also remained strong and physically coordinated. The damage was remarkably ‘selective’; it seemed to have affected only his practical reasoning, especially with respect to operating smoothly and

[133] Both Joyce and Greene (2002) follow the same strategy in arguing for moral projectivism. Some of the authors they cite (e.g., Blair, Damasio, Haidt) I was familiar with and would have figured into my own argument for moral projectivism. Others I thank Greene and Joyce for providing. Almost all the studies that Joyce cites appear in (the earlier) Greene. It would be cumbersome to acknowledge Greene and/or Joyce for all the studies individually, so I will just cite the original authors, and acknowledge Joyce and Greene here for their help in putting together the inventory.

[134] Damasio (1994) gives the definitive account, pp. 3 - 10.

successfully in the social realm (though all forms of long-term planning were also seriously compromised if not ruined).

Subsequent patients with similar injuries have led to greater understanding of the nature of Gage's drastic change in personality. Damasio describes his patient 'Eliot' who suffered similar injuries to Gage's (due to surgical removal of a tumor). Eliot's capacities to reason about politics, the economy and business were seemingly intact. He performed above average on intelligence tests and had a normal showing on personality tests. But like Gage, he became irresponsible and made very poor decisions in the practical and especially social domain (though he was not crude and cantankerous as Gage had become—Damasio found him witty and engaging). Soon he was without the job, wife and respect of family and friends he had had prior to the onset of the tumor (though fortunately for him I suppose, the damage also seemed to prevent him from being bothered by his fate).

Eventually Damasio gave him a test designed to explore emotional responses and found the root of Eliot's problem. When showed pictures that typically elicit emotional reactions in people, such as bloody injuries, imminent drownings, earthquakes, etc., Eliot did not have any emotional reaction, but recalled that he would have had them in the past. To rule out the possibility that Eliot had lost access to the 'rules and principles of behavior that he neglected to use day after day' (46), Damasio and colleagues ran a wide range of tests on Eliot to assess his ability to describe what sorts of actions would be appropriate in different social circumstances, what sorts of means conduce to a variety of ends, and what kind of actions and justifications for those actions were appropriate in sociomoral contexts. He performed average or better on all tests. After coming up with an impressive array of genuine potential solutions to a social dilemma, Eliot declared, '"And after all that, I still wouldn't know what to do!"' (49). Based on Eliot's reports of his lack of emotion and Damasio's observations of his very 'cool' unemotional style, he hypothesized that Eliot's problems were the result of an emotional defect which had left his 'higher reasoning' abilities intact. Subsequent tests have confirmed this diagnosis.

These two cases (since Gage's injuries were similar to Eliot's) suggest that emotion plays a crucial role in both prudential and socio-moral functioning. Studies of sociopaths and psychopaths generate even more striking findings about the relationship between emotion and moral judgment and behavior. Damasio, Tranel and Damasio (1990) found that people with ventromedial damage in adulthood no longer had normal emotional arousal (as measured by electrodermal conduction) to socially meaningful stimuli, but retained it for nonsocial emotionally charged stimuli. Their behavior subsequent to their injuries has resulted in what has been termed 'acquired sociopathy'. In a fascinating study, Anderson et al. (1999) showed that two subjects with early (three and fifteen months) ventromedial damage developed 'psychopathic' behavior, including remorseless theft and violence. Interestingly, these early-onset patients lack explicit awareness of moral and social norms, unlike those with late-onset damage. While this could suggest that the damage affected the parts of the brain responsible for such declarative knowledge, the hypothesis that Anderson and his colleagues propose seems at least as likely—that the inability to experience emotion in early life interfered with learning the explicit norms and rules.[135]

[135] It is worth noting that all the ventromedial patients discussed so far had marked inabilities to pass up short-term rewards or desires for greater ones in the future (Joyce 2006, 124).

The seeming requirement for emotional responses to aid in sociomoral decision-making is manifested in other (non-ventromedial-damaged) psychopaths as well. Two studies (Blair, 1995; Blair, 1997) showed that psychopaths, while retaining their general rational capacities, had decreased emotional reactions (again, measured by electrodermal conductance) to normally distressing stimuli, compared to control prisoners who were, like the psychopaths, convicted of murder or manslaughter (their emotional responses to threatening and neutral stimuli were the same as the controls). The psychopaths also fail to distinguish moral from conventional transgressions, again, unlike control prisoners. Rather than consider all transgressions conventional, they do the reverse. They treat all norm-violations as independent of authority (which in all cases was a teacher in a school) and have a much greater tendency to explain the wrongness of the violations in terms of rule-violations than harms inflicted. Given the sub-normal electrodermal responses to distress cues, these results are most plausibly explained by the psychopaths not having the (greater) emotional responses (especially empathy) to what we consider moral violations than to conventional ones.

Finally, fMRI studies indicate that ‘emotional centers’ in the brain are particularly active when people are asked to make decisions in morally ambiguous situations, especially where the decisions are ‘personal,’ such as involve directly harming another person (Greene et al. 2001; Greene and Haidt 2002; Moll 2002, 2003). Emotion centers are also more active than normal when subjects punish opponents for defecting in the ‘Ultimatum Game’ (Sanfey et al. 2003).

The evidence presented above suggests a strong link between moral judgments and behavior, and emotional capacities and responses. But it does not directly address the question of the causal relationship between judgments and emotions. Fortunately there are several sources of evidence that the causal arrow generally runs fom emotions to judgments.[136] Jonathan Haidt has marshalled evidence from various sources to support this claim, including research of his own that has led to his Social Intuitionist Model of moral judgment. I want to give a sketch of the evidence for the importance of emotion-backed intuitions in moral judgments, then discuss the relevance of some of Haidt's findings.[137] My aim will not be to defend the specifics of Haidt's model, but rather to use his work to bolster the case for the centrality of (especially subjective) commitment in moral judgments.

The primacy of affective intuition

I'm going to start with the empirical evidence Haidt provides to be skeptical of the causal importance of reasoning in moral judgments as compared to that of affect-backed intuition. Before addressing the relative roles of reasoning and intuition, it's worth getting clear on what Haidt means by these terms:

> The words "intuition" and "reasoning" are intended to capture the contrast made by dozens of philosophers and psychologists between two kinds of cognition. The most important distinctions are that intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, while reasoning occurs more slowly, requires some effort, and involves at least some steps that are

[136] The evidence could also be read as saying that the emotions constitute the judgments (Prinz 2007).

[137] I'll be drawing on Haidt's (2001) for a lot of the evidence of the role of intuition and emotion in moral judgments, as well as a description of his own model. All page numbers are to this work unless otherwise noted.

> accessible to consciousness (818).[138]

An 'affective valence', either positive or negative, is included in these intuitions, often but not necessarily accompanied by or including emotion. Intuition is to be contrasted with reasoning but not with cognition. Intuitions are cognitions, and Haidt sees his account as employing the 'affect as information' hypothesis, which is drawn from research showing that people use transitory moods and feelings to make decisions and judgments (G. Clore & Schwarz, 1994; Loewenstein, Weber, Hsee, & Welch, 2001; Schwarz & Clore, 1983).

The first reason Haidt gives to doubt the causal importance of reasoning in moral judgments is what he calls the 'dual-process problem' (819). Rationalist psychology researchers as well as philosophers have long-focused on the reasoning process. Haidt on the other hand draws attention to the fact that by the beginning of this century social psychologists, taking their lead from researchers such as Zajonc (1984), Bargh (1994), and Greenwald and Banaji (1995), had come to see a great deal of behavior and judgment as 'automatic', i.e., lacking conscious awareness, effort or intention (819). This research followed closely on the heels of Zajonc's (1980) synthesis of evidence from several fields to modernize Wundt's (1897/1969) 'affective primacy theory'. Zajonc's conclusions were that conscious reasoning and affective states are subserved by distinct systems in the brain, and that the affective system has 'primacy' in every way. It is first phylogenetically and ontogenetically; it makes quicker judgments in real-world situations and those judgments tend to persist in the face of conflicting judgments from conscious

[138] These definitions of 'intuition' and 'reasoning' correspond fairly closely to Hume's terminology in Book III, Part 1 of *A Treatise of Human Nature* where he begins his famous and influential discussion of the relationship between 'reason' and 'passion' in morality.

reasoning. Haidt argues that this ‘intuitive’ system is ‘ubiquitous and understudied’. (819).

The second source of skepticism comes from the surfeit of evidence that much of our reasoning is motivated and biased in the direction of conclusions we antecedently wish to reach. He says that our reasoning processes are more akin to lawyers defending their clients than a ‘scientist seeking truth’.[139] Haidt surveys numerous studies that support a broad spectrum of motivated thinking in the moral domain. He cites two classes of motives that ‘bias and direct reasoning’, namely ‘relatedness’ and coherence’ motives. First, people are strongly motivated to ‘agree with friends and allies’. Second, people are strongly motivated to avoid or resolve cognitive dissonance, especially with respect to their attitudes about themselves, their ‘cultural world views’ and moral commitments (821). Quite a lot of evidence is presented to indicate that we use motivated reasoning to reach conclusions we are antecedently motivated to reach. Kunda’s (1990) review argues that this motivated reasoning works by searching only for evidence in favor of preferred conclusions. Pyszczynski and Greenberg (1987) argue for a process in which biases operate throughout the reasoning process, from the selection of initial hypotheses to inference-making to searches for and evaluation of evidence and finally, how much evidence is required before reaching a conclusion.

The details of this work are interesting but the general conclusions are, I dare say, unsurprising. Most of us are aware that people often reason in biased and motivated ways so as to reach or support desired conclusions, though perhaps we don’t know the

[139] Haidt is apparently more sanguine about individual scientists’ impartial search for truth than some others.

extent to which it happens or (all) the kinds of strategies we use. Less intuitive is the extent to which we come up with post-hoc explanations for our behavior, moral or otherwise. This is the third source of skepticism--post-hoc justifications of intuitive judgments which cause the illusion of objective reasoning (822). Nisbett and Wilson (1977) convincingly argue that people often provide post-hoc explanations for a wide variety of behaviors. Our 'introspections' seem like, and perhaps in some sense really are, searches for the actual cognitive processes that caused the behavior in question; but those processes are (at least often) not consciously accessible. What people can and do access is a 'pool of culturally-supplied explanations for behavior' (Nisbett and Wilson 1977, p. 248). In effect, they generate a hypothesis about why they would have done something, then generally search (only) for evidence in favor of that hypothesis, and stop once they've found the evidence (Kunda, 1990; Pyszczynski and Greenberg, 1987).

Subjects engage in post-hoc confabulation when hypnotized (Zimbardo, LaBerge, & Butler, 1993) and when their behavior was affected by subliminal presentation (Kunst-Wilson & Zajonc, 1980). Subjects reliably and readily provide false but plausible reasons for their actions. And the most striking examples of confabulation are the split-brain patients, whose neuronal connections between their brain hemispheres have been severed. These people's verbal centers in their left hemispheres easily and unknowingly create explanations for what their left hands are doing as a result of stimuli presented to the right hemispheres, which control the left hands (Gazzaniga, Bogen, & Sperry, 1962). The patients are so proficient at this, and their explanations so resemble the confabulations of normals, that Gazzaniga (1985) has hypothesized that we have an ‘interpreter module’ in our left hemispheres that generates rationalizing explanations for behavior the true causes of which it simply cannot access.[140]

One of Haidt’s central claims is that post-hoc reasoning about our behavior is also quite common in the moral domain. Just as people draw on culturally-provided explanations for their behavior in the nonmoral domain, they draw on ‘culturally-supplied norms for evaluating and criticizing the behavior of others’ (822). Because of the close relationship between people’s moral judgments and their justifications, researchers have assumed that the reasons people give in their justifications are the causes of the moral judgments they make. But if people lack awareness of the true causes of their judgments, these reasons are likely post-hoc rationalizations, and it is more likely that the judgments give rise to the justifications.[141]

The final kind of evidence with which Haidt supplements his case is that there is a weak link between moral reasoning and moral action and a strong link between moral emotions and moral action. Haidt contends that the positive correlations that Blasi (1980) found between moral reasoning and moral action and Kohlberg (1969) found between moral judgment tests and I.Q. is not to be interpreted as reasoning causing judgment. He draws on Metcalfe and Mischel (1999) to argue that the correlations are most plausibly explained by the increased self-regulatory, or self-control mechanisms of the higher-I.Q.

[140] Gazzaniga considers it a strict inability, but I think very often the lack of ‘access’ is motivated and corrigible.

[141] Again, I think as a general claim, this is probably true, but that does not show that norms and principles do not sometimes or often play an important role in generating the intuitive judgments.

subjects. Frontal cortex is thought responsible for both impulse-inhibition and the kind of intelligence that shows up on moral reasoning and I.Q. tests.[142]

Adding to the plausibility of this explanation is the fact that in Blasi's (1980) review, moral reasoning 'was most predictive of negative morality—refraining from delinquent behavior' (824). Haidt also mentions the long-standing inverse correlation between criminality and I.Q., even when controlling for socioeconomic status (Hirschi & Hindelang, 1977).[143] The evidence of correlations between moral reasoning and 'positive morality', i.e., actively helping others, is much less clear. Some studies show a correlation while others do not, though the correlation with negative morality remains.[144]

I'll be much briefer in presenting the evidence for the strong link between moral emotions and moral action. The bulk of it concerns the work I discussed above regarding psychopaths and Damasio's work on people with damage to VMPFC, and their behavioral problems that seem to result from lack of emotional/affective responses, though they retain discursive moral and practical reasoning abilities. The remaining evidence can be summarized by saying that whatever the specific hypothetical mechanisms, researchers in the field are in agreement that in the (not unusual) cases in which people help each other, they are motivated primarily by affective states, including

[142] Haidt says that the best interpretation is that the common cause of both better reasoning and better behavior is intelligence. But I think this is misleading. At minimum, there are two kinds of 'intelligence' if the capacity for impulse-inhibition is to be considered a kind of intelligence per se. There is the kind that can produce moral (and presumably other kinds of) reasons, and the kind that inhibits impulses. These might be connected, but as the case of Eliot and many others show, they can come apart.

[143] He doesn't explicitly mention the other well-documented strong correlations between impulsiveness and criminality I have cited above.

[144] It's worth noting that the correlations are based on scores on tests such as Kohlberg's scale, which are of dubious merit at best. Most importantly, the high scorers on Kohlberg's scale (5 and 6) regard morality as based on legalistic contracts and "principles of choice involving appeal to logical university and consistency" respectively (1984, 44). It shouldn't be surprising if such people are more inhibited than others, but it doesn't follow from this that they are better moral reasoners, *pace* Kohlberg.

empathy, reflexive distress, guilt, shame and sadness (Cialdini, 1991). To argue that emotions motivate behavior seems superfluous.

What I want to do now is give some of Haidt's own oft-cited evidence that affective intuition drives moral judgments and that private (as opposed to social) reasoning plays a relatively rare role in causing moral judgments, but it is much more commonly in the business of acting like a lawyer in that it seeks justifications for judgments that are made via emotional intuitions.

Here's a story Haidt gave to some subjects:

> Julie and Mark are brother and sister. They are travelling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decided that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love but decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love? (Haidt, Bjorklund, and Murphy, 2000)

Haidt reports that most people unhesitatingly reply that it was or would be wrong for them to make love. When asked for their reasons, they give justifications that are purposefully blocked by the presentation of the story, such as the likelihood of the children being deformed or the harm that will come to Mark and Jill's relationship. When it's pointed out to them that those justifications don't apply in this case, they do not change their views, but many end up claiming to know that it is wrong while admitting they cannot explain why.[145] The fact that they stand by their judgments while willingly abandoning their reasons is good evidence that their reasoning processes were not the causes of those judgments.

Notice that the justifications used above were drawn from the culturally-acceptable repertoire of justifications, in this case those of harm to self and harm to others. The ostensible requirement to both stick to one's initial moral judgment and draw from available justification norms appears to cause people to find victims to apparently victimless crimes. For example, some partipitants in another study claimed that cleaning the toilet with the national flag held out the possibility of harming the person doing so (Haidt, Bjorklund and Murphy, 2000). Haidt thinks that these cases are examples of a wider phenomenon that we were probably already familiar with, which he calls 'moral dumbfounding' and defines as 'the stubborn and puzzled maintenance of a judgment in the absence of supporting reasons,' (Haidt, Bjorklund and Murphy, 2000).

In 2005, Wheatley and Haidt hypnotized (highly hypnotizable) participants to experience disgust when they heard the word 'take' or 'often'. Then they were to read short fictional episodes, some involving morally evocative scenarios, some neutral, and some using their 'disgust word' and others not. The stories which did involve moral violations were responded to with greater condemnation when disgust was prompted. Incredibly, even in stories such as the one below with no moral valence whatever, subjects in whom disgust was triggered frequently regarded the fictional character as

[145] It might be thought that the stipulation that no harm came to Mark or Jill is unrealistic and that the subjects were responding to this. They might have been responding to it in an intuitive way, but presumably not in a reasoning way, since they would simply have had to say that they did not find the scenario plausible. Also, if the judgment were the result of such reasoning, they would not also offer harm to the child as a reason, since the blockage of that justification is not at all implausible. Further, their confidence in their justifications were clearly not as strong as that of their judgments.

morally dubious or worse:

> Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He [tries to take] <often picks> topics that appeal to both professors and students in order to stimulate discussion.[146]

Subjects whose version included their disgust word typically regarded Dan as a 'popularity-seeking snob' or as seeming 'so weird and disgusting' or as being 'up to something'. Others admitted that although they didn't know why what he was doing was wrong, this did not prevent them from claiming that his manner of scheduling discussion topics was wrong. It could be maintained that there is something unusual about highly hypnotizable people that (partially) explains this, but I find it just astounding that people (what else can we call it?) projected their emotions onto the character in the vignette despite the clear lack of anything remotely unwholesome in his behavior.

Here ends the case for projectivism. As Joyce puts it, [the] evidence concerning … what really causes moral judgment and … how it seems to us virtually adds up to a statement of moral projectivism' (130).[147] Again, I mean this only as a causal account of moral experience. The account explains how, in terms of emotional projections, we could have the experience of perceiving intrinsic (moral) value even if there weren't any.

Though I will no longer be using Haidt to argue for projectivism, there are still parts of his model and the evidence he marshals in favor of it that I find useful for my larger project (if sometimes to disagree). First and most importantly, I want to address the role of reasoning in the social intuitionist model. Up to now, I have given the

[146] The bracketed words are the target words, one given to some subjects, the other to the rest.
[147] Though he does not go into the detail I do about Haidt's model.

impression that Haidt thinks there is little to no role for reason to play in moral judgments, but that is not so (though the role he assigns it is unlikely to satisfy any rationalist). Second, there is more to say about how intuitions are formed, especially with regard to the role of culture. All this is intended to contribute to a subjective-commitment-centric account of moral judgment. But I want to get all the material on the table first, then show how we can understand both the intuitive and reasoning components of the model in terms of their role(s) in generating and/or maintaining commitments.

2.5 Expanding the Commitment Model of Moral Judgment

2.5.1 Reasoning in the Social Intuitionist Model

So far we've focused on two processes (or 'links' as Haidt calls them) out of 6 that Haidt incorporates into his model. The first two are the intuitive judgment link and the post-hoc reasoning link. We are now familiar with these. The next two combine with the first two to form the four primary links of the model as Haidt sees it (818). They are the reasoned persuasion and social persuasion links. The (dubiously named) reasoned persuasion link operates when a person communicates the post-hoc justification they have come up with for a moral judgment to others. Haidt thinks these justifications can influence other people through the content of the justifications, but regards this as relatively uncommon, due to the rarity with which moral argument or discussion results in persuasion. The social persuasion link is hypothesized as a result of people's extreme sensitivity to the norms of their communities, friends and allies. When a member of a (especially a close-knit) community makes a public moral judgment, that fact in itself can influence the judgments of others even in the absence of any reasons being provided for the judgment (Berger & Luckmann, 1967; Davis & Rusbult, 2001; Newcomb, 1943; Sherif, 1935).

The following two links are the only ones that give a role to what might traditionally be thought of as 'genuine' moral reasoning, where the reasoning is playing a significant causal role in generating the judgment. I'll discuss them in reverse order (6, then 5 as Haidt numbers them), beginning with the private reflection link. One might come to have conflicting intuitions about a moral situation by, for example, coming to see the situation from a point of view that one had not yet considered. Cognitive developmentalists including Piaget (1932/1965) and Kohlberg (1969, 1971) have cited perspective-taking as a primary form of moral reflection. In adopting another's (or just another) perspective, emotional responses can change immediately and markedly, and the availability of multiple perspectives can generate multiple conflicting intuitions. In some cases the strongest intuition will carry the day, but in others the reasoning process can play an important role in forming a stable judgment. This might be due to a 'conscious application of a rule or principle' (819) or more generally, because the reasons one has available to draw on are more successful in building a case for one intuition than the other(s). That one will begin to seem (or 'feel') right, and once it does, the search for additional competing reasons will tend to stop (829).

Finally, there is the reasoned judgment link. This is the quintessentially philosophical reasoning process, whereby one may reason oneself to a conclusion in contradiction to one's initial intuitive judgment. There is evidence that this kind of reasoning is quite rare, with the exception of philosophers (!), who 'have been extensively trained and socialized to follow reasoning even to very disturbing conclusions' (829). He cites Socrates and Peter Singer as examplars of this capacity. I think all these links can be illuminated in the light of commitments. But to make this task easier for me, I'll need to include sketches of how reasoning often works by means of embodied metaphors as well as that of the role of culture in forming intuitions and making them seem like perceptions of mind-independend realities.

2.5.2 Embodiment and Culture:[148]

As we saw, hypnotized subjects seemed to project their emotions of disgust onto Dan in the form of making a moral judgment about him. Striking as this is, there was clearly some confusion and uncertainty on their part as to why they should be thinking (feeling) this way about Dan. When they looked to what they knew about Dan from the vignette, they could find no properties or actions in him to support what plausibly seemed to them as their perception of some property in him--call it disgustingness. We might say that their post-hoc reasoning resources were maximally thin in this case, in that there were no norms which they could draw attention to his having violated ('popularity-seeking' was one such attempt however).

This is a highly artificial situation. Surely the overwhelming majority of moral judgments are made where there is at least some publicly-accepted norm that is not

[148] Much of the following discussion is from pp. (825-7) of Haidt (2001).

obviously and absurdly misapplied. Especially in the context of disagreement about judgments, it is reasonable to think that the availability of such norms has a stabilizing effect on these judgments. Though the study did not test this, it is plausible to think that the judgments made by the hypnotized subjects are not (nearly) as stable as those judgments in which there is a means of justifying one's judgments to oneself and others.

In this section I'm going to use the results of some important recent work in anthropology, neuroscience, moral psychology and cognitive science to begin an account of how the emotion-backed subjective perception of intrinsic value acts as a committment mechanism, which is aided by the explicit rules and norms available to members of a society. I will then go back through the S-I model and attempt to show that it can be seen as providing mechanisms for commitment at multiple levels.

In outlining the distinction between intuition and reasoning above, I mentioned the 'affect as information' hypothesis. Haidt notes (825) that Damasio's 'somatic marker hypothesis' fits well with this idea, in its thesis that many (or most) of our experiences are due to (or identical with) changes in the body that our brains perceive and feel. Over time and with repeated exposure (training), brain regions which perceive these bodily changes come to respond when relevantly similar situations arise. Eventually, the mere thought of an action or experience of a particular kind can occasion an attenuated version of the feeling that the brain has come to expect from experience (in addition to whatever innate responses might be present).

After years of studying patients with damaged ventromedial prefrontal cortices, Damasio proposed that the role of the VMPFC is to integrate these felt 'somatic markers' with the other decision-making resources (knowledge-base, goals, particular circumstances) in order to make effective decisions with limited information and time. When this system is matured, the input from the VMPFC (as well as other areas) is fast and easy (in the sense that no subjective effort is required). The results of this activity are generally consciously available in the form of an affective state, but its processes are inaccessible to (or at least not in fact accessed by) consciousness. In a word, it is intuitive.[149]

Lakoff and Johnson (1999) expand the role of experienced perceptions in our thinking by arguing that our bodily experience underlies much of what they call 'embodied cognition'. Our physical and emotional experiences seem to provide a kind of scaffolding or template for thinking and talking about many complex topics such as war, love and morality. This helps to explain why so much moral reasoning uses metaphor and/or attempts to categorize the action(s) in question with other kinds of actions that already have strong emotional associations either for or against. Such strategies are surely far more common (whatever their merits) than those employing a series of purportedly true premises leading validly to a therefore necessarily true conclusion.[150] Saying that abortion and/or eating animals constitutes murder is a nice example. Murder is highly negatively valenced, somewhat conceptually rich and has clear implications for action, mostly having to do with avoiding committing or supporting it. Of course categorizing restrictions on abortion (or eating) as instances of governmental intrusion on personal liberty have all the same general characterisitics, but accepting this latter

[149] The study mentioned above in which participants were hypnotized to feel disgust, and then reacted with moral judgments, is (by hypothesis) an example of the direct manipulation of somatic markers giving rise to moral judgments.

[150] Though these styles of argument are by no means mutually exclusive. In fact, they seem very common in normative ethics. I'll argue that Singer has had great success with blending these styles.

categorization leads to contradictory normative views than if one accepts the former categorization. I'll have more to say about this later. For now I will agree with Haidt in that such arguments do count as reasoning, but they nevertheless operate by attempting to induce (emotion-backed) intuitions in their audiences.

Now I want to turn to anthropology and moral psychology to fill in a picture of how these intuitions are developed and how they come to have the perceptual phenomenal character that they do. I'll begin with Schweder et. al's (1997) theory that cultures worldwide seem to be organized around three general kinds of ethics, which all cultures seem to accept to some degree, but which vary considerably in their relative importance from culture to culture. They are an ethics of autonomy, an ethics of community and an ethics of divinity.[151] Each kind of ethics is said to reflect, or create, its own type of moral 'goods'. The first is centrally concerned with individuals conceived of as such, and its associated goods are individual rights, freedoms and welfare. The second focuses on the goods of collectivities as such, whether families, tribes or nations, with the corresponding ethical goods of loyalty, duty, honor and self-control. These two ethics and their associated goods are probably most familiar to philosophers and secular Western society in general. The final ethic is concerned with 'the spiritual self', and its quintessential goods are piety and purity, both mental and physical. The theory plausibly holds that infants are born with the capacity to fully develop all three kinds of ethics, but that local cultures tend to emphasize one or two of them.

These three kinds of goods, and the norms that exist to protect them, are

[151] For reasons that it would be distracting to elaborate, I agree with Prinz (2007) that this last ethics seems better described as that of 'the natural order' (73). I will refer to it as such henceforth.

associated with three distinct emotions (Rozin et al., 1999). Violations of norms of autonomy tend to inspire (moral) anger, violations of community norms tend to elicit contempt, and transgressions against the natural order ('divinity') occasion (moral) disgust.[152] As it happens, the typically elicited emotions start with the same letter as their corresponding ethical good. This is therefore conveniently called the CAD model.

There is excellent reason to believe that moral disgust derives from physical disgust, as it has similar bodily responses and employs a similar logic of contamination (Rozin et al., 1993). The details of this needn't detain us here, beyond noting that the experiences associated with food that can become spoiled provides a model for the concepts of purity and cleanliness as goods in the physical, and then by extension, moral domains. These attitudes differ by culture of course, but many cultures have strong 'purity-based' ethics in which children (especially women) begin life pure but can be contaminated and corrupted by mere exposure to sex, violence or the devil (Haidt, Rozin, McCauley, & Imada, 1997). The contaminated persons are either difficult (e.g., exorcism) or impossible (loss of virginity) to make pure again, reflecting the fact that food, once spoiled or contaminated, is also very difficult to make clean again.

Now I want to recapitulate a description of how children in Orissa, India, a culture that emphasizes conceptions of purity and (therefore) contamination, come to perceive values in their world. A child growing up in Orissa is exposed daily to objects, their arrangements and people's behavior with regard to them that are structured by the concepts and rules of purity and contamination. Without requiring being told, the child

[152] These are, to be more precise, the emotions that are felt when the transgressor is someone other than oneself or a loved one (this isn't to say that these emotions never apply in the latter cases, only that they are not paradigmatic).

sees that 'foreigners and dogs may be allowed near the entrance to a temple complex, but only worshippers who have properly bathed are allowed into the central courtyard (Mahapatra, 1981). In the inner sanctum, where the deity sits, only the brahmin priest is permitted to enter' (827). In households, the room where the household god is kept, as well as the kitchen, are treated as more pure and clean than other parts of the house, and the human body is graded along a purity/pollution scale, with the head at the top and feet at the bottom of this scale.

By both imitation and instruction, children learn when to remove shoes and how to signal and interpret signs of deference, such as someone placing their head at another's feet. The cultural knowledge they gain is 'a complex web of explicit and implicit, sensory and propositional, affective, cognitive and motoric knowledge' (D'Andrade, 1984; Shore, 1996). They develop an intuitive sense that pure things must not mix with the impure, and cultivate (and/or have cultivated in them) emotional responses to perceived transgressions of the implicit and, especially as they get older, explicit norms. This immersion in 'custom complexes' results in a physical embodiment of the 'moral order' such that they experience the world in terms of (moral) purity and pollution. When they are older and the explicit concepts of asceticism, transcendence and sacredness are taught to them, such concepts will not seem strange, but rather clarifications and/or extensions of what will by then have come to seem like perceptions of the world they live in.

It's worthwhile to sketch how this process occurs in a culture fairly removed from ours, which latter does not emphasize purity norms or violations of the natural order (though these norms are certainly not absent). It also serves to highlight the fact that that the moral intuitions are anything but simple knee-jerk reactions. Rather, they are conceptually rich, with those interconnected concepts being employed often preconsciously in a way that involves the affective/emotional system(s). Nevertheless, those intuitions can be supplemented with an indefinitely deep and complex stock of explicit concepts, arguments and justifications.[153]

In the next section I'm going to make use of the Social Intuitionist Model to argue that these concepts and the justifications that they figure in appear to have the function of maintaining or stabilizing the norms and intuitive judgments both of individuals and the culture as a whole; they in effect help to commit individuals and cultures to their own moral judgments. However, as we will see, justifications exist for conflicting norms and judgments, and the materials used in such justifications can be employed not only to support antecedently-held intuitions, but to generate new ones, which can, in turn, conflict with others.

2.5.3 Assimilation of the Social Intuitionist Model to My Commitment Model

Though I did not appreciate this fact this when I decided to discuss Haidt's work, a person looking for a model of moral judgment centered on the importance of commitment could not hope for anything better than the four links that provide the core of of the S-I model. Let's review them. The first link is the intuitive judgment link. This is the fast, felt reaction to a real or imagined circumstance. The emotions backing or

[153] That the affective system can employ conceptually rich and complex concepts shows that even if projectivism is true, it does not imply that affective intuitions are not responsive to perceptions of moral facts, e.g., violations of legitimate or true moral principles. For that we would need an independent philosophical argument, which we will get in Chapter 3.

(partially) constituting these judgments are just what Frank, Joyce, and I have in mind when talking about emotional commitments. The intuitions Haidt discusses are rich in conceptual content, but as we'll see this content tends to strengthen the associated commitments.

The second link is the post-hoc reasoning link. This process searches in a biased way for arguments and justifications to support the intuitive judgment already made. An unbiased or less biased process could only tend (on average) toward the undermining, or destabilization, of the initial judgment. Further, this post-hoc reasoning generally only takes place in the context of a challenge or request for justification from someone else. But whether the doubt arises from within or without, the post-hoc reasoning link can only tend to (further) commit one to the original judgment. It is in the business of answering *why* one's judgment is correct, not *whether* it is. Asking the second question has a destabilizing effect. The first one is purely a search for *supporting* reasons. These reasons are like the ropes tying down the statues of Daedalus in Plato's *Meno.* Bypassing the question of whether these reasons contribute to knowledge, we can recognize the descriptive fact that judgments supported by conscious reasons are less apt to 'fly away' than those without, other things being equal. I'll have more to say about the stabilizing effect of reasons on judgments below.

The 'reasoned judgment' link

The third link is the 'reasoned persuasion' link. Post hoc reasoning can influence others' intuitions and judgments. Haidt thinks that the content of the reasons is rarely what affects others, but the affect that typically attends the justifications is likely influential. But Haidt neglects the crucial fact that the content of the judgments can provide and/or reinforce commitments of the speaker, as Joyce noted. They can also make and/or reinforce commitments of the hearer. Haidt says that the persuasion doesn't work by providing arguments, but misses that even if the affect is what in fact 'persuades', *the reasons the person provides are taken as justificatory*. Since people aren't *aware* of the post-hoc reasoning phenomenon, they *take* these reasons to be both causal and justificatory, whatever their actual role. Having taken them as justificatory, they are now available to feed back into the person's own intuitions and judgments and reasonings about their own and others' behavior.

To see the crucial difference that the 'reasoning' component makes, imagine two different very simple scenarios. In one case, Billy steals from Jill. Jill is angered and intuitively judges that Billy is immoral for stealing. She communicates this judgment, with its attendant anger, to her friend Bobby, who is influenced by Jill's affect to also form an intuitive judgment against Billy. But Jill has provided no rationale for her (moral) anger, only the affect. She has made no appeal to publicly available norms that could be presented as both cause and justification of her anger.

In the other scenario, she does provide such a justification. 'Stealing is wrong' she says. She is angry, and condemns Billy, so she says and believes, because what he has done is wrong. The others are influenced by her affect, but they are also members of a community which accepts that stealing is wrong. They all (roughly) believe and publicly express that their moralized anger and judgments against Billy are both caused and justified by the fact that Billy has stolen and stealing is wrong.

In the first case, even if we imagine that the relevant social norms are in place, without any implied appeal to them as the appropriate cause of the relevant emotional reactions and associated judgments, those norms are not reinforced by the judgments. The people are plausibly no more committed to the norm(s) against stealing than they were before. If they do not have any explicit norms, they may yet be at the level of alliances we see in primates, where Joyce rightly notes that there seems to be no notion of a *transgression*, or *deserved* (as opposed to expected) punishment. The justifications people give for their moral judgments, post-hoc or not, are in the business of saying why it is that the Billys of the world *deserve* to be punished, not why they *will* be.

This is not a story that gets purchase on how such public norms and their usage arise, but it is meant to show (at least one important means of) how they are stabilized and maintained. When one consciously and publicly declares one's judgment that another person *ought* to be punished *because* he stole, one commits oneself (further) not to steal. Joyce focuses on such declarations operating at the interpersonal level, since if one has publicly expressed that 'it would be immoral to do X', then that acts like a contract by which one accepts that one should be punished for doing X (2006, p. 122). This is to view the commitment as more 'secured' than 'subjective'.[154]

These secured interpersonal commitments (if that is the best way to understand them) might be important, but I think they are not as important as the subjective component of the commitments, at both the personal and interpersonal levels. The most obvious reason for this is that it seems that the vast majority of transgressions, across all

[154] In this context Joyce is explicitly providing ways in which moral judgments can create interpersonal commitment without emotional backing. The establishment of such a 'contract' would be a paradigmatically secured commitment.

the cultures I am aware of, do not require anything like a stated willingness to be punished should one violate a norm. People are punished for violating moral norms whether they ‘agreed’ to be or not. Nevertheless, a person who has consciously and publicly justified their own moral outrage (and calls for punishment) in terms of the violation of a norm is plausibly more subjectively committed to avoid violations of that norm than they were before having done so. If Joyce is right that the emotion of guilt requires the thought that one has transgressed against a norm (typically, a norm of autonomy), then the conscious attribution and justification of one’s own moral anger to the perceived violation of a particular norm only makes that norm more salient, and whatever guilt one is disposed to feel at the thought of violating it both more likely and acute.

Still, Joyce is right to think that such declarations can constitute interpersonal commitments as well, even without emotional backing (though they will be weaker without it). I said above that cultures don’t require their members to make implicit verbal contracts in order to punish them. But this is not to say that making such ‘contracts’ do not have important committing functions.

There is often disagreement about what sorts of actions are appriopriate in what circumstances. No culture has a decision procedure that produces clear and agreed-upon judgments in all actual or potential cases of conflict between individuals or subgroups of the culture. According to the S-I model (and in my view any serious model), people will often make moral judgments motivated by their particular (perceived) interests and/or concerns. Parties to a conflict then will often be called upon to publicly justify their judgments. In claiming ‘revenge is my right’ or ‘revenge is barbarous’, they thereby reduce their options in the future with respect to revenge.

Moral hypocrisy is perhaps the most common and most commonly denounced of sins. Moral hypocrisy just is the trait/activity of making commitments (in the form of moral judgments) and then acting or arguing contrarily to them when one’s perceived interests or concerns are differently aligned. As Kurzban and Aktipis (2007) have argued, the functional modularity of our minds allow for extreme—and extremely common—inconsistencies between personal beliefs, public declarations and action. The fact that hypocrisy is very common does not prevent there being significant costs associated with one’s hypocrisy being discovered. One cannot publicly argue for the immorality of revenge and the next day take public revenge or argue that one should be allowed to, at least not without surmounting the significant obstacles that one has created by arguing in the opposite way.

In fact, the only way that such declarations, in the absence of emotional backing, could be a signal *at all* is because of the potential cost associated with discovered hypocrisy. Otherwise, such declarations could be made whenever they seemed to suit one’s interests, in which case they would contain no useful (relevant) information and would therefore tend to be ignored. And it’s important to note that even if we understand such commitments as secured, in that they involve a contract of sorts, the existence of such a ‘contract’ is founded on the fact that people are subjectively committed to react negatively to hypocrites.

We’ve seen how post-hoc reasoning, when used in public justifications of intuitions, can create and/or strengthen commitments to the norms employed in those justifications. This particular story allows, but does not require, that the person was originally outraged *because of the violation of the norm*. The norm existed of course, and the person was aware of the norm, and was presumably already disposed to some extent to avoid violating it, perhaps only out of fear of punishment, which seems to be the case in young children. My (addition to the) account has it that the emotions of the speaker's audience are aroused (mostly) by the former's affective state (emotional contagion is common, especially among those with pre-existing emotional ties), and then the intuitive emotional judgment is *attributed* to the perceived violation of a norm.

This process works for both speakers and audience. Such a process plausibly stabilizes norms not only by increasing individuals' subjective commitments to avoid their violation, and the punishments those violations issue in. It also plausibly contributes to the uniquely human phenomenon of 3rd-party norm-enforcement (moralistic punishment). But I won't argue for this claim here. What is important to note is that this process is both committing and interactive. The more the negative responses are attributed to perceptions of norm-violations, the more that perceived violations of those norms in the future will tend to cause, or be used as leverage to cause, the associated negative responses.[155]

Link 4 is the social persuasion link. This states that people are often directly influenced by the moral judgments of their friends, allies and other group members even in the absence of 'reasoned persuasion'. The committing properties of such a mechanism are obvious. They are a simple extension of the emotional commitment involved in the

[155] We'll see below how Singer employs this type of leverage by attributing our negative reactions to someone who lets a child drown in front of them as resulting from a perception of their having violated a moral obligation.

original person's judgment to the others in her group. Such a process does not provide a mechanism for stabilizing norms or subjective commitment to them, but it does provide a mechanism for committing groups to particular actions. The extent to which this mechanism operates (that is, in the absence of any justifying reasons) is far from clear, but it does not matter for my purposes. It is just link 3 with a very important committing component removed, leaving a mechanism that simply spreads one or more individuals' commitments to others.

The Reflective and Reasoned Judgment Links

I bypass link 5 for now and move to link 6. This is the reflective judgment link, and the first that involves what we would recognize as anything like genuine moral reasoning. It is the private moral reasoning that occurs when moral intuitions clash. A person may, by hypothesis, go through cycles of links 6, 1 and 2 indefinitely in an attempt to settle on a stable judgment (829). Haidt emphasizes the role that perspective-taking and/or attentional focus has in generating clashing intuitions. For example, abortion may seem permissible when our attention is on the mother's wishes and ostensible rights but not when considering the gory death of the fetus. Likewise, empathizing with the points of view of different people may give rise to conflicting intuitions.

When intuitions clash, reasoning can cause a judgment in an individual[156] in that it can be more successful in building a case for one intuition than another. I want to focus here on the kinds of clashes that come from the two different kinds of ethics that seem to be emphasized in western culture, those of autonomy and community. Intuitions can easily clash within individuals when presented with a scenario which requires violating individual rights for community benefit. So long as the intuitions clash, the person's preferences or judgments are unstable with respect to this situation. To achieve stability, which is of course highly desirable in real-world situations, reasoning can make use of whatever justifications the culture provides and/or it can invent.

It's important to note that the more one's reasoning begins to make one intuition 'feel right', the less people tend to actively seek out reasons in favor of the other intuition (829).[157] But which side wins will presumably often be largely due to idiosyncratic features of the scenario, as well as the individual (at the time). For example, the intuition that a doctor shouldn't sacrifice an innocent person to save the lives of five sick ones who need different organ transplants will typically be justified in terms of individual rights. The intuition that one should allow a village to be bombed so as not to reveal the breaking of a code in wartime will be justified in terms of 'community' or other collective consequences, but not rights. Likewise, if one is personally invested in the outcome, one will be motivated to find justifications of the sort that lead to one's preferred conclusion. So the direction in which stabilization is achieved need not be even

[156] I argued that it was importantly causal in creating and/or strengthening commitments in links 2 and 3, and therefore potentially causal in generating future judgments. But it does not give rise to the judgment in the individual at the time those links are invoked.

[157] The more they do this, the less stable their judgment will be. This increased destabilization will only require more reasoning to achieve stability. This process will be important when we get around to discussing philosophers.

close to determined by the person’s dispositions with respect to providing one kind of justification or another in general, or by the features of the situation. But the more one kind of justification ‘wins’, the more prone one will be to call on that kind of justification in the future.

Though the model holds that reason is causal in these cases, the nature of its causal power is still not appreciated by the person. They see reason as having beaten a deadlock not because on this occasion it happened to have been able to come up with a more effective case for one side than another, and certainly not that it was employed in a motivated way. For these reasons to seem genuine justifications, the reasons need to seem stronger, and not in a merely descriptively motivational sense. As a result, there should be some momentum to rely on reasons of the same kind in the future, which momentum should be greater as the explicit awareness (which is often greatly increased by a public declaration) of the justifying reasons increases.

If explicit awareness is tied to great intensity of feeling, the commitment to make future appeals to justifying reasons of the kind in question will be yet greater. Let me explain. If there is an emotionally intense conflict of intuitions, we should understand that to mean that the intuitions represent strong and conflicting commitments. That means that in the absence of either of the competing intuitions, each intuition would provide a powerful and often sufficient emotional incentive to act in accordance with it, the nature of these emotional commitments being such that acting contrary to them is likely to result in painful emotional consequences. In order to act on one commitment in preference to another, either the force of the ‘losing’ commitment must be lessened or the person should expect to suffer some significant emotional consequences for violating that commitment.

Sometimes this is unavoidable, as in the film *Sophie's Choice,* in which a mother has to choose which of her children to sacrifice to the Nazis on pain of losing them both. But I gather that very much of the time the justifications we provide ourselves with (and others provide us) help to lessen the force of the pain we would otherwise feel. In fact, even Sophie must have felt less intense and enduring pain than she would have if she did not have access to any justifying reasons for giving her daughter to the Nazis.[158] After all, there is quite a lot of difference between believing that we have acted rightly in regrettable circumstances and that we have acted wrongly. When Churchill had to let British towns be bombed in order not to reveal that the British had broken German codes, it was unlikely an easy or remorseless decision, but believing the justifications he and others provided for his actions surely make the process much more bearable than if one believed one had acted wrongly in allowing one's countrymen to be bombed to death.

The point here is that in such situations there will be *strong motivation to accept the legitimacy of the justifications* that one (feels that one) used in arriving at one's judgment and action. If one has once allowed rights to be violated in the pursuance of the common good, then the kind of justifications that led to this decision will be difficult to disavow in the future on pain of feeling guilt for past actions. And the reverse would also

[158] So far the flavor of this account might suggest that justifications are nothing more than commitments to behave in certain ways and ways of making ourselves feel better about what we've done. This raises the question whether behaviors are ever 'really' justified. In the context of such a nightmarish example, I feel I must say both that there are excellent, normative justifications for behavior and that Sophie's choice to give up one of her children (I don't say which one) was justified. Which is not to say that she reasoned herself to it, but rather that she perceived, largely via her emotions, that giving one of them up was the right thing to do. Of course there will be more on these issues later.

seem to be true (if one had upheld rights at significant cost to overall welfare). Of course sometimes people regret their decisions afterward, and in those cases may be rather *disinclined* to trust the styles of justifications that led to them. Add to this the general pressure for consistency (that is, the commitment to consistency) in one's judgments and justifications, which varies tremendously from person to person depending on temperament and circumstance,[159] and there are multiple pressures to commit to one style of justification over another having done so in the past, which will then affect the kinds of judgments one makes in the future.

We should not think of this effect as restricted to occasions in which independently-arrived-at intuitions clash, but rather one's mode of justification can feed back into one's intuitions.[160] But when they do clash again, the preferred justificatory style will have an advantage, pushing one's judgments and intuitions ever more toward the 'ethic' to which one is increasingly committed. Therefore we can see this kind of moral reasoning as providing stability both in the context of individual judgments where intuitions initially conflict, as well as stability in one's style of justification and therefore the kinds of judgments one is likely to make.

[159] That is, this hobgoblin may take the form of a kind of personal conscience and/or be imposed on one by the necessity of making public justifications.

[160] Again, this is why Haidt's evidence should not be thought to lead directly to projectivism, understood as entailing antirealism. Accepted principles and justifications can feed into intuitions, and can presumably do so even when they cannot be successfully called upon in defense of the judments. But more importantly, there are large differences between different people. Just because some people can be morally dumbfounded doesn't mean we all can. Some people may have integrated principles or justificatory styles into their intuitions far more than others, and be far more capable of calling on those principles or justifications explicitly in defense of their judgments.

One would predict from this that the more explicit justifications one has to make, the more one will favor one style over another.[161] This seems to be borne out at least insofar as philosophers both have to defend their judgments vastly more than average and tend to adhere to normative theories such as consequentialism or deontology far more often as well. These 'foolish consistencies' are predicted by the commitment model.

I said that link 6 was 'the first that involves what we would recognize as anything like 'genuine moral reasoning'. That said, it is certainly not paradigmatic reasoning in that it does not begin from propositions and inference-making mechanisms (or rules) and work its way to conclusions. We started with the (conflicting) conclusions and tried to see which one could find better support from reasoning. Link 5, the reasoned judgment link, allows that 'a person could, in principle, reason her way to a judgment that contradicts her initial judgment' (829).

Haidt says nothing about how this process would work, but only that the evidence (Kuhn, 1991) indicates that philosophers may be special in their ability to do this, trained as we are to 'follow reasoning even to very disturbing conclusions' (829). It is in these cases that he thinks that reason 'cannot be said to be the 'slave of the passions''. Unsurprisingly, I think that commitment is at work here too, and so granting this kind of reasoning as an exception to Hume's dictum is too hasty.[162]

I mentioned above that Singer would use the attribution leverage I talked about. Haidt cites Peter Singer as one of the philosophers capable of following reasoning to

[161] That is, perhaps until one makes enough of them that one begins to suspect that both styles represent different commitments, at which point one might welcome back the previously marginalized considerations.

[162] Hume allowed for 'calm passions'. All Hume's 'passions' are perhaps best read as 'desires', where these needn't be consciously available. So the reason I say Hume's dictum should not be thought to have been violated is that I consider commitments desires. But I won't elaborate or defend this until the next chapter.

disturbing conclusions. Singer is as good an example as any in showing the importance of emotion-backed commitment in moral judgment and the will. In the next section I'm going to provide a case study of Singer's most famous argument as an illustration of the importance thereof. It will also exemplify the themes that we have been interested in, namely 1) our experience of and belief in intrinsic value, 2) the deflection of attention away from one's own motivations, and 3) how the attribution of emotional responses to conceptually rich norm violations can be leveraged to generate novel responses. It is my aim that with the tools we've been developing, we will have a way to understand what is happening in this argument (and it is meant to be representative of many persuasive moral arguments) that reinforces the commitment model of moral judgment, including and especially its explanations of these three thematic elements in terms of their motivational functions.

2.6 Singer's Commitment—A Case Study

Singer's famous *Famine, Affluence and Morality* (1972), is a representative example of his style of argument, a style very common in philosophy by which we sometimes arrive at counterintuitive results. The argument goes like this:

1) Death from starvation and/or lack of medical care is bad in itself.
2) If we can prevent something bad from happening without giving up something of comparable moral importance, then we have a moral obligation to do so.
3) There are people dying from starvation and lack of medical care (just for example), and we can prevent some of that without sacrificing anything comparable.
4) Therefore, we have a moral obligation to prevent these people from

> dying and to help the poor and desperate generally.

Given the situation of the world,[163] the argument's conclusion entails that most of us should be working full-time to prevent such circumstances. Roughly speaking, we should give whatever we have until we are at or just above a subsistence level, or a level at which a sacrifice on our part would be equal or greater than the benefit we could bestow on another. This characterization, though slightly rough, is more than accurate enough for my purposes.

This has been a very influential argument, eliciting many kinds of response, including agreement, of course. The argument has only two 'moral' premises, which when combined with the state of the world, generate the 'disturbing' conclusion. The second he thinks requires very little argument, for it 'seems almost as uncontroversial as the [first] one.' The first one is that starvation is bad. Singer allows that 'people can hold all sorts of eccentric positions, and perhaps from some of them it would not follow that death by starvation is in itself bad. It is difficult, perhaps impossible, to refute such positions ... [t]hose who disagree need read no further' (231).

He does nevertheless go on to defend the second principle, or rather a significantly weakened version of it, which states that 'if it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it' (231). He defends it in the following way. Suppose you're walking by a pond and see a child drowning. You could save the child at the cost of ruining your shoes. In such circumstances you morally ought to save the

[163] The immediate context was a famine in East Bengal in 1971, but the relevant circumstances are no different today.

child.

This certainly does seem uncontroversial. But now notice that accepting just these two premises, combined with the state of the world, entails that most people in western society (but not only western society) should be living quite different lives than they are. Since any affluent person can prevent very bad things from happening without giving up anything 'morally significant' (expensive cars and luxuries of all kinds will presumably not count as morally significant), such people should be doing so up to the point at which this is no longer true.[164]

Notice how such a disturbing conclusion is reached. First, we start with intrinsic value. This is taken for granted to the extent that anyone not recognizing that people starving is, 'in itself' a bad thing is simply dismissed with an *ad hominem*. It is not worth the trouble to convince anyone not able or willing to see this simple fact. Second, we operate on principle. Actions are classed (harkening back to Ainslie, we might say bundled) together into kinds by means of collecting them together under a principle. Further, Singer calls the example of saving the drowning child an 'application' of the principle. This suggests that such actions as saving drowning children are motivated and performed by applying statable (I do not say stated) principles of action such as the one described. Finally, notice the role of emotion-backed intuitive judgment.

A person who declines to save a drowning child in order to save her shoes would receive quite a harsh moral judgment. By assimilating not feeding children across the globe to allowing one to die in front of you, he encourages the emotions toward the latter

[164] This is the much weaker version of the argument. The stronger one uses the original premise involving moral comparability, not the weakened one of moral significance, the former having much more counterintuitive implications.

to be employed toward the former. And this is what is *disturbing*. It's not the sheer conclusion that you should give (much) more money to the poor, but that in not giving you are, morally speaking, no different than someone who lets the child drown.[165]

This recalls the point that arguing in metaphor is more common and presumably at least as effective as arguing from premises to conclusion. But here we have the strategies effectively combined. We have the *emotional* force that comes with using the metaphors, as in identifying meat-eating and abortion with murder.[166] But we also have the *stabilizing* force of an explicit appeal to publicly and personally accepted moral principles. These and the other considerations Singer provides have the effect of reducing the availability of post-hoc justifications of the intuition that one has no obligation to give. Which is of course why he is careful to provide them.[167]

[165] I realize that this might seem to conflate judgments of moral correctness with judgments of moral character. But I think that part of the force of the argument derives from the juxtaposition of the motivation we would have to save the child in front of us—or to punish someone who did not—and the relative lack of motivation to save distant, 'invisible' children and/or punish those who don't. So I think the argument is meant to imply not only that saving the distant children is the right thing to do, but to encourage a judgment of negative moral character toward those, including and especially ourselves, who perceive the situation and are yet not similarly motivated.

[166] Of course the people using these arguments don't conceive of them as metaphors. 'It's not a metaphor—meat really *is* murder! And so is abortion!' (Though the same person rarely makes both claims). But whether they think they are metaphors or not, these are far from the paradigmatic instances of murder. One seeks to recruit the highly negative emotional reaction that attends 'exemplar' cases of murder and have it called forth in response to these cases as well. This analysis can hold equally well even if we were to (confusedly) agree that meat-eating and abortion really are murder.

[167] Anyone who has spent a discussion section on this argument with undergraduates has seen post-hoc justifications at their most obvious and painful. Students will give excuses for not giving that have already been clearly undermined by Singer in the argument, and repeated in class. People will make wildly implausible claims about what kinds of things are 'morally significant', what they need to be motivated to work hard enough to make enough money to give away in the first place, etc. To be sure, there are non-implausible ways to lessen the dramatic conclusion Singer reaches, but few students conclude that significant giving to the poor is a moral requirement. Just as in Haidt's cases, when you point out the weaknesses of their arguments, they eventually fall silent, but do not give up their judgments. But some do, at least for a while, seem somewhat disturbed. They are plausibly (much) more disturbed and for longer than if Singer had either simply asserted the moral equivalence of not saving a drowning child and allowing a foreign child to die of hunger without argument, or argued in purely 'abstract' terms for the equivalence, without any real-world example to arouse emotions of blame, guilt, anger, contempt, or whatever people feel at someone allowing a child to die in front of them to save their clothes. The combination of the two is what makes the argument powerful.

I introduced the notion of practical clout with a thin and invented example employing two versions of an anti-abortion activist, one who expressed or reported her emotions and one who made characteristic moral judgments. I said that it seemed that more motivation seemed to be recruited by deflecting attention away from the emotions. I conclude the main argument of this chapter with a representative philosophical argument that attempts to get practical clout by employing the notion of intrinsic value, directing attention away from the emotions while using them to generate the judgment, and employing the use of principles to group (hypothetical) actions together.

Finally, notice that Singer *attributes* our extremely negative responses to a person letting a child drown (in order to save their shoes) to their having *violated a moral principle*. I think it is *extremely implausible* that the perceived violation of Singer's stated principle is the cause of most people's negative responses to such an imagined case.[168] However, it does seem like a socially accepted (or at least acceptable) norm that we should prevent bad things if we can without sacrificing anything morally significant.

Now, recalling the discussion of the committing function of the 'reasoned persuasion' link above, if a (strong) negative response is attributed to the perception of the violation of such a norm or principle, then that acts to reinforce the principle. To deny the principle appears to excuse allowing children to drown to save one's shoes. And since people do not reliably have access to their motivations or the causes thereof, and we are accustomed to believe that moral action proceeds from principles, Singer can have us believe that his principle is both cause and justification for saving a drowning

[168] Again, it could be that for some people this is the case, in the sense and to the extent that they have internalized such a principle. But the evidence is good that most people are not doing so, and I personally doubt whether this is anyone's primary reason for their response in such a case.

child in the typical cases. Then this rationalization is fed back into a judgment-producing process, in this case, one that has counter-intuitive results.

At this point we find ourselves in the situation described in the reflective judgment link. We have two conflicting intuitions (though the new one is likely to be much weaker) and so reasoning is employed to find support for our prior judgment. If Haidt is right, people will stop looking for reasons when one side 'feels right'. This is a difficult position to achieve with respect to Singer's judgment, because it will almost necessarily be uncomfortable to think that one morally should be doing much more for the sick and dying than one is. But people can be 'forced' to (grudgingly) accept this position because of a commitment to justifying their moral judgments and an inability to find a way to do so in the face of Singer's arguments.

There is a lot more to say about this, but for now I want to show that a number of worthwhile points are illustrated in Singer's argument and the common responses to it. The first is that people overwhelmingly accept the premise that some things are intrinsically bad (good). Peter Singer has been using this argument for over 35 years, including in his latest (2009) book; I have taught it to many tens of undergraduates, related it to friends, and debated it with philosophers and I can't recall anyone simply rejecting the premise invoking intrinsic value.[169] The second is that its invocation seems to deflect or direct attention away from one's own emotions in the case. I won't argue more for this beyond commenting that we have negative emotional reactions to the idea of children starving, which emotions plausibly form the basis for the judgment that it is a

[169] Update: since this was originally written I taught Singer in a class where the students had read arguments for and against 'objective values' earlier in the course. In this course some students objected to the intrinsic value claim.

bad thing. But these emotions are conspicuously absent in the description of the situation as a bad thing in itself.

The third thing is that Ainslie has given us a way of understanding Singer's reasoning in terms of principles as paradigmatically commitment-based. This is not to say that this particular principle has arisen by representing individual decisions as 'test-cases' as might happen in other contexts. But the 'technology' of acting according to principle, whether learned culturally, individually in an Ainsliean fashion, and/or in some other way, is widely believed to have an 'anti-akratic' motivational function achieved by subjectively removing differences between the relevant cases, treating them as effectively all of a kind.

I am arguing that moral concepts play an important role in achieving this identification-in-kind. After all, there seems to be quite a lot of difference, psychologically and otherwise, between allowing a child to drown right in front of you in order to save your clothes and not sending money to help children on the other side of the world. But Singer can use moral concepts to abstract away from these differences. In addition to the principle about intrinsic value, another principle about which he does 'not … need to say much in defense' (232) is that distance (whether physical, social or cultural) is morally irrelevant. As he notes, these things will make it more or less *likely* that we help, but they cannot make a difference to whether we *morally should* help. Presumably, the differences in (multiple kinds of) distance make a descriptive psychological difference because of differences in our emotional reactions. But if it is true that 'from the moral point of view' (232) we are to ignore all such differences then it seems that just to that extent we are to ignore what are ostensibly the primary inputs to our faculty of moral judgment, our emotions.

This example can help show why such judgments seem to be both the products of reason and unmotivated by emotion or desire. But this is not so. For the desire to avoid the feelings of guilt aroused by the attempted assimilation and the desire (commitment) to be able to justify one's moral intuitions and judgments are just two of the desires which seem to be clearly operative in this case, which I take to be, in these respects, a quite typical case.

I want to be clear that nothing I've said is meant to imply that many people should not be giving (much) more money to the poor than they are. In fact I find that very plausible. But I do think that conclusions to this effect reached in the paradigmatically moral mode of argumentation distract us from why this is true, to the extent that it is. Our attention and focus is led away from the sources of value that are nevertheless hard at work behind the scenes.

2.7 The Levels of (De)Stabilization

So far I have presented moral reasoning as having a stabilizing or committing function, both at the individual and social levels. But this could be very misleading. Let me now try to clarify the senses in which moral reasoning can be stabilizing, as well as destabilizing. In the case of Singer's argument, I argued that the moral reasoning played a role in committing us to a judgment which intuitively for most people is very weak.

There are a few things to be said about this, most of which will apply much more generally.

First, when I say that the reasoning commits us to the judgment, I mean to be employing the conception of commitments as option-reducers, as placing obstacles in the way of actions or judgments that conflict with the commitment. So although most people do not come away ‘committed’ in the usual sense to help the poor more than they were beforehand, in order to go on as before (including feeling the way they did about it), they will have had to overcome obstacles that Singer put in their way (assuming that they engaged his argument—the obstacles don’t arise if one simply ignores or doesn’t care about the considerations). And people do try to overcome them by attempting to show why his argument isn’t sound. To be sure, the vast majority soon go on roughly as before, but this is because the obstacles are subjective in nature. For them to be effective (even to exist), one must care about whether one can justify one’s judgments and actions to one’s own and/or others’ satisfaction, for example. People differ quite a bit on this score, and the extent to which Singer places obstacles to their continuing on as before will vary mostly as a function of this difference. The vividness of their imaginations, the extent to which they are susceptible to arguments based on principle, the intensity of the emotions that are called up will all affect the effectiveness of the obstacles.

Second, to say that such reasoning has a committing or stabilizing effect is only to say that it has it with respect to the particular intuition or judgment at issue. The vast majority of the time (so it seems), reasoning is employed on the side of the initial intuition and/or judgment. In these cases, the stabilization is at the the intuition- (or sub-personal) level *and* at the person-level. In the case of Singer’s argument, reasoning has a stabilizing effect at the level of the intuition and/or judgment generated by the argument, but a destabilizing effect at the personal level.

The same dynamic applies at higher levels. Two members of a group with conflicting judgments[170] (perhaps motivated by simple conflict of interest) will tend to employ reasoning that supports and stabilizes their own judgments. This will tend toward instability in the judgment of the group, especially as members are swayed to opposing sides. And the same dynamic scales up as distinct (sub)groups of increasing size justify their (between-group) conflicting judgments internally, stabilizing them and potentially destabilizing whatever roughly shared judgments might have existed prior to the origin of the conflict. And at all levels there is the potential for analogs to link 5 to be enacted, whereby whatever shared justificatory norms exist can end up supporting one side more than another, or creating a hybrid.

But this is not the only means by which moral reasoning can destabilize. Since by hypothesis we do not typically arrive at our initial judgments by reasoning--much less principled reasoning--demands for (principled) justifications for our actions will often be difficult to meet. We justify actions or judgments by appeal to norms, but it can always be asked why that norm is itself justified. Admittedly, this activity is not carried on often or far by non-philosophers. But it is a feature of justification (as well as explanation) that it can seem as if requests for justification (or explanation) can always be made in response to the last offering.

To continue to use our example, not only can we ask what justifies spending

[170] I think it makes sense to say that when there are significantly conflicting intuitions at the personal level, then there is typically not yet a (stable) moral judgment. At any rate, as I go to higher levels, I will use 'judgment' instead of the more cumbersome 'intuition and/or judgment'.

money on luxury goods when there are starving children, we could ask what justifies the moral principles Singer employs. Singer says that ‘If we accept any principle of impartiality, universalizability, equality, or whatever, we cannot discriminate against someone merely because he is far away from us’ (232). But why should we accept any of these principles? Suppose we come to see that accepting these principles conflicts with our (strong, stable) first-order intuitions. Why not reject the principles rather than the intuitions? This is an example of conflict not between competing intuitions but between intuitions and principles of justification.

We should see two things. First, that rejecting the principles and rejecting the intuitions are of a piece logically. Second, once we’ve seen that even principles of equality and impartiality can be called into question in cases of conflict, we can call those or any other moral principles into question, absent any conflict. And to the extent that we both call those principles into question and rely on them to guide our actions and justifications, we risk losing our normative footing.[171]

2.8 Conclusion

We have drawn on Ainslie, Joyce, Frank, Haidt and associated scientific evidence to provide a commitment model of moral judgment. This model explains our beliefs in

[171] Something along these lines is what Bernard Williams (1985) meant when he said that ‘in ethics, *reflection can destroy knowledge* (148, emphasis his). Williams had in mind ‘hypertraditional societies’ who could lose knowledge not of objective ethical truths but of how to conduct themselves in their particular way of life. Reflection on the (‘thick’) concepts underlying some of the beliefs by means of which they navigated their social worlds could come to disturb and unseat those concepts, rendering them unavailable to do the work they once had. But I think that the same general points apply not only to ‘hypertraditional societies’ but to our own, including that of Western academics and intellectuals.

intrinsic value, the ‘practical clout’ of moral judgments, moral principles and moral judgments generally in terms of the motivational power that comes with the capacity for commitment. I presented a hypothesis that the teleonomic function of moral judgments lies in their contribution to personal and interpersonal commitments. The more plausible this or some similar hypothesis is, the more we should expect that they have a similar instrumental function, along with whatever other instrumental functions they might have.

My assimilation of Haidt’s Social Intuitionist Model of moral judgment to my commitment model gave us, I hope, some good independent reasons to think that moral judgments have the instrumental function of committing us to ways of thinking and feeling. Finally, we can understand Ainslie as having provided both teleonomic and instrumental functions, where the teleonomic function is not given by evolutionary forces, but selective forces based on reward.[172] We can also see, even more clearly perhaps, that the committing function of moral judgments in social contexts has significant potential for pathology at the level of interactions of groups.

I hope that on the whole and in the main, the commitment account of moral judgment is plausible. But I have no illusions that I have provided an airtight case. The crucial point is not that I have definitely gotten the correct (and certainly not the complete) explanation of these things, but rather to show that whether at the biological and/or cultural and/or psychological levels, plausible explanations exist for our beliefs in intrinsic value and the practical clout of moral judgments that do not rely on postulating those very things, but rather advert to the motivational roles played by belief in them.

[172] The instrumental functions are nearly the same as the teleonomic ones, though in cases that appear pathological (such as compulsions) it’s not clear whether we would want to identify the commitments as instrumental functions (since those are understood in terms of what something is ‘good for’).

My goal has been to provide motivational explanations for these things that are as strong as possible. One needn't think these explanations have been *established*, but only that taken together they represent a more plausible hypothesis than that peculiarly moral imperatives or intrinsic value exist independent of any of our motivations. That the explanations involve commitments gives us purchase on how to proceed insofar as commitments imply goals, and goals can at least often be evaluated against one another. But just understanding many of our moral intuitions as (emotion-backed) commitments of various kinds is already illuminating.

There are no direct arguments here against intrinsic value or any of the concepts for which I attempt explanations (away), but if much of the evidence I have been reviewing is correct, the motivations for those beliefs are independent of the arguments put forward in favor of them. Or rather, those arguments support beliefs that were neither arrived at by means of those arguments nor are supported by them in anything other than a descriptive psychological sense. In the context of these explanations, I hope the post-hoc flavor of those arguments will be more readily perceptible.

However, it is not enough to hope that people see the post-hoc, or motivated nature of certain philosophical arguments. In Chapter 3 I'll undertake the philosophical work of arguing directly that all values and reasons—and therefore justifications—are end-relational. I've argued that moral judgments reflect commitments. I will now argue that commitments are best understood as kinds of desires, understood as goals or ends. Then I'll argue that in response to the sorts of pressures for justification described above, the only reasons we can give, and therefore the only ways of establishing and maintaining our normative footing, are ultimately relative to our actual goals.

Chapter 3: Reason for Desires

3.0 Introduction

In this chapter I'll argue directly for a neo-Humean conception of motivating and justifying reasons. Specifically, I'll argue that all reasons (and values) are relative to, or 'from the perspective of' some desire(s). The last chapters argued that even if there were no such thing as intrinsic value or categorical rational authority, we would be motivated to believe that there were. Though I hope those arguments helped to implant some suspision in the reader regarding the legitimacy of these concepts, they cannot substitute for a substantive engagement with the relevant philosophical literature. Nevertheless, much of what I have said in the first two chapters will be relevant to making out my case for an end-relational theory of reasons and values.

It is a crucial aspect of my project to convince my readers that all reasons and values are relative to what I will call 'motivated perspectives'. That is to say that whenever we accept a reason for action or a value-claim, that acceptance proceeds from some aspect of our motivational psychology (not that that's *all* it proceeds from). Therefore when someone makes a claim to the effect that something is valuable, there is something about their actual motivations playing an ineliminable role in that judgment, and likewise when someone accepts a reason for action.

This is a crucial aspect of my project for two reasons. The first is that I think that such a conception of reasons and value fundamentally threatens moral discourse. Normally this threat is couched in terms of providing a crucial premise in an error theory about morality. The error-theoretic strategy proceeds by attempting to show that the concepts employed in moral discourse are essentially committed to some features or propositions that are not true, and therefore the entire discourse is systematically flawed. For all I am claiming, a moral error theory so understood is correct.

But that is not my primary concern. I think that the conception of values and reasons that I put forward in this chapter, if generally accepted, threatens moral discourse because of the motivational role that a lack of awareness of our motivations plays in peculiarly moral discourse. And I think that peculiarly moral discourse is severely threatened by a general awareness of the relational nature of reasons and values. These specific claims, and others related to them, will be elaborated and defended in later chapters. The point of these comments is to highlight the philosophical significance of this chapter for my entire project.[173] Of course I don't expect to demonstrate beyond the possibility of rational doubt that my view is correct, but I do expect to show that my view is at least as plausible, and I think more so, than any of the leading rationalist (anti-Humean) proposals.

My general strategy is to show that a version of neo-Humeanism that avails itself of some of the empirical and theoretical resources that I've developed can handle all the rationalist objections of which I am aware. In the course of responding to these objections, I'll describe what is to my knowledge an unrecognized and yet very deep problem affecting all of the most prominent rationalist proposals. I'll argue that all these proposals are committed to a fundamental misconception of the relationships between

[173] I think the arguments of this chapter also significant general philosophical significance, since the nature of reasons and values is a large philosophical topic.

practical rationality, evaluative belief, desire and the will. All, in their various ways, are committed to the in-principle rational authority of evaluative beliefs over desires in a way that I will argue is untenable,[174] and indeed bizarre. None of these rationalist authors have, to my knowledge, even attempted a response to this problem, though it is potentially fatal to all their proposals. My view, and the model of the will I employ to defend it, not only help to point out this common rationalist flaw, but also has the resources to explain how such a large problem could have gone unrecognized so long.

My specific strategy is as follows. First, I'll argue for the particular conception of desire with which I'll be working, the 'direction-of-fit' conception. Then I'll make a case for a Humean theory of motivating reasons (which is far less controversial than that of normative reasons). Both of these first tasks will borrow heavily from Smith (1994). At the end of the argument for a Humean theory of motivating reasons, I'll recapitulate an important objection to this view by R.J. Wallace, which shows that the arguments Smith provided are inadequate. That rationalist objection will not be met until the end of the chapter.

In the interim I'll make the case for my neo-Humean theory of normative reasons. I begin by describing the attractions of Humean views, attractions which are great enough that if any version of such a view could overcome the prima facie serious problems of Humean views, then that view should be considered the default champion. The next step then is to describe those problems. For this I draw upon David Brink's (2007) systematic attack on Humean theories of normative reasons. I use Brink's paper because I think it

[174] At least untenable without allowing for a potantially large gulf between rational and normative behavior.

provides some of the most difficult problems for a neo-Humean to deal with, and so in dealing with them makes for the strongest possible neo-Humeanism.

One of those problems is that Humeanism cannot recognize robust forms of fallibility in the practical realm. I think this is the most difficult and important issue for any version of Humeanism to address—so difficult that most Humeans have been led to develop versions which appeal to some form of idealized (as opposed to actual) desires to ground practical reasons. My version only appeals to actual desires for such 'grounding' and so will be thought especially incapable of accommodating robust practical fallibility. Therefore in addressing Brink's challenge on this score, I take a long detour through Korsgaard's famous (1986) challenge to Humeanism on just this point.

There I show why a specifically end-relational—as opposed to instrumental—conception of practical reason is important in meeting some of Korsgaard's challenges. In addition, I identify some commitments that Hume had that made his view vulnerable to the objection from robust fallibility, but they are commitments that we can and should give up. Chief among them is that one's strongest desire provides one's strongest reason. Giving up this claim is not only consistent with Humeanism, but is a purer form of it, since the temptation to make that claim stems from the failure to see that reasons are relational 'all the way down', i.e., there is no nonrelational fact of the matter which of one's reasons is the strongest.

Next I compare my view head-to-head with Korsgaard's and show that mine wins along several important dimensions. I argue that in the places where Korsgaard goes right, especially in her conception of the relationships between practical identities and the will, my Ainslie-inspired view shows why she is right. And in the places she goes wrong, it explains that too.

Having demonstrated that my view can recognize and explain robust forms of practical fallibility, I return to address the remainder of Brink's criticisms. I do my best to show that my view is strengthened by its having been required to deal with Brink's forceful critiques. Then, having built up the view in the process of defending it, I proceed to directly compare it to the rationalist proposals of Michael Smith, Tim Scanlon and R.J. Wallace.

In this comparison, I defend my view from some of their attacks on desiderative conceptions of motivating and justifying reasons, but I am most concerned to highlight the deep and serious problem that all their proposals have in common. I argue that their in-principle rational prioritization of evaluative belief over (first-order) desire amounts to what I call a 'willpower conception of rationality'. As I said, I argue that the commitment that underlies this conception is untenable and bizarre. Further, my view can explain how, despite its bizarreness, it could seem so obviously correct that none of the authors spend almost any time arguing for it. Since Wallace's early critique of Smith's Humean theory of motivating reasons depends on this bizarre commitment, showing that the commitment is untenable undermines that critique.

I end the chapter by directly engaging arguments to the effect that certain desires are intrinsically irrational. It is a central feature of neo-Humeanism that I wish to keep that no desires are intrinsically irrational. Or perhaps more accurately, I argue that the only sense in which desires can be intrinsically irrational is predicted and explained on the hypothesis that all reasons are relative to motivated perspectives.

3.1.1 Commitment, Desire and Direction of Fit

Direction-of-Fit

I'll conceive of desires as dispositions manifesting a particular 'direction of fit' with the world. On this conceptualization of belief-desire psychology, beliefs are (representational) states of an agent such that a mismatch between the content of the representation and the aspect of the world which the belief represents tends to be resolved by changing the belief to fit the world. A desire has the opposite direction of fit. Where a perceived mismatch exists between the content of a desire and the relevant state of the world, the agent does not tend to accommodate the desire to the world, but tends to change the world to match the content of the desire. This is reflected and summarized in the common idea that beliefs aim at the true, while desires aim at realization.[175] I will briefly rehearse the virtues of such a conception, provided by Michael Smith (1994), but the view has many adherents.[176]

This conception of desire is attractive for several reasons. First, it provides a plausible epistemology of desire. We commonly make desire attributions where it is not

[175] Though I will be characterizing desires according to this model, I am reluctant to do so with beliefs. I think it is at best misleading to say that (all) beliefs aim at the true, especially the kind of beliefs I am primarily concerned with in this dissertation, moral beliefs. This could be grounds for not considering them beliefs, but I want to remain uncommitted about whether they are (best thought of as) beliefs, while remaining committed to the thesis that they employ the representational trappings of belief to do their work.

[176] Examples include: Elizabeth Anscombe, *Intention* (Ithaca: Cornell University Press, 1963), sects. 36, 40; I.L. Humberstone, "Direction of Fit" Mind 101 (1992), pp. 59-83; David Velleman, "The Guise of the Good" Nous 26 (1992), pp. 3-26; Bernard Williams, 'Consistency and Realism', reprinted in his *Problems of the Self* (Cambridge, England: Cambridge University Press, 1973), 187-206; John Searle, *Intentionality* (Cambridge, England: Cambridge University Press, 1983), 7-9; and Richard Wollheim, *The Thread of Life* (Cambridge, Mass.: Harvard University Press, 1984), 52-3.

assumed or entailed that there is any phenomenology associated with the desire, and certainly not a phenomenology that is straightforwardly introspectable as evidence of that particular desire. Since at least Nietzsche and Freud, we have become comfortable attributing desires to people that they would sincerely deny. This isn't to say whether most or any of those claims are true, but it is to say that there is no general requirement for an agent to have phenomenological or otherwise conscious access to her own desires. Smith gives the example of 'John', who buys a newspaper from a relatively inconvenient stand that has a mirror. If the mirror were removed and he stopped going there, and especially if we had other evidence of vanity, we might reasonably conclude that the reason he went there was because he wanted to see his reflection, even if he were to sincerely deny that that is the reason (106). Examples of this sort can be multiplied indefinitely. Any adequate conception of desire has to accommodate the possibility of desires that are not (straightforwardly) introspectable.

On the direcion-of-fit conception, desires are dispositional states. An ascription of a particular desire to someone is an ascription to them of dispostions to act in certain ways under certain circumstances, where those circumstances include, among other things, other desires, as well as beliefs about how to obtain them. Call these circumstances the 'manifestation conditions'. So the epistemology of desires is just the epistemology of counterfactuals involving what a person would do in a variety of circumstances which include her own belief- and desire-set (113).

The dispositional analysis of desire has several other attractive elements. First, it allows us to claim that some desires have phenomenological content, while others have none whatever. That is possible just in case some desires are dispositions to produce, among other things, certain feelings, while other desires are not dispositions to produce any feelings, though they do dispose their bearers to certain kinds of behavior (114). Second, the dispositional account allows, but does not force us, to conceive of desires as causes. It does not force us just because whether dispositions are causes is a contested view (114).[177] Third, we can capture the claim that desires often involve beliefs as parts, since the conditions under which some desires are manifested might require the presence of certain beliefs, including but not limited to means-ends beliefs.

Finally, and most importantly for my purposes, a dispositional account fairly straightforwardly implies that one can be quite mistaken about one's own desires. For if desires are dispostions to behave in certain ways under certain counterfactual circumstances, there is no general reason to think that all such behavioral dispositions are known to their possessors. Nor is it the case that believing that they are true implies that they are true or will render them true (114).

Commitment and other Pro-Attitudes

Despite everything I've just said, I want to distance myself from the claim that this conception of desire reflects all or even most of what we normally have in mind when talking about desire. We might not 'want' to study but nevertheless be motivated to do it, and we might not 'want' to do our moral duty as wee see it, but do it nonetheless. Somehow we are motivated to do it, and one of the questions that separates Humeans about motivation from non-Humeans is whether there is a requirement for any desire to

[177] Like Smith, I do conceive of them as causes.

do the motivating, and whether that desire has to be antecedent to the judgment that one ought to do it. I am taking the Humean position that there is an antecedent desire, but all I mean by this is that there is an antecedent motivation of a relevant kind. The term 'pro-attitude' seems best-suited to perform this catch-all function. In fact, Smith addresses the critique that 'desire' is not broad enough to capture all of the states we have with the appropriate direction-of-fit. His response is to recommend that the Humean 'simply define the term 'pro-attitude' to mean 'psychological state with which the world must fit', and then claim that motivating reasons are constituted, *inter alia*, by pro-attitudes (compare Davidson, 1963, p. 4)' (117).

This is exactly what I propose to do. I follow Davidson in thinking of pro-attidues as including:

> desires, wantings, urges, promptings, and a great variety of moral views … [and] values in so far as these can be interpreted as attitudes of an agent directed toward actions of a certain kind … The word 'attitude' does yeoman service here, for it must cover not only permanent character traits that show themselves in a lifetime of behavior, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action, like a sudden desire to touch a woman's elbow.

When I speak of 'desire' in what follows, I will always mean 'pro-attitude', where both will mean ''psychological state with which the world must fit', unless otherwise noted. So these terms will not have exactly their usual meanings or connotations. I frequently use 'desire' because I find 'pro-attitude' clunky, 'desire' is the traditional way of speaking, and 'desire' is usually close enough to what I mean anyway. I also use many of the 'species' names above, as well as 'goals,' 'concerns' and things one 'cares about', where I think these terms better express what I have in mind than the catch-all 'desire'. All terms other than 'desire' retain their normal meanings, as I think they can all be understood as forms of 'pro-attitude'. And remember, a pro-attitude on this conception can be had without one having a 'positive feeling' toward the relevant action, or any feeling at all.

A particularly important species of pro-attitude is commitment. I will be using 'commitment' in the way I've been using it, with all the theoretical trappings. Like desire, having a commitment on this understanding is consistent with not believing that one has it, and believing that one has it is consistent with one not having it.[178] In fact, it is in the context of commitments that it is often the case that we have a desire to do something even if we don't 'feel like it'.

3.1.2 A Humean Theory of Motivation

Before beginning my defense of a (neo)-Humean theory of practical reason, I think it is worthwhile to rehearse the virtues of a Humean theory of motivation, also articulated by Smith (1994). The former project aims to show that all normative, or justifying reasons are relative to desires. The latter aims to show that the reasons which explain actions, the reasons that in fact motivate people, are all relative to desires. I'll summarize these arguments before moving on to the more difficult task of defending my

[178] Technically, I think that if one believes one has a commitment, that in itself tends to create a commitment, due to features of the will that we have discussed and will discuss further. I think one will not often if ever be wrong in thinking that one does have a commitment at all, but rather about the nature or strength of the commitment.

particular neo-Humean theory of practical reason. All mentions of reason-explanations in this section will refer to motivating reasons unless otherwise indicated.

The direction-of-fit conception of desire is well-suited to defend a Humean conception of motivating reasons. That is because of the essentially teleological nature of reason explanations. Any action done for a reason must be an intentional action. We don't attempt reason-based explanations for involuntary (unintentional) twitches or spasms. Intentional actions are goal-directed. To be motivated to perform an intentional action is then to be in a goal-directed state. Therefore the crucial question between the Humean and anti-Humean theorist of motivation is the question of which theory is best able to account for the fact that motivation for action consists, at least in part, in the pursuit of a goal (104).

The direction-of-fit conception of desire is ideally suited to this task. Anti-Humeans typically argue that beliefs of one sort or another can motivate without desire, or if desire is present, the desire is not the source of the motivation (Nagel, 1970, p. 29; McDowell, 1978, p. 15). But on this conception of belief and desire, the contents of beliefs tend to be adjusted so as to match the perceived features of the world, while desires tend to lead to the adjustment of the world to match the contents of the desire. It follows from this alone that beliefs cannot, by themselves, be the sole source of motivation for action, since they have just the opposite direction of fit that goals do. A person with a goal tends to attempt to bring the world in line with that goal, rather than bringing her goals in line with the world.[179] But of course the direction of fit of a desire is just the direction of fit of a goal.

So the argument for Humeanism about motivation, as Smith notes, is quite simple and powerful. It says

a) Having a motivating reason is, *inter alia*, having a goal.
b) Having a goal is being in a state with which the world must fit.
c) Being in a state with which the world must fit is desiring. (116)

I agree with him that the first two seem fairly 'unassailable', leaving only the third as a potential target. One way of attacking this argument we have already allowed to win, namely the claim that desires, in their normal sense, are not the only states with the appropriate direction of fit. We have acknowledged this and are simply calling all these states 'pro-attitudes' and 'desires'. As Smith says, this attack is not at the heart of the Humean view, but only a detail (117).

The second kind of attack comes from 'besire' theory, which claims that there are psychological states with both directions of fit. If so, then they could constitute being in a goal-directed state, but they would not be desires, since desires have only the one direction of fit (118). The heart of Smith's argument against it is as follows.

A besire theorist must claim that 'it is *impossible* for agents who are in a belief-like state to the effect that their Φ-ing is right not to be in a desire-like state to the effect that they Φ; that the two cannot be pulled apart, even modally' (120). But it seems that people's beliefs both about prudence and moral rightness do not *necessarily* motivate, but

[179] This is not to deny that goals are sometimes adjusted to facts about the world. It is just to say that the dispostions associated with goals are dispositions to adjust the world so as to realize the goal.

only ceteris paribus, i.e., if the motivation is not blocked by weakness of will or body or some such thing. This argues against the 'entailment' from the belief-like state to the desire-like state. More damaging is an argument against the entailment from a lack of the desire-like state to a lack of the belief-like state. For example, McDowell (1978, p. 18) argued that a virtuous person has some beliefs which are necessarily accompanied by appropriate dispositions of her will (121). Suppose this virtuous person, upon perceiving that someone is shy and sensitive, necessarily has the disposition to treat that person protectively, or to make special efforts to make her comfortable. Then it follows that should the virtuous person be overcome by weakness of will or some other motivational malady, and thereby were no longer motivated to care for the shy, sensitive person in the normal way, then it would follow that she no longer was able to see that the person was shy and sensitive. And this seems ridiculous.[180]

This concludes Smith's argument that motivating reasons are, among other things (such as means-ends beliefs), desires. But this claim does not defeat the anti-Humean! For as R. Jay Wallace emphasizes, the 'central point between the Humean and the rationalist … is the extent to which rational processes of thought—those which are governed by rational principles or norms—can contribute to the explanation of motivation' (2006, p. 20, n.12). And from the argument above (what Wallace calls the teleological argument), we cannot conclude that motivation cannot be accounted for entirely by adherence to rational norms. In order to conclude this, we would seem to

[180] Personally, I do not think these arguments against besire theory are as powerful as Smith does, largely because I think he does not distinguish between not having much motivation, or sufficient motivation for action, and not having *any* motivation. However, though I am somewhat sympathetic with 'besire theory' in some form (specifically, where besires are tokens, not types, of psychological states; see Michael Bedke, 'A Case for Besires' (unpublished manuscript)), I will not be advocating it here.

need a further assumption that these rational principles are excluded from playing an explanatory role to the extent that desires play such a role (20). But if some desires are explicable as the pure products of reasoning, then the explanation of the actions in which those desires issue will not stop with that desire, but with the reasoned belief(s) that gave rise to the desire. And in such a case, not only the motivations, but the justifications for the actions would also seem to rest upon the reasoning that led to the belief, as opposed to grounding out in the desire.

This can appear to be a big problem for me. For now, I'm just going to explain why this is so, before making a promissory note to deal with it later in the chapter. The problem is this. I've argued that evaluative beliefs are commitment devices, a kind of 'technology' for the generation, stabilization and strengthening of commitments.[181] But even if one were to agree that some, or many evaluative beliefs were commitment devices, it might seem that they could not all be. For to say that they are commitment devices seems to presuppose that there is always a commitment or other desire held prior to its associated evaluative belief. This is because I'm arguing that the evaluative beliefs are commitment devices in the sense that they either generate or strengthen commitments with respect to some particular antecedent desire(s). But this seems problematic, for it is apparently the case that, at least sometimes, we are argued from one evaluative belief to another by what appear rational means. Prior to being convinced of this new evaluative belief, we did not already have it. And if we did not already have it, then there seems to be no non-question-begging reason to suppose that there must have been some prior

[181] That's not to say they *are* commitments. It's important to be clear that they are not themselves commitments, for then I would be claiming that a class of beliefs are also desires, which would either be incoherent or would in effect be the claim that they are really neither, but rather besires. But this is not what I am arguing.

desire (commitment or not) to do the thing that my new evaluative belief says is the right thing to do.

This would be big trouble for me, since I am advocating a neo-Humean view of motivation and justification. For even if we accept that evaluative beliefs are commitment devices, the desires to which the beliefs commit us had better be, in some robust sense, prior to the beliefs. That is because if an evaluative belief were to generate a novel commitment (associated with no prior desire) in accordance with it, then our explanation, and plausibly our justification, of a person's action in accordance with that commitment would make ultimate reference to the evaluative belief and the reasons for which it was held by the agent, not the commitment that was consequent on that belief, especially if the commitment itself was rationally required, given the evaluative belief.

Several rationalists have maintained a thesis along these lines, i.e., that even if it is true that desires are always required for motivation, it is not the case that those desires must be the ultimate source of the motivation. The judgments of our pure practical reason can generate desires, at least insofar as we are rational, in accordance with those judgments. And this allows for the possibility of both explanation and justification of our actions that makes no (ultimate) reference to desire. Wallace employs just such a strategy in offering a rationalist response to the Humean teleological argument, drawing on Nagel (1970). I will address that argument in due course, but it will have to wait until after the elaboration of my view. For now, I just want to acknowledge my awareness that the teleological argument given by Smith does not in fact by itself defeat a rationalist about motivation, and I promise to address this issue (much) later.

3.1.3 Attractions of Desiderative Conceptions of Practical Reason

Smith has helped us to see that Humeanism about motivating reasons is compelling, even if we have yet to cinch the case for it. Unlike Smith, I (like many others) find neo-Humeanism about normative reasons attractive as well. I say neo-Humeanism because I want to remain uncommitted to any particular claims about Hume's actual views. What I want to retain is what I take to be the core of Hume's views about practical reason, which also seem to be the features that have inspired the tremendous amount of interest in these views in philosophy and the social sciences. I take these two core ideas to be that desires are never intrinsically irrational and that all reasons for action are in relation to some desire(s).

In conjunction with the claim that desires are required for motivation, these two features account for the attractiveness of neo-Humeanism about practical reason. Humeanism about motivation says that beliefs alone cannot motivate; desires are required.[182] This sits well with a division that Hume made popular and plausible, that beliefs are in the business of judging of matters of fact, i.e., they aim at the true. Desires on the other hand aim at realization. We saw above that this view of belief/desire psychology captures the respective directions of fit of these two psychological states. This view of belief/desire psychology leads to the view that desires are not themselves truth-evaluable, or rationally criticizable in the way that beliefs are. For a desire to be

[182] Further, as noted above, Humeanism seems committed to the proposition that not only are desires required, but that explanation of motivation must make *ultimate* reference to desire and not rational principles or norms.

'mistaken' on this view of the respective roles of belief and desire, it seems as if some relevant belief related to the desire must be mistaken.

Plausible though these claims are to many, of course others reject them. Perhaps the least controversial aspect of practical reason that neo-Humeanism can deliver quite easily is the fact that judgments of practical reason typically, if not always, have the power to *motivate*. This follows straightforwardly from the fact that, on a Humean view, reasons are always *instrumental* to desires, which are intrinsically motivating. The claim that practical reasons *can be* instrumental is accepted by almost everyone. Disagreement tends to arise over whether that is *all* it is. Another feature of practical reason that many contemporary philosophers find very plausible is that different people can have different reasons for action. That is, there can be considerable *diversity* in the reasons people have to act. This diversity is most neatly explained in terms of the fact that different people have diverse concerns. There seems to be a wide variety in the things that people care about, and many (if not all) of these things do not seem to be intrinsically irrational.

Another reason for being attracted to neo-Humeanism stems from the idea that, at least sometimes, a motivating and justifying reason can be the same reason. It seems very plausible, if not a conceptual truth, that '…when an agent acts for a (specific) reason that very reason is also the explanation (or at least part of the explanation) of why she did what she did. Normative or justificatory and explanatory reasons are the same reasons in such a case and not different kinds of reasons altogether' (Heuer, 2004, p. 45) . Dancy (2000), Korsgaard (1986), (Garrard & McNaughton, 1998) and others have argued that the 'fact' that normative and explanatory reasons can at least sometimes be the same reasons shows that Smith was wrong to argue that they are categorically different kinds of reasons. These authors use this claim to argue against Smith's 'psychologistic' account of motivating reasons as causes. Since normative reasons are (ostensibly) not causes, and normative reasons can be the same as motivating reasons (when the agent is in 'sound normative shape'), motivating reasons must not (always) be causes.[183]

I think the claim that motivating reasons and normative reasons are not categorically different is correct, even, as Heuer claims, a truth that 'leads right into the centre of our understanding of rational agency' (2004, p. 59). But I think this argues not *against* motivating reasons as causes, but *for* normative or justifying reasons as potential causes. That is, the strength of motivational Humeanism, specifically the version in which reasons are causes, combined with the thesis that motivational and normative reasons can be the same (or at least intimately related), gives us reason to pursue a conception of (at least some) normative reasons as potential psychologistic causes. Neo-Humean views see normative reasons as relative to desires, and if we see desires as causes, then normative reasons will be relative to some subset of the desires-as-causes that move one to act. Such a view, if it can be made plausible, promises to capture what I agree is the intimate connection between explanatory and justificatory reasons. I will be doing my best to make this view plausible in the following sections.

Finally, neo-Humean conceptions of practical reason accommodate modern orthodox rational choice theory and rational decision theory quite well. This is a considerable advantage insofar as these theories are the most-developed and systematic accounts of how various, differently ranked ends, and (limited) information from multiple

[183] Lenman, James, "Reasons for Action: Justification vs. Explanation", The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/reasons-just-vs-expl/>.

sources, all contribute toward rational decisions. They do this by conceiving of rational agents as having sets of desires (or preferences) and beliefs (subjective probabilities) about the relevant facts. In orthodox versions, a rational choice is one that promises to achieve an outcome at least as high in the agent's preference ordering as any alternative. The 'Hume-Ramsey' theory of rational choice is also utilizable in empirical inquiry, insofar as making assumptions of rationality in its sense lets us infer 'revealed' preferences as well as beliefs (Railton 2006, p. 270).

We have already seen that there are significant problems with the empirical uses of such theories, despite their widespread usage in economics, cognitive science and elsewhere. We have focused on these problems especially insofar as they do not recognize the ubiquitous phenomenon of preference-switching as a result of hyperbolic discounting (but not only as a result of that). However, the nature of these failures does not, at least not obviously, reduce their support for neo-Humeanism, though it does complicate matters. For the Ainslean criticisms of orthodox rational choice theory imply not only a multitude of ends (and subjective probabilities) with preference-orderings, but that those preference-orderings (at least with respect to concrete objects of desire) can and do change over time. The Ainsliean critique problematizes the idea that there is a single 'rational' agent with stable preferences that can, in principle, be acted on in a perfectly rational fashion. Instead, there seem to be a constellation of desires ('interests') at different time-ranges. If preferences can change (and change back) in this way, then there will be no straightforward sense in which it will always be rational to act on one's preferences. One simple but quite deep problem is that very many of these temporary preferences will be for objects that we simply will not (stably) regard as rationally desirable.

If we could solve this and neo-Humeanism's other ostensible problems, it would be a satisfying and worthwhile accomplishment. Neo-Humeanism offers many attractions, and is, especially if we go beyond philosophy to the social sciences, cognitive science and psychology, the dominant view of practical reasoning. It nevertheless has received intense philosophical criticism from a number and variety of sources. If these criticisms could be rebutted, or even significantly blunted, neo-Humeanism in some form should be considered the default champion, especially in the absence of another contender that offers as many or more attractions with less cost. The remainder of this chapter will be concerned to defend a specific version of neo-Humeanism, a version I will follow Stephen Finlay in calling end-relationalism.[184]

I believe that the complication(s) that the Ainsliean model of the will introduces into neo-Humean conceptions of practical reason are also the source of considerable explanatory power in accommodating some powerful intuitions that seem to tell strongly against neo-Humeanism. In addition, neo-Humeanism needs to be conceived of somewhat differently than it is normally done, so as to avoid other common objections. Fortunately, my end-relational conception (which I will develop below) fits very nicely with the Ainsliean model of the will. It also fits nicely with the general picture I

[184] We will discuss Finlay's views below, but for the moment I want to say that Finlay's view is about the semantics of claims about practical reason, while mine is a view about what reasons actually are, or perhaps more perspicuously, what is really going on when we make practical reason claims. Although the view is broadly Humean, I prefer the term end-relationalism because it avoids some of the central conceptual mistakes of classical Humeanism, e.g., that one's strongest desire is one's strongest reason. More on this below. Still, I consider end-relationalism a species of neo-Humeanism, and will continue to refer to it as such.

described from Boyd and Richerson (2005) of humans as inherently conflicted due to their having evolved multiple deep and powerful sources of motivations that often run contrary to one another.

Broadly, my strategy will be two-pronged. The first prong will be to develop my view and defend it against criticisms. The second will be to criticize severeal leading rationalist proposals, explaining the *apparent* attractions of those views within the apparatus of my own. As we will see, much depends on how we conceive of the relationship between willpower and rationality. To get started, I want to describe some deep and systematic critiques of any possible desire-based conception of practical reason. My view will be developed and defended in the course of answering them.

3.2 Criticisms of Desiderative Conceptions of Practical Reason

David Brink (2007) has recently provided a systematic and wide-ranging attack on desiderative conceptions of practical rationality as well as of one's own well-being, in favor of a perfectionist view. Brink's arguments against a desiderative view of practical rationality are especially appropriate to address for three reasons. The first is that they are clear, very general attacks right at the heart of any (reductive)[185] desiderative view. The second is that an important part of those attacks relies on some intuitive appeals to categorical goods and imperatives. I think that the resources I've been developing in the

[185] A 'reductive' view analyzes or 'reduces' facts about reasons to non-normative, or value-neutral, facts about desires. A non-reductive view would attempt to employ desires as part of an analysis of reasons, but import normative features into the analysis. If one were to analyze reasons for example in terms of what it would be good to desire, this would be non-reductive. Brink, rightly in my view, holds that such non-reductive efforts run serious risks of circularity and/or a lack of explanatory value (23-26).

prior two chapters and will develop here can help to explain why his alternative perfectionist view would be attractive, given the Ainsliean framework that helps motivate my desiderative view. I'll also take on some of his arguments against desire-satisfaction views of well-being, since given my particular view, those are also applicable. The third reason I focus on Brink's arguments is that he conceives of desires on the 'direction-of-fit' model. This is helpful insofar as his arguments are explicitly directed against the conception of desire with which I'm working. Arguments against other forms of desiderative views would not necessarily apply to my view and/or my responses would run the risk of moving past rather than meeting these challenges.

3.2.1 Against The Normative Adequacy and Authority of Desire

The heart of Brink's arguments against desire-satisfaction views center on their putative failure to provide a satisfactory account of what Brink terms the 'normative adequacy' of desiderative conceptions of practical reason. There are two dimensions along which an account of practical reason can fail or succeed to provide such an account. The first is normative accommodation. This dimension is concerned with the particular guidance that a view of practical reason gives us. That is, '[h]ow well does it accommodate what we are prepared, on reflection, to think about the normative valence of various actual and hypothetical situations?' (26). The second dimension is normative authority. The question of authority is the question why we should care about or be governed by the advice that a conception of practical reason gives us. An account of practical reason 'must provide a *rationale* for the normative authority of its demands'

(27, emphasis his).

We start with the problems of normative accommodation.[186] Simple desire-satisfaction views first of all don't seem to provide for 'robust forms of fallibility' (26). They seem to give significance to desires based on erroneous beliefs about how to satisfy other desires, as well as on faulty inferences. Also, we often seem not to have desires to do things that we do in fact have reason to do (27). Among the things that we *seem* to have reason to do that we may not have desires to do are 1) comply with basic moral duties and 2) be temporally-neutral. Intuitively, the sheer fact of not wanting to be moral, or not wanting to be temporally neutral at some given time does not affect the rationality of these things (especially the latter, one might think, for rationality and morality are perhaps not so tightly linked conceptually as rationality and prudence).

Moreover, if something's being rational depends on its being desired in some way, then it seems as if there can be no incorporation of normativity into the content of acceptable (or acceptable content) of desires.[187] This appears to entail 'unrestricted adaptation' of one's desires to the circumstances so as to maximize the number of one's desires that are likely to be fulfilled. For instance, if one could make it that one's greatest desire was to collect lint, and then went on to be a highly successful lint-collector, and thereby satisfied one's desires fully, that would not seem to make the person's life good, or the resultant actions rational. Brink also evokes the Deltas and Epsilons from

[186] I am going to focus on the problems that apply to 'simple desire-satisfaction' views, since I aim to defend such a view. These are opposed to idealized desire-satisfaction views, which filter desires through some idealized version of oneself, typically armed with full information and rationality. I'll argue that much (but not all) of the appeal of idealized desire-satisfaction views can be claimed by simple desiderative views, without the costs.

[187] If the acceptable content included any normative notions, the view would seemingly not be genuinely reductive.

Huxley's *Brave New World* to suggest how contemptible an image of humanity such a view seems to entail, or at least allow for. The fact that they have an extremely low bar for having their desires satisfied doesn't seem to make their lives better, but worse. We should not accept a view that recommends lowering ourselves by lowering our desires in this fashion.

Now we move to the second dimension along which Brink argues that desire-satisfaction views fare poorly; they fail to provide an account of normative authority. The idea is simple; 'it is not at all clear why we should care about the satisfaction of desires independently of the way in which they were formed or of their content' (32). One might think that this objection only has force against simple desiderative views, i.e., those that hold that practical reasons are given by our actual desires. But most desiderative conceptions of practical reason appeal to some form(s) of *idealization*. Idealized desire views attempt to do what simple desire views *appear* not able to do, namely to accommodate robust forms of fallibility in our desires. By filtering our desires through processes such as full exposure to relevant information and logic (Brandt, 1979) or what an epistemically idealized version of the agent would want for her non-idealized self (Railton, 1986), idealized desire views attempt to retain the attractions of desiderative views without falling afoul of deep intuitions that just any old desire provides us with a reason to act on it.

However, Brink argues that all forms of idealization fail to render desiderative views capable of passing the test of normative accommodation. In addition, they take on deep problems that don't afflict simple desire views. I won't discuss these additional problems, but I will say that I agree that they are very serious and render idealized desire views unattractive. But I do want to say why Brink thinks idealization fails to improve on the problem of normative accommodation. The essential point is that certain things, such as temporal neutrality, as well as emotionally and intellectually rich lives, are *unconditionally* good, that is, whether one wants them or not. But it is ultimately contingent whether an idealized version of yourself would in fact desire these things (for you). If we are tempted to respond that an idealized version of ourselves couldn't or wouldn't want us to lead shallow, stupidly contented lives, then we are simply presupposing rather than explaining aspects of rationality (35). If 'the life of a contented swine' (35) is *categorically* bad, then an account of rationality that leaves it an open question cannot be right.

3.2.2 Response

Normative Accommodation

I will defend a simple desiderative conception of practical reason. I have two primary reasons for doing so. The first is that to most people, simple desiderative views are less plausible than idealized versions. For just the sort of considerations that Brink raises concerning the variability and contingency of people's actual desires, it can seem quite implausible that what it is rational for us to do can depend on them.[188] However, if this is the most implausible form a desiderative view can take, and it can be well-defended, then more plausible ones will be yet more plausible. The second and more

[188] This sense of implausibility lessens when we think in terms of pro-attitudes rather than desires, I gather, but still does not go away.

important reason is that I think this one is the most plausible, not the least.[189] In order to chip away at the sense of implausibility attaching to it, the first thing I need to do is make the following distinction between objective and subjective instrumental reasons[190] as presented in Joyce (2001, 53):

> Objective Reasons:
>
> S has an objective reason to Φ if and only if Φ-ing will further S's ends.
>
> Subjective Reasons:
>
> S has a subjective reason to Φ if and only if S is justified in believing that she has an objective reason to Φ.

Typically, when we criticize people for failures of practical rationality, we criticize them on the basis of their failure to follow their subjective reasons. However, a number of philosophers[191] conceive of an ideally rational agent as one who follows not their subjective, but rather objective reasons. This builds in ideal epistemic circumstances into one's conception of practical rationality. I find this a potentially confusing way to speak of rationality. Certainly it is absurd to praise or blame someone for their failing to act on the basis of reasons of which they are justifiably unaware.

[189] I don't think idealization can deliver the increased plausibility it promises, at least not enough of an increase to be worth the serious costs.

[190] It's important to note the objective and subjective reasons here are both *instrumental*. Brink deploys the notion of an objective reason in his target article in an anti-instrumentalist way: 'a conception of practical reason or the good is objective just in case it identifies things as reasonable or valuable independently of being the object of the agent's actual or informed desire' (21).

[191] E.g., Brandt (1979), Parfit (1984), Smith (1994).

However, defenders of objective rationality can distinguish between 'rational criticism' and praise or blame, reserving the latter for failure to act on subjective reasons, but allowing the former to encompass 'failure' to act on objective reasons. Fortunately, we can avoid confusion without insisting on one or the other conception of practical rationality by simply making sure we're clear which sorts of reasons we take agents to have, and therefore in what sense they are irrational, if and when they are.

The second step toward reducing the apparent untenability of a simple desiderative view of practical rationality is to restrict this view to underived desires (that is, to them and derived ones only insofar as they contribute to underived ones).[192] If I want to drink the liquid in the glass because I'm thirsty and I think it's water, then my underived desire is quenching my thirst, whereas my derived one is drinking this particular liquid. Notice we can already see a sense in which there can be fallibility. First, the liquid could be gasoline, though I think it's water. Now, to answer the question whether it's rational to drink it, there are two things to consider. First, do we mean objectively or subjectively? In this case, that is perfectly clear; there is not objective reason to drink (and objective reason not to, one supposes). There is subjective reason to drink iff I'm justified in thinking there's water in the glass.

The distinctions between subjective and objective reasons and between derived and underived desires provides us with two sources of fallibility. First, we might have subjective, but not objective reason to act, whenever we are justified but mistaken in what will serve our ends. Second, we might not have subjective reason to act, in the sense that

[192] Cf. Parfit (1984, p. 117).

we are not justified in believing that some action will serve our ends. Our derived desires might be based on mistakes of fact or reason, and these mistakes might be rationally justifiable or not. These are sources of fallibility that my simple desire model can countenance. Further, we can see that the basic desire view under consideration here does not ‘give significance’ to desires based on faulty inferences or mistaken beliefs. For it to do that, there would have to be underived desires that are based on some such mistakes. I know of no compelling reason to think that underived desires are based on any inferences or beliefs, mistaken or otherwise.

Suppose it is granted that the simple desire view doesn’t give significance to desires based on factual or reasoning errors. It can still be maintained that the kind of fallibility discussed so far is not ‘robust’ and/or not of the right kind. The first kind of fallibility, that of acting on one’s subjective but not objective reasons, is not, as I noted above, the kind of thing for which we can reasonably *criticize* people. We might be able to *correct* people in such cases, but not criticize them as having committed a rational failing. And though we are rightly concerned with a person’s objective reasons, and can appropriately advise or correct them from an epistemically superior position, it also seems true that people are sometimes, often in fact, appropriately subject to full-blooded *criticism* (where that includes a kind of blame) as to their choices and actions, from the standpoint of practical rationality.

The second sort of fallibility illustrated does promise to leave plenty of room for criticism of a person’s rationality, but one can object that the kind of rationality that appears to be at issue in these cases is that of theoretical, not practical rationality. Suppose we see a water-seeming liquid poured into a glass from a tap. Under normal circumstances, we are justified in thinking it is water. If we see it poured from a containter labeled 'gasoline' or 'HCL', presumably we are not. But questions as to the justification, or rationality of beliefs are questions of theoretical rationality.

This was part of the reason for Hume's (commonly interpreted) skepticism about practical rationality. On his view, only desires, not reason, can motivate. What we call 'practical reasoning' is only instrumental reasoning, concerned with questions as to discovering efficient means to given ends. But which means are efficient routes to which ends are facts which are true or false independently of anyone's desires or actions. These facts are true or false, not good or bad. The kind of rationality that pertains to such facts and their discovery is theoretical. Therefore there is nothing for 'practical rationality' to be, beyond theoretical rationality applied to one's desires. But many people, including Brink and myself, think that there is such a thing as practical rationality, and therefore a view which implies that there is not must be wrong. Therefore it is important for me to say how it is that my view doesn't imply that there is no practical rationality.

Accommodating Akrasia

Though we have allowed for two forms of fallibility, which might well encompass most instances of practical error, the view on offer does not yet seem to countenance at least some kinds of robust fallibility, or criticizability, that judgments of practical rationality often imply. In fact, the kinds of fallibility that we've seen that the view can countenance do not seem to be paradigmatic cases of practical irrationality at all. Turning to those cases will return us to one of our central themes.

For many philosophers, (synchronic and/or diachronic) akrasia is the paradigm of practical irrationality. While I do not discount the possibility of synchronic akrasia, I take it that its instances are vanishing in comparison to diachronic akrasia. Socrates denied the possibility of the former, largely based on a realization that people in fact (for him, always; for us, perhaps only overwhelmingly) judge their actions to be reasonable, or at least do not judge them irrational or wrong, *at the time of action*. We are all more familiar than we would like to be with the phenomenon of ‘rationalization’. This suggests that what we do at these times is to convince ourselves that our aims at the time are rational, or at a minimum not irrational (or wrong) on the whole. And that means that we think that certain ends *can* be irrational, and further, we often judge both prior to and after the fact that some of our actions would be and/or were irrational, and this does not seem to be due to mistakes involving theoretical irrationality, or epistemic limitations.[193]

The simple desire view offered here does not countenance the fallibility of underived desires, nor the *intrinsic* irrationality of any desires. But it seems that there are at least many cases of weakness of will, arguably the paradigmatic form of irrationality, wherein the SS desires that sometimes outcompete LL desires are (also) underived. The desire for (the pleasure of) a cigarette doesn’t seem derivative on any other desire. And if it is not intrinsically irrational, then we will presumably have to say it is ‘merely’ relationally irrational, that is, relative to the LL desire.

If I were advocating not a basic, but some form of idealized desire view, I could pursue a strategy of claiming that such desires are irrational in relation to the desires they

[193] Though Socrates disagreed. His diagnosis was precisely that people who acted akratically do not, even prior to the action, *know* what the right thing to do is. Rather, they believed that something was right, but without having those beliefs secured by correct reasoning (and thereby turned into knowledge), they were susceptible to switching as desires became stronger.

would have under some idealizing condition(s). If their desires were made coherent and/or if they had full information and/or were deliberating soundly or something else, then they would not have this desire, and since it is inconsistent with one's suitably idealized desire set, which is identical to one's fully rational desire set, then this desire would be irrational without being intrinsically irrational.

I don't like that strategy though, because it takes on significant problems that would only be justified by the ostensibly greater plausibility they offer over basic desire views. If I'm going to show that this claimed greater plausibility is an illusion, I'll need another answer for the person who wants to know how to understand the prima facie irrationality of some of our underived desires, many of which seem to figure into episodes of weakness of will. In fact, I still need to say why my view doesn't entail skepticism about practical rationality, which I maintain it does not. Doing so will require and provide an opportunity for me to deploy an improved conception of 'instrumental rationality'[194] that will avoid many common criticisms of basic desiderative views. For this, I'll have to postpone my responses to Brink's criticisms in order to turn to another influential critic of Humean conceptions of practical rationality.

3.2.3 Korsgaard, the Will, and Practical Rationality

Christine Korsgaard (1986) is perhaps the locus classicus for a rationalist argument against Humean skepticism of practical reason.[195] Korsgaard sees skepticism

[194] I put it in quotation marks because I call it end-relational rationality.

[195] Brink also provides an argument to the effect that Hume denies the existence of practical reason as part of his critique of basic desire views. Joyce (2001), drawing on Korsgaard (1986) and (1997) advocates a

about practical reason as skepticism that there can be what she calls ‘true irrationality’. By this she means ‘a failure to respond appropriately to an available reason’ (378). As we saw, on a common interpretation of Hume, the only kinds of mistakes people make in the arena of ‘practical reason’ are those involving errors of fact and/or relations of ideas. But as we also saw, drawing good inferences from mistaken facts is not in itself irrational, and mistakes of reasoning belong to theoretical, not practical rationality. However, Korsgaard notes that Hume is ambiguous in claiming that a passion can be unreasonable ‘when, in exerting any passion in action, we chuse means insufficient for the design’d end, and decieve ourselves in our judgment of causes and effects’ (T416, quoted on 377). There are at least two readings of this passage.

Hume might mean that we are caused to act in part by a false belief about causes and effects. This would not be a case of genuine practical irrationality. But Hume could also mean that a person could know which causes would lead to his desired end(s) and choose insufficient means and/or not choose obviously available and sufficient means. In this case, we would have an example of ‘true irrationality’ according to Korsgaard. This irrationality would consist in a ‘failure to be motivated by the consideration that the action is the means to your end’ (378). Even someone who thinks that instrumental rationality is all there is to rationality should allow for this form of ‘true irrationality’ (378).

And there seem to be many things that can disrupt such motivation. ‘Rage, passion, depression, distraction, grief, physical or mental illness: all these things could

non-Humean instrumentalism based largely on Hume’s purported denial of practical reason generally and weakness of will in particular.

cause us to act irrationally, that is, to fail to be motivationally responsive to the rational considerations available to us' (378). Suppose for the moment that we exclude an interpretation in which true irrationality is possible. Then it seems that the Humean cannot count 'failures' of prudence as rational failures at all. Korsgaard asks you to suppose that you have a choice between your greater and lesser good and you choose the lesser. Since true irrationality is excluded, this would be evidence that you either do not have your greater good as an end or it is at least not as important to you as your lesser good. Of course, if you are motivated by the option that leads to your greater good, then we take that as evidence that you do care about your greater good. This implies that one's greater good is an end one might care about or not, and one's reasons are relative to these concerns (379-80).

However, since we have 'seen' that there seem to be many routes to true irrationality, Korsgaard claims that there is no particular reason to accept this view. Whether one accepts it, Korsgaard says, depends on whether one already accepts that reason is only instrumental. If so, then one will interpret the choosing of a lesser good as arising from a *'stronger desire'* for it, and *so a stronger reason'* (380, my emphasis). If one does not limit oneself to instrumental rationality, one is likely to say that to choose the lesser good is a case of true irrationality, because prudence is unconditionally rational. Given that Korsgaard has already purported to show that there are myriad ways that one could be truly irrational even where that is limited to instrumental irrationality, there seems to be no special reason to suppose that true irrationality must be limited to that one form. Further, given the counterintuitiveness of the suggestions that one might, with equal rationality, choose one's lesser good over one's greater, and that wherever a *motive* for one's greater good disappears, so does the *reason* for it, it seems that we should be willing to find true irrationality in both instrumental and noninstrumental contexts.

There are two obvious strategies to resist this conclusion. The first is to deny the possibility of (true) instrumental irrationality. The second is to admit it but argue that there are good reasons to think that true irrationality is limited to instrumental irrationality. The first strategy might seem hopeless. To deny the possibility of true instrumental irrationality seems to deny, among other things perhaps, the possibility of weakness of will.[196] Since a stronger desire necessarily provides a stronger reason, and one necessarily acts on one's strongest desire, then one necessarily acts for one's strongest reason. Therefore there is no weakness of will. But we all have had the experience of weakness of will, so this strategy seems hopeless. Further, it seems to deny genuine practical reason.

Despite these seeming difficulties, I will maintain that there is no such thing as 'true' *instrumental* irrationality, as Korsgaard understands it. That is, there is no categorical requirement to be motivated to take the means to one's ends. I will however grant that there is 'true irrationality' in the sense of a 'failure to respond appropriately to an available reason'. But I will understand what an 'appropriate reason' is in an end-relational fashion. I will not deny the possibility of weakness of will or of practical reason generally. On the contrary, an improved understanding of weakness of will will help us understand why no such categorical requirement of reason exists. Further, I see

[196] In footnote 10, Korsgaard says that weakness of will, in addition to self-deception and rationalization, are what she has in mind as causes of (true) instrumental irrationality.

no reason to deny practical rationality, but rather to improve our understanding of what practical rationality consists in. I'll argue that reasons are end-relational, and that judgments of (ir)rationality are made from the perspective(s) of some of our ends, typically our long(er) term goals. The end-relational view holds not only for reasons but for normative judgments generally. This will be as robust as fallibility gets, which I hope to convince the reader is plenty robust, though getting used to 'merely' relational normativity will take some … getting used to.

There is no rational requirement to be motivated by one's desires

Finlay (2008) argues (persuasively in my view) that there is no instrumental norm of reason, where that is understood as a requirement to *will* the means to the ends that one has *willed*. He shows that no attempt to make sense of this requirement has been able to satisfy two conditions, and argues that none ever will be. First, it must be possible to violate the requirement, and second, the requirement must meet a 'reasonable endorsement' condition. For something to be a categorical requirement of reason, it must be that a reasonable person thinking clearly can see that (if not why) it is such a requirement. Finlay shows that attempts to meet one condition run afoul of the other. The same arguments apply to an instrumental norm of reason, conceived of as a requirement to be *motivated* to satisfy one's *desires*. Any *plausible* 'requirement' of this sort will not be possible to violate, and therefore is no (normative) requirement at all.[197]

[197] I'm indebted to Finlay's discussion, especially pages 24-26, for help in framing the end-relational view presented here.

On the conception of 'desire' (pro-attitude) employed here, to have a desire just is to be disposed to fit the world to it, which just is to be disposed to take the available means. I've said that this is analytically necessary on our conception of desire. This might make it seem like there's something wrong then with that conception, since it seems we are certainly not always motivated even to take the means to our desires (which I will also refer to as 'ends' in this connection). We might have a desire to be healthy but not be motivated to take the means of stopping smoking to achieve that end. My aim in the following few paragraphs is merely to blunt the potential reaction that something is going badly wrong, before moving on to a substantive engagement with Korsgaard.

On our conception of desire, something counts as a desire just in case we are disposed to (attempt to) bring the world in line with the content of the desire. To make such an attempt entails that one is motivated to do so. So to have a desire just is to be disposed to be motivated to realize the desire. And to be disposed to be motivated to realize a desire just is to be disposed to be motivated to take what one regards to be the available means to bring the world in line with it.

Suppose one is mildly thirsty and there is a cup of water on the table. If one wants the water, then ceteris paribus one will be motivated to get it. The ceteris paribus clause marks the fact that desires are *dispositions* to be motivated to bring the world in line with them. For if the water is in an electrified metal cup, one might lose this motivation, and indeed gain motivation to stay away from it. If one gets much thirstier, one will presumably have both motivations, to drink and to stay away, which will be in conflict with each other. This is very often the case (not with electrified cups, but more generally). We might have some motivation to study, but effectively more motivation at the time to party. Having conflicting motivations is something everyone can relate to. And it should not be inferred from the fact that one motivation is stronger at a time that the other motivation does not exist. And it should not be inferred from the fact that one is only weakly motivated that one is not motivated.

However, as we saw, it is plausibly the case that we can have desires without being motivated at any particular time to take the perceived means to fit the world to them. Perhaps *this* is the categorical requirement then, not to simply be *disposed* to be motivated, but to *actually* be motivated to take the means to our ends. But of course this is wrong, as we can see with the electrified cup. If I am mildly thirsty and so would like some water, but I know I will be severely shocked if I attempt the means to satisfy my thirst, it is not irrational not to be motivated to do so. And it can't be irrational not to be motivated under the manifestation conditions for the desire, since this is impossible, analytically. If the manifestation conditions are realized, then one will be motivated, analytically, if the desire-attribution is accurate.[198]

We might then be tempted to impose 'appropriateness' conditions of manifestation under which it is irrational not to be motivated. This makes no sense either. If one is not motivated under some circumstances, then either one does not have the relevant desire or those circumstances do not fall under the manifestation conditions for the desire (for many desires, attending to them will be part of the manifestation conditions). In addition, attempts to build in substantive appropriateness conditions would seem to undermine the ostensible uncontroversiality of the categorical

[198] There is no suggestion here that these conditions are known, or need to be known in order to attribute a desire to someone.

requirement. Attempts to build in requirements of strength of motivation will have the same problem.

We saw that it is not necessarily irrational not to (be motivated to) drink the water if the cup is electrified. If one is dangerously thirsty, and the shock will be mild (and one knows it), then our judgment of rationality changes. Now it does seem irrational not to. It's hard to imagine not being motivated by strong thirst, but we can imagine a desire for healthy teeth that does not lead to a motivation to go to the dentist. But when we say this, we might mean that we have no motivation to go to the dentist in and of itself. Or one might not be convinced of the (causal) requirement to go to the dentist in order to have healthy teeth, or one might not be attending to one's desire, or the causal necessity of the dentist, or one might be confusing the fact that one has a (much) greater motivation not to go with the idea that one has no motivation at all.

It is both intuitive and part of the account of the will developed here that one's attention plays a very important role in which desires motivate one at a particular time. If one's attention is on the pleasures of smoking, the motivation to smoke will tend to be strengthened, and if one's attention is on the long-term harms of smoking, the motivation not to smoke will tend to be strengthened. A large part of the motivational effectiveness of desires stem from their tendency to repeatedly or even continuously draw our attention to them. Perhaps the categorical requirement then is to attend to all of one's desires, or more promisingly perhaps, one's long term desires.

The first option is of course out. It is obviously not rationally required to attend to all of one's desires. Is there perhaps a categorical requirement to attend to one's long-term desires (and thereby be motivated by them)? Surely not at all times. This is not only impossible but undesirable. Making love with one's attention on one's long-term desires doesn't seem rationally required. Then perhaps the requirement could be to attend to them when (one perceives that) they conflict with shorter-term desires.

This is more plausible. This looks like a requirement to take a moment to 'think about what you're doing' before you do something that might run counter to your long-term interests. When we 'stop and think' (in the prudential context at least) we try, perhaps inter alia, to distance ourselves from the felt tuggings of desire or passion, and attend to the likely long(er)-term consequences of our potential actions. Perhaps it is a rational requirement to do this. I think there is some truth in this, but that truth is not that there is a categorical requirement of reason to do so, but rather that our judgments of rationality, especially with respect to our own rationality, are typically made *from the perspective of our long(er)-term desires*. I hope to make this important claim more palatable in the pages to come.

The strategy just employed might have simply traded one problem for another. One strength of the 'direction-of-fit' conception of desire is that it allows me to deny that there is a categorical requirement of reason to be motivated by one's desires. A potential weakness is that a purely analytic-functional conception of desire seems to have no ability whatever to rationalize action.[199] Therefore, it might seem I am only avoiding one thing going horribly wrong just to substitute another. But my account does not suppose that the fact of having a desire in itself 'rationalizes' the action that would satisfy it. My view is that all reasons are relative to motivated perspectives, but not all motivations are associated with reasons, and certainly not with our 'stronger' reasons.

[199] See Warren Quinn (1993, pp. 236, 246-7).

Again, these remarks are not meant to be nearly sufficient to establish that an instrumentalist needn't countenance any non-instrumental rational requirement to be motivated to take the means to one's ends. Putting this requirement in terms of desires is not the usual way of understanding it anyway. Ends are normally understood in terms of something like intentions. I mentioned that Finlay (2008) has argued against this common understanding as well, but I won't recapitulate those arguments. Instead, I'll begin by showing how the end-relational view on offer can deliver a plausible conception of practical reason, including weakness of will. In its attempt to take on no objectionable baggage (such as idealizers take on), it opens itself up to the charge that it is unable to explain why it is irrational not to drink the water in the second but not the first case presented above, or why not going to the dentist is irrational. These (non)actions are, or at least are normally considered irrational, even if one has no false beliefs. Any account of practical reason that denies this takes hits to its plausibility, as does any view that supposes that the mere fact of having a desire provides an agent with a reason to satisfy it. I'll try to show that the view presented here avoids these and other difficulties, and is not only plausible, but more plausible than its contenders.

Strength of desire determines neither efficacy of motivation nor strength of reason

I presented a brief argument above that the Humean doctrine that we always act on our strongest desires and that our strongest desires constitute our strongest reasons implies that there is no such thing as weakness of will. But we all experience weakness of will, so the Humean doctrine must be wrong. The clear answer to this challenge is to reject the claim that we always act on our strongest desires or on our strongest reasons, both because it is wrong and because it makes for a much more plausible Humeanism.[200] It is wrong both because our strongest desires do not provide us with our strongest reasons and because in any case we do not always act on our strongest desires, at least if 'strongest' retains anything like its intuitive meaning. We should even more emphatically reject the claim that if neo-Humeanism is correct, then when we act on 'desires for our lesser good', that shows that we care more about our lesser than greater good, or even that we don't care about the latter at all.

Indeed, the causes by which Korsgaard maintains that we often act instrumentally irrationally show that this isn't correct. When we are in the grip of rage or some other passion, or our attention is distracted in some other way and as a result we act contrary to our greater good, it would be perverse to claim that it follows from this that we care, even at the time, more about our lesser good than our greater. Of course one might think, and many do, that this shows that Humeanism is perverse, since Humeanism entails this result. To the extent that this view is the result of Hume's ostensible commitment to the claims that we always act on our strongest desires and that our strongest desires constitute our strongest reasons, this reaction is understandable. But we can keep the heart of Humeanism without these claims.

Korsgaard alluded to self-deception and rationalization as quintessential mechanisms of irrationality. When we rationalize, in the context of prudential

[200] I am now separating from 'Humeanism' the claims that we always act on our strongest desires and that our strongest desires constitite or provide our strongest reasons. I think one can make a case that these aren't really central to Hume's actual views, but that is not my concern. From now on, by 'Humeanism' (about practical reason) I will only include the claims that desires cannot be intrinsically irrational (or unreasonable), and that all reason is instrumental (more accurately, end-relational, as I'll discuss below).

considerations, what we tend to do is precisely to try to convince ourselves that the desire we propose to pursue is in fact contributory toward, or at least not inconsistent with, our greater good. When we are successful in this, we often judge at a later time that we deceived ourselves. We make the same kinds of judgments about others. Alternatively, or in conjunction with rationalizing self-deception, our attention can be distracted by short-term desires such that we are simply not attending to what would in fact serve our greater good. In fact, 'violent' passions, as well as other forms of short-term rewards, have the ubiquitous feature of making insistent demands on our attention.

In summary, the ubiquity of rationalization and self-deception in these contexts points up the fact that even in apparently egregious cases of imprudence, there is no general reason to suppose that people intentionally act contrary to their greater good, while conceiving of it as such at the time.[201] Where these two sources of imprudence are absent, the myriad forms of distraction to which we are susceptible also help to explain why imprudence is not obviously to be explained in this way. Moreover, we can see that these mechanisms belie the claim that we always act on our 'strongest' desire. Hume opposed the 'violent' to the 'calm' passions, and it is true that very many cases of imprudence are ones where we are pulled by desires whose phenomenology lends itself to description in terms of strength. But, especially in modern life, we are plausibly seduced most often not by lust or rage (not that these have gone away!), but by the lulling distractions of television, the internet, and myriad other ready-to-hand diversions. Under these circumstances, it is often the case that our actions are determined more by our 'sneakiest' than 'strongest' desires.

[201] I'm not saying this never happens, only that it does not follow from the fact of egregious imprudence.

It might be objected that I'm misrepresenting Korsgaard's views. And that is correct, in a sense, but it is also for a reason. Korsgaard asked us to suppose that we are confronted with a choice, 'and though informed that one option will lead to your greater good, you take the other' (379). It is in these circumstances she claimed that absent true irrationality we would have to interpret you as caring not at all or less about your greater good than your lesser. But the point of the preceding comments is largely to illustrate how underdescribed and/or unusual such a situation would be. For the scenario to have the import she would like it to have, it would have to be a situation in which we recognized clearly that the option chosen would in fact be worse for us than the other. Not only that, but considerations of prudence would have to be the only relevant considerations, ruling out the possibility of some moralistic reasons (or rationalization). Is such a thing possible? I certainly won't try to rule it out, but my point is only to make it clear that the fact that we often act for our lesser good is not remotely so telling against the Humean as it might seem, once we take into account the ubiquity of self-deception, rationalization, and the myriad additional forms of distraction, in addition to rejecting the claims that we always act on our strongest desire and that our strongest desires constitute our strongest reasons.

But how can the Humean reject the claim that our strongest desires constitute our strongest reasons? Or rather, if he does reject that, then what *do* constitute our strongest reasons? The reader might agree both that our strongest desires do not determine our actions (again, where 'strongest' is thought of in its typical sense),[202] and that very often

[202] I suggested above that its 'typical' sense was bound up with phenomenology. This needn't be the case for the rejection to go through. I think it can be rejected on a 'quasi-hydraulic' theory of belief and desire

even our strongest desires, the strong passions, operate in part by distracting our attention away from where our greater good lies—but aren't they at least sometimes really just *stronger*, even if we are aware that they might (or will) go against our greater good? The reader might even admit that sometimes strong passions are the sources of our best reasons, but it also seems clear that sometimes this is not the case. Further, it appears that the strong passions don't always or even generally require self-deception or distraction to move us. Can they (or something else) not move us contrary to what we take our greater good to be? And when they do so, is this not irrational? In any case, I already admitted that I won't rule out a clear-eyed favoring of one's lesser good over one's greater good. Such cases are intuitively irrational, so I had better have a story that accommodates or explains away this intuition.

3.2.3.1 End-Relationalism

I see no good reason not to describe such behavior as irrational, but this is not because the relevant desires are *intrinsically* irrational. We can and should yet maintain that no desire can be rationally faulted in and of itself. These desires can however be faulted *relative to other desires*. For those who are accustomed to search for non-relational normativity, this will seem to miss the point, since 'irrationality' entails (rational) *failing*, not 'merely' conflict. We do not experience our desires as *merely* in conflict with one another, but, in Brink's terminology, some of them seem to have

as well, which does not imply any particular phenomenology, but I won't pursue that here. The term is from McDowell (1981, 155).

(rational) *authority* over others. How can a simple desire theorist capture our sense of the *normativity* that prudence has over desires for our lesser good (if in fact such desires exist)?

I'll begin to answer this question by quoting Hume:

> What we commonly understand by *passion* is a violent and sensible emotion of mind… By *reason* we mean affections of the very same kind with the former; but such as operate more calmly, and cause no disorder in the temper… (T437).
>
> When any of these passions are calm, and cause no disorder in the soul, they are very readily taken for the determinations of reason … What we call strength of mind, implies the prevalence of the calm passions above the violent" (T417-8).[203]

Insofar as the 'calm', relatively stable desires hold sway, we act prudently, and not otherwise.[204] Korsgaard recapitulates the same understanding of Hume, following it with the interpretive claim that if this 'general desire for the good does not remain predominant, not only the motive, but the reason, for doing what will conduce to one's greater good, disappears' (379). We've already seen that it is highly misleading to describe typical cases of acting for one's lesser good as cases in which one has 'lost the motive' for one's greater good generally.[205] But all this aside, we can answer the question why imprudent actions are irrational, where that entails that they are wrong, by recognizing that these judgments are (typically at least) made *from the point of view* of the longer-term desires and commitments. That is, relative to the ends that they

[203] These quotations appear in Finlay (2008, p. 25).
[204] This temporary simplification is deliberate.
[205] Further, lack of 'predominance' does not entail the *absence* of motive, but a *relative* weakening.

represent, which ends we tend to identify with as (rational) agents who *will* those ends, actions that run contrary to them are 'irrational'.

Although most of our long-term, relatively stable preferences have a 'calmer' quality than many of the desires that seem to assail us from time to time and under particular circumstances, this isn't always the case. Our focus should be on the relative stabilities of the desires, not their 'violence' or 'calmness'. Our relatively stable desires are generally for long(er)-term goods[206], and further, it is generally in pursuit of these goods that we experience ourselves as willing, and our experience of ourselves as agents is bound up with our experience of ourselves as willing. Let me elaborate on these claims.

Agency, Rationality and the Will

We have and form commitments, intentions and resolutions. These, by their nature, attempt to control what we do in the future. As such, they are at least partly constitutive of what it is to be an agent. 'The view of [agency] as active *now* essentially involves a projection of itself into other possible occasions' (Korsgaard, 1996, p. 229). Insofar as we are agents we are not carvable into indefinitely many time-slices. Korsgaard also views the will as a lawgiver whose laws are to govern oneself at other times (or one's future selves).[207]

[206] I am restricting the desires under consideration to be those connected with our own good. I do not mean to suggest that our stable desires are for our own (long-term) good as opposed to others' good.

[207] It's worth noting that Korsgaard's conception of the will, aside from its rationalism, is very amenable to Ainslie's.

> When I will an end, I must *ipso facto* will that even on another occasion, even when I am tempted not to, I will stay on the track of that end … So when you will an end, the form of the act of your will is general: you will a kind of law for yourself, a law that applies not only now, but on other possible occasions. (Korsgaard 1996, pp. 230-1)

Our sense of ourselves as agents is essentially bound up with our sense that we can will ends (in the future). Our sense of ourselves as agents is also bound up with our sense of rationality. On a Kantian/Korsgaardian view, our fundamental essense just is rational agency. Even if we don't take quite such a rationalistic view of ourselves, most of us, especially those who are inclined to resist the kind of view I'm espousing, think of their agency and their rationality as essentially connected, if not different words for the same thing.

The view I'm advocating also recognizes important connections between agency, rationality and the will. But it explains them differently. This view is that we identify ourselves as agents primarily with our long(er)-term, relatively stable desires, which are the desires that often require the phenomenon of willing to procure, given the otherwise hyperbolic discounting that would consistently undermine them. The sources of imprudence, understood as failure to realize our long-term good, often involve factual and reasoning errors, as well as distraction by a thousand means from what is conducive to our greater good. Therefore, reflection on our ends, removed from sources of distraction and with time to correct for factual and reasoning errors, is an excellent way to promote our greater good, in general. We call this slow, conscious reasoning toward a judgment about what to do using our (practical) rationality, and acting in arrordance with these judgments is (at least pro tanto) to act rationally.

But on this view there are no intrinsically irrational ends, and there is no agent or executive that stands apart from all our desiderative perspectives and pronounces on the rationality of some and the irrationality of others. Rather, such pronouncements come from the perspective of some end(s) or other(s), and the perspective associated with our wills and long(er)-term goals, with the deliberative, conscious reasoning about what to do that generally occurs from that perspective, is the perspective of what we call 'rationality'. Practical judgments are inherently made from a motivated perspective. Borrowing from Finlay (2008, 26), ought-statements are teleological, *end-relational* statements.[208]

My goal will of course be to argue that my end-relationalist view is more plausible than competing rationalist views. If it is at least as plausible, then we will have dealt with the ostensibly powerful objections that such a view denies both weakness of will and practical reason itself. But how are we to judge which of these two kinds of view is more plausible? The right place to start is the will.

3.2.3.2 Korsgaard on the Will

Korsgaard understands the will as a lawgiver, the essence of which is to govern choice at other times. She regards the source of normativity as residing in the nature of the will, in 'the fact that we *command ourselves* to do what we find it would be a good

[208] I should repeat that Finlay means this as a semantic thesis, whereas I don't (I'm uncommitted and mostly uninterested in the semantic question—see Chapter 5). However, Finlay notes that the semantic thesis is in Foot (1972) and Mackie (1977), and he argues for such a view in his 2008. I only endorse the view that normativity is *in fact* end-relational, a view for which Finlay also argues in his 2004 and 2006.

idea to do' (1996, 105, her emphasis).[209] And we *must* command ourselves to act: '[t]he reflective structure of human consciousness requires that you identify yourself with some law or principle that will govern your choices. It requires you to be a law to yourself. And that is the source of normativity' (104-5). Such commandments necessarily take the form of a universal law: 'The claim to generality, to universality, is essential to an act's being an act of the will' (1996, 232). Further, the fundamental threat to these laws and their governance of us is when we make 'an exception of the moment or the case' (103). 'In cases where a small violation combines with a large temptation, this has a destabilizing effect … You may know that if you always did this sort of thing your identity would disintegrate … but you know you can do it just this once without any such result' (102).

The point of all this is to illustrate how the core elements of Korsgaard's conception of the will (and normativity) are paralleled in Ainslie's model. But on the end-relational view, the 'sources' of normativity are one's ends. More perspicuously, normativity just is end-relational. My view sides against Korsgaard and (roughly) with Hobbes in thinking that 'the role of the legislator is to make what is *in any case* a good idea into *law*' (1996, 24, her emphasis). That is, from the perspective of one's long(er) term ends, it is good that one study, brush one's teeth, etc., but without the power to 'impose sanctions',[210] the 'subjects' (our competing interests) will not obey.

[209] Until further notice, all quotations will be from her (1996) *Sources of Normativity.*
[210] That is, to offer greater reward *at the time of action* than the short-term desires can, which often involves present unpleasant feelings, such as anticipatory guilt and the like.

3.2.3.4 End-relationalism vs. Korsgaard on Practical Rationality and Will

The point of these comments has been to establish the central elements of Korsgaard's thinking about the will, rationality and normativiy, and the prima facie similarities and differences between her and my conceptions thereof. In this section I'll show how my end-relational conception compares equally well or favorably to hers along several dimensions.

Much less potential for pathology in prudential vs. religio-moral willing

Both Korsgaard and Ainslie think of the will as legislative, of laying down law-like principles for future versions of ourselves to obey. What sorts of things do we tend to command? We do not experience ourselves as commanding our future selves to have a large bowl of ice cream[211] or smoke a cigarette or stay out late drinking or cheat on our spouses. That is, we tend to command our perceived long(er)-term goods (not necessarily conceived as such) over our shorter (in the prudential realm at least). Of course Korsgaard would say that we do this '*because* prudence has rational authority', whatever our desires or commitments might be, and our wills are responsive to this fact (1986, p. 380, my emphasis). But this can't be quite convincing by itself. People command themselves to do things that others, including themselves at different times, find outlandishly irrational. People resolve themselves ever so firmly and piously to such

[211] Except perhaps in cases of compulsive eating disorders, which are in competition with both shorter- and longer-term rewards. Here, the command also comes from the longer-term preference of eating healthily.

a bewildering and, to the outsider at least, utterly insane variety of religious and moral proscriptions that it can hardly be maintained that the apparent rationality of the commanding will in cases of prudence is the result of the fact that this will is inherently, or even reliably, 'rational'.[212]

There seems to be a much greater range for 'irrationality', or pathology of the will outside of what are taken to be (at least purely) self-regarding contexts. This is not to say that such pathologies don't exist in circumstances where one's own good is the sole or overwhelming consideration, but it is to say that they seem much less rampant, and the variability in what strikes observers as irrational seems considerably less than in religio-moral contexts. It would take more space to argue for this than I want to spend here, so if the reader does not accept the premise, this point will have no effect on him. However, if he does agree to it, then on the hypothesis I'm offering, this difference is explained (at least in part) by the pathologies of the will that are made possible by the fact that the principles and evaluative beliefs employed by the will deflect attention away from one's own concerns. But insofar as we remain within what we take to be a purely self-regarding context, there is much less potential for pathology of this kind. The potential for large discrepancies between one's underived desires and one's conception of her own good is much less than that between one's underived desires and one's conception of religio-moral Good.

I am not sure what sort of explanation(s) Korsgaard or other rationalists would offer for this phenomenon, insofar as they agree that it exists. But if we do agree that

[212] Let this be clearly distinguished from the the claim that the will *takes itself to be* rational, or normative. I think this is generally, if not always true. At any rate, this could be true and it would not help the objector here.

there is something to be explained here, my account has an explanation, while Korsgaard would seem to have trouble with this apparently true generalization.

Long-term-goal explanations of paradigmatically irrational behavior

Now let's look at some mechanisms underlying many of the paradigms of irrational behavior. Foremost among these are rationalization, wishful thinking and self-deception. We take the last to be involved in the first two, at least most of the time. Rationalization is conceived of as 'fooling ourselves' into thinking that some desire of ours is reasonable when it really isn't. When we judge ourselves to have rationalized some behavior, it can seem as if our reasoning, justifying capacities have been hijacked in the service of desires which are not in fact, but wish to present themselves, as rational. These desires are very often for shorter-term rewards than the rewards associated with the judgments they (if only temporarily) displace. Wishful thinking is intrinsically and immediately rewarding in most cases. Believing that I am likely to win the lottery is quite a pleasant thought. That my spouse is not cheating on me is a (relatively) pleasant thought. But if it seems that I do or should have good reasons for not believing these things, then others (perhaps including my later self) will judge that I am engaging in wishful thinking, deceiving myself.

When we ask what is *wrong* with wishful thinking, rationalization or self-deception, one kind of answer we might get is 'those things are irrational'. But what is the cash value of such a claim?[213] It seems clear that *in general*, at least one thing that is wrong with these things is that they trade long-term for short-term satisfactions. Facing the truth about an unfaithful spouse or about the real possibility that one has cancer[214] is unpleasant in the short-run, but *tend* to be better in the long-run than not doing so. When we judge that we have rationalized ourselves into having a cigarette or a drink, the question before us is the following. Is it a faculty of pure practical reason, in the service of no ends or set thereof, that is seemingly hijacked? When we judge that we have fooled ourselves, or that others are fooling themselves, is it the case that in each person there is a rational self which stands apart from all her desires and judges of their rational authority, but that rational self (somehow) deceives itself, or is (somehow) deceived by some desire as to whether that desire has authority? Or do we have ends at different time-ranges (as well as different rankings at the same time-ranges), most of which have the capacity to influence our judgment in their service?

It seems at least as plausible that we have long-term goals, which often result in intentions and commitments of various kinds and degrees, which by their nature attempt to influence or control our actions in the future. And these long-term preferences, commitments and so forth are the ones we endorse on (not necessarily calm) reflection, and are (generally) the ones we identify with as rational persons, agents or ‘selves’.[215]

[213] One might critique wishful thinking for example from the standpoint of theoretical rationality, but such a critique would only show why this belief-formation process was unreliable, non-truth-tracking, or employed invalid inferences. But the question we’re wondering is what is the *problem* with doing those things in the practical realm? In other words, from the practical point of view, we are asking in response to these ‘critiques’, ‘So what?’

[214] Jon Elster’s (2007) examples.

[215] I realize that the commitments that we identify with go beyond simple stability of preference, but now is not the time to enter into a discussion of which commitments, in general, constitute our sense of who we are, at least qua rational persons. I doubt that any such general characterization is possible. For the time

Violations never rational vs. violations sometimes rational

According to Korsgaard's conception of the will and its relationship to rationality and normativity, one's own commandments to oneself are the source of normativity. On this conception, those commandments have the form of a universal law. Since these commandments issue from one's own practical identity, violation of them is unconditionally irrational, unless this identity and its commandments conflict with a more fundamental source of one's practical identity.[216] But as we've seen, she recognizes that one or a few violations can be tolerated without loss of that identity, but not indefinitely many. Korsgaard says that this shows that the 'depth' of obligation is limited, though it is always unconditional.

I want to make a comparison here between Korsgaard's and Ainslie's understandings of the relationship between rationality and the will as lawgiver. On Korsgaard's view, it is always and unconditionally irrational to violate the laws that one's will has laid down. On Ainslie's view, things are quite a bit more complicated and elusive:

> [T]here can be no hard and fast principle that people should follow to maximize their prospective reward. Thus "rationality" becomes an elusive concept. Insofar as it depends on personal rules demanding consistent valuation, rationality means being systematic, though only up to the point

being, it suffices to note that we tend to identify with our long-term commitments over our short-term desires, and generally experience ourselves as willing the former and not the latter.

[216] Of course I haven't allowed the reader to make full sense of the 'since' here, since I haven't explicated her views on why this follows, but the essential point is that acting in contradiction to one's practical identity threatens to destroy that identity.

> where the system seems to go too far and we look compulsive. Even short of frank compulsiveness, the systemization that lets rules recruit motivation most effectively may undermine our longest-range interests. (2005, 645)

Korsgaard views violations as unconditionally wrong (irrational), though not necessarily 'deeply' so. '[W]here a small violation combines with a large temptation, this has a destabilizing effect on the obligation' (102). Since our identities are fairly stable, such violations, though necessarily wrong, may not be 'deeply' wrong, in that small, infrequent violations needn't undermine the aspect of our identity associated with the relevant obligation. However, for Ainslie, matters are not so clear-cut. Whether or not some violation is 'irrational' can depend very heavily on various and subtle features of the agent and her history and how she conceives of her own obligations. In many cases, it just won't be decidable whether it's 'irrational' or not. For example, most of us have as part of our practical identities a commitment not to steal, but I hope the reader joins me in finding it insane to suppose that it would therefore *always, necessarily* be wrong (irrational) to steal.

This might seem to ignore the fact that Korsgaard allows for the rationality of violations in cases where one's more fundamental practical identity is at stake. But on Korsgaard's view, the practical identity associated with a commandment not to steal is the most fundamental. It is the identity one has as a rational agent per se, which identity requires you to value unconditionally the rational agency of others, which requires you not to steal from them, for that is to undermine that agency. Perhaps one can get around this somehow. Perhaps one can exempt stealing from the list of things that is forbidden by one's most fundamental identity as a rational agent.[217]

I don't think such a strategy is promising, for two reasons. First, it seems unlikely that one could keep in enough obligations for it to come close to capturing 'our' pre-theoretic judgments and yet not be open to indefinitely many counterexamples. It just seems crazy that we shouldn't violate one of our own commandments where the violation is small enough and the benefits high enough, even in the 'moral' realm. I expect that attempts to accommodate these intuitions by reference to rankings of practical identities (that are consistent over time!) without resulting in triviality will be unsuccessful.

But, and this is the second reason, there is no need to stick to the moral realm. Remember, an obligation arises from the fact of having commanded oneself; we shouldn't think that they involve some independent obligation that we 'really' have. We can command ourselves to run every day, or once every week without fail, or write at least 2 hours every day, or a thousand other things, and it is just absurd to suppose that it is irrational to violate such commandments no matter how large the 'temptation.'

Not writing for a day in order to take some fabulous opportunity that is inconsistent with doing so is not irrational. Of course the first reaction to such criticism is to try to build in exceptions to the rules, but attempts to do this that both avoid counterintuitiveness and yield something better than 'except when it's (certainly? almost certainly? probably?) a good idea not to' seem doomed.

Life, and rationality, require *judgment* (quite a bit of which is preconscious and/or intuitive). And yet rules and commitments seem to have a very important role to play.

[217] I don't think that the strategy of finding a different fundamental identity than 'rational agent' will be at all promising insofar as one wants to keep the basic Korsgaardian/Kantian framework and (contentful) unconditional obligations. The same kinds of counterexamples will be easily discoverable for any putative fundamental identity.

Ainslie’s model accounts for the importance of these things without committing itself to seemingly absurd implications.

One might think that all this is talking past Korsgaard since it goes without saying that Korsgaard wouldn’t agree to the premise that rationality is in the business of maximizing reward (neither do I). But this response would miss the point. The point I want to make is that Ainslie’s model *explains* why Korsgaard’s intuitions (if we call them that) are roughly what they are, and why she doesn’t have to spend any time arguing for claims to the effect that our ‘practical identities’ can take a few knocks without collapsing, but will do so if we keep violating the ‘commitment to our own integrity’ (103). These and the other features of the will she recognizes, such as its inherent ‘generality’ and intimate connection with principles of behavior, are fairly intuitive. And these features, among others, have a compelling and unifying explanation in Ainslie.[218]

There is another point to be made here. Ainslie’s model, at least insofar as it captures these features of the will which Korsgaard recognizes, should be attractive to her and those who also recognize these features. It is of course a vastly more empirical, unifying, sophisticated and richly elaborated model of the will than Korsgaard has to offer, capturing vastly more ‘data’, in addition to giving more intuitive answers to what counts as ‘irrational’, at least in some cases. And this model has no executive rational agent residing beyond all desire or reward, who decides which actions get taken and which don’t on the basis of their perceived rationality. The model is inherently end-relational. And further, there is no reason to think it has been constructed with a mind

[218] It’s worth reminding ourselves that the model, in a non-ad hoc way, entails that the reward system is set up in such a way so as to highly discourage conceiving of at least many kinds of rewards (especially emotional ones) as the object of our pursuits.

toward supporting any particular theory of rationality or morality. Insofar as we think of Korsgaard as having an independent desire to defend a Kantian conception of morality, normativity and obligation, we can understand why she would be motivated to take violations of the will's commandments to be unconditionally irrational.[219] For the account to work, it *has* to be that way. But, when we have no such prior view to support, we can even more easily see that it just isn't that way.

I said that the model has no extra-desiderative rational agent judging of the goodness or rationality of actions and desires. It does however predict and explain why long-term motivations can benefit from not being represented as 'merely' such, but rather objectively and/or intrinsically good or rational. Korsgaard recognizes the basic idea behind this feature as well. She criticizes Hume because he 'forgot that knowing that our hatred of injustice was based on general rules would have a destabilizing effect on the obligation always to be just' (p. 103, n. 12). I don't know whether Hume forgot this or not, but I do know that it is true. This *is* destabilizing. To the extent that a person is committed to *always* be just, there will be pressure against coming to believe any such proposition about the basis of justice. One intuits its seditiousness. Likewise, to the extent that a person is committed to always act on prudential principles, there will be pressure against coming to believe that such principles are explicable as commitments based on general rules.[220]

[219] And of course that there is a rational agent which transcends all desire.

[220] Later I'll try to make up for these destabilizing effects.

Summary so far

Here are the beginnings of a defense of end-relational practical rationality. Note that it does not entail a mindless relativism. There are still numberless facts of the matter as how to best[221] achieve goals. Very many derived desires are presumably based on false beliefs and/or bad inferences. There is abundant room for criticism, not ‘simply’ from one perspective against another, but of the sorts we canvassed above. We can criticize another culture’s footbinding or ‘genital-mutilation’ practices, not only from our perspective, but from the perspective of their own concerns.[222] Others (and we) can criticize our materialism, commercialism and wastefulness not only from their perspective, but from the perspective(s) of what we care about.

The presumption in favor of an objective standpoint from which to judge normativity can be explained on the basis of the very robust forms of fallibility we’ve discussed, in addition to the independently-motivated account of the will according to which motivational resources can be marshalled more effectively by diverting attention away from one’s own introspected desires. We canvassed other explanations as well in Chapter 2. To the extent that these or some other nonrationalist explanations are plausible, such a presumption carries little to no weight against an end-relational view with multiple independent attractions.

My Hume- and Ainslie-inspired end-relational view of practical rationality seems more plausible and certainly more empirically grounded than one of if not the most

[221] ‘Best’ will of course be cashed out in terms of other desiderative considerations.

[222] Gerry Mackie (1996) has an excellent account of not only how footbinding was ended in China in a single generation after existing for 1,000 years, but of how the motivation to end it is very easily explained (and justified) from within their own desiderative set.

influential rationalist conception of practical reason. Probably the biggest apparent cost of my view is that it seems to have no room for any fallibility that is not in relation to some desire(s). But the model of the will upon which my view draws provides a non-ad hoc explanation of why this result would be so counterintuitive.

Now, having satisfied myself that I've accommodated our intuitions regarding the possibility for robust fallibility, including the paradigmatic cases of practical irrationality, I continue my responses to Brink's arguments.

3.3 Categorical Requirements

3.3.1 Temporal Neutrality

I can empathize with the feeling that there is something objectively rational about prudence that goes beyond 'mere' difference of perspective between our long- and short-term interests. One problem of course is that we are accustomed to say 'mere' before 'difference in perspective'. It is going to be a challenge for us to come to appreciate that whatever significance and importance there is in the world is from some motivated perspective(s).[223] Another source of this feeling is very plausibly the idea that to be prudent is objectively better for us than not to be. This is very plausibly true. Insofar as we remain the same person over the course of a life, to the extent that we achieve our long-term goals, our lives generally go better. To the extent that short-term desires

[223] Of course, many philosophers have advocated such a view, probably most famously and influentially, Nietzsche.

undermine this, our lives go worse, generally speaking. Our lives seem to go best of all when we are able to find a judicious balance between indulging short and medium-term desires of various kinds, while not allowing them to significantly threaten or undermine our life-plans.[224]

These comments stop short of accommodating Brink's contention that temporal neutrality is a categorical rational requirement. Instead, they point toward quite general, even abstract desires. Most of us have desires that our lives go as well as possible, that we're as happy as possible, that we have as few regrets as possible, and so on. We can take up the motivated perspective that encompasses our whole lives, conceived as such.

I said that these desires are abstract. They are more or less abstract depending on the specificity and concreteness of one's ideas about what will be likely to make and/or constitute one's life going well/best. Nevertheless, the consideration that course of action A will result in our lives going better than course of action B, by our own lights, is a motivating consideration to most of us. That's not to say it isn't effectively outcompeted by other motivations, only that that most-encompassing of perspectives is a motivated perspective, representing concerns that, when we are reflective and calm, we tend to regard as more rational than conflicting desires and their associated courses of action.

We can see the rewards of prudence in our own lives and in those of others. We can see the suffering that imprudence brings to us and to others. We want the rewards of prudence and we do not want to suffer the harsh consequences of imprudence. This does not mean that perfect temporal neutrality is rationally required, much less unconditionally

[224] Of course, sometimes our life-plans can be changed in the course of pursuing what seemed shorter-term goals. Whether this counts as undermining will be a matter for interpretation by others, especially the later self.

rationally required. From the perspective that encompasses our whole lives as such, it is rational to be perfectly temporally neutral, as far as this is possible. That is to say that in order to make our lives go as well as possible on the whole, it is necessary to be temporally neutral insofar as this is possible.[225] So, *relative to that end*, we are required to be temporally neutral.

This is all consistent with the view I'm proposing. In fact, I think it is plausible that rational prudence requires temporal neutrality even relative to ends other than living the best overall life we can. This is because it is far from clear that personal identity is what grounds what Jeff McMahan (2002) calls 'egoistic concern'.[226] He agrees with Parfit (1984) that identity is 'not what matters' for rational prudence. What McMahon calls the 'prudential unity relations' are defined as whatever grounds rational egoistic concern, and McMahon argues that these relations are not constituted by personal identity but rather psychological unity (75). Psychological unity encompasses both psychological connectedness and psychological continuity. "The degree of psychological unity within a life between times t1 and t2 is a function of the proportion of the mental life that is sustained over that period, the richness or density of that mental life, and the degree of internal reference among the various earlier and later mental states" (75).

The details of McMahon's arguments in favor of psychological unity rather than identity as grounding rational egoistic concern, and indeed the details of what psychological unity amounts to in the first place, would take us too far afield. The important point is that it might be that rational egoistic concern comes apart from our

[225] Where temporal neutrality is thought of as a standard, not a decision procedure.

[226] Like McMahon, I agree that it could be less misleading to call this 'special, egoistic-like concern' but that phrasing is awkward and I will also use egoistic concern as a shorthand for 'special, egoistic-like concern' (2002, 42).

personal identities. In fact I agree with McMahon's view that psychological unity is what matters for rational prudence, but I cannot recapitulate his extensive arguments for that conclusion here.

I do want to point out one feature of his view however, and that is that McMahon thinks that rational egoistic concern has a discount rate built into it; the discount rate is not temporal however, but tied to the strength of the prudential unity relations. Where the prudential unity relations are maximally strong, there should rationally be no discounting, but where they are weak then significant discouting can be rational, even rationally required. Still, the defender of temporal neutrality can rightly claim that even on this view, there is no rational defense of discounting sheerly on the basis of time, but only on the basis of psychological unity. Yet we clearly have desires, which we sometimes act on, that do not conform to the kind of discounting that McMahon argues is rationally required by prudence. In other words, we often display discounting that is explained by the relative temporal (or physical) location of rewards, not by the degree of psychological unity between various (versions of our) selves. It seems that such discounting is irrational on McMahon's view as well, no matter one's desire set. How can my view accommodate this?

In the same way as before, but now with more details. I said above that we had desires directed toward our lives as a whole, and that we plausibly make judgments of irrationality from the perspective of these desires. But I think that what arguments by people like McMahon and Parfit help to show is that identity is not what actually matters *to us*, even when it comes to the special kind of concern we have for how our own lives go. If, as McMahon (2002) argues (and I agree), we are essentially embodied psychological beings, with a certain kind of special concern for ourselves (egoistic concern), it should not be surprising that as the degree of psychological unity lessens between ourselves now and our (imagined) selves in the future, that kind of concern is weakened.

Parfit says, 'My concern for my future may correspond to the degree of [psychological] connectedness between me now and myself in the future. Connectedness is one of the two relations that give me reasons to be specially concerned about my own future. It can be rational to care less, when one of the grounds for caring will hold to a lesser degree' (1984, p. 313). While Parfit and McMahon presuppose that there are objective grounds that provide reasons to be concerned about one's own future, their methods for finding these grounds crucially depend on 'intuitions', which I regard as fairly clearly tapping into what we actually care about. Discovering that connectedness is one of the two relations that give me reasons for egoistic concern should be understood as discovering that this special kind of concern *actually tracks* psychological connectedness. Perhaps it also tracks psychological continuity, and what McMahon refers to as psychological unity.

So when McMahon or Parfit argue that such and such prudential unity relations ground rational egoistic concern, to the extent that their arguments are persuasive they will have given grounds for believing that we have a special kind of (egoistic) concern that *actually tracks* these relations.[227] They will effectively be arguing that the kind of concern we are all familiar with is really a concern about or based on X (psychological

[227] If it did not track them, the intuitions upon which they rely to build their theories would presumably be different.

unity) and not Y (identity). Once we have shown that egoistic concern is in fact 'grounded in' X, we can then say that motivations which can *appear* egoistic but in fact, say, discount temporally in a way that is inconsistent with X, are irrational (or acting on them is irrational). And in a very important sense—which is very easily confused for something it is not—such behaviors and such discounting can be described as *categorically irrational*. Recall that I am saying that these judgments are made from a motivated perspective. In some cases, the perspective is that of X, i.e., whatever it is that egoistic concern is actually concerned with.[228] Now, *from this perspective*, it seems *categorically irrational* not to be temporally neutral. That is, no matter what *other* desires one has, temporal neutrality is required. But since the motivated perspective from which judgments of rationality are made does not take itself to be a motivated perspective, it understands that to mean that the rationality judgment is independent of *all* motivated perspectives.

Practically speaking, there is nothing necessarily wrong with this. This is the perspective from which we *should* make such judgments—just insofar as we want to further our egoistic concerns. Further, the methodology of plumbing into this egoistic source of concern with thought experiments designed to reveal whether and how we care about an innumerable variety of counterfactuals is a fantastic way to discover the nature of this very source of concern, when practiced by a talented and careful philosopher-investigator. However, the instinctual and unquestioned assumption underlying these otherwise valuable investigations is that one is discovering what 'really matters', where

[228] McMahon's and Parfit's points, recall, is that it might not be fundamentally concerned with how our lives go as such, but with something easily mistaken for that, namely considerations of psychological unity.

that is understood as independent of what people actually care about.

But why think that my analysis is correct? Why not think that these investigations serve to uncover what constitutes what is *objectively* rational egoistic concern, no matter what anyone's *actual* concerns? The most obvious answer is that it is far from clear how the 'intuitions'[229] garnered by these investigations would or could be derived from something independent of the very source of concern that I argue is (whether the investigators know it or not) under investigation. Of course this is just a special case of the larger issue I am addressing in this essay, which has to do with the best explanation of our beliefs about nonrelational values. Therefore I am not going to get into the details of this more specific debate about how such intuitions could be about things other than people's actual concerns. But I submit that 1) the primary explanation of why people think they are investigating a realm of motivation-independent facts is that it just seems that way, 2) I have done a lot of work already in trying to explain why it would seem that way,[230] and 3) in the face of such explanations, my view is the most straightforward explanation of how such a methodology can serve a valuable purpose, which I think it does. While it is far from clear how we can intuit facts about what is valuable or reasonable in the practical realm that don't depend on what we actually care about, it is not surprising that we can conduct an investigation into what we actually care about.[231]

In sum, what I am arguing is that *whatever* the prudential unity relations are, judgments of prudential irrationality can be understood as judged from the perspective of

[229] A term which prejudices the mind toward thinking that they are about something mind-independent.

[230] This explanation doesn't just rely on Ainsliean psychology, but also strong traditions and inertia in academic philosophy.

[231] Though perhaps to some it is surprising that we should *need* to. This topic will be discussed in Chapter 6.

the special concern that we call 'egoistic', but the specific nature of which is a matter of substantive disagreement. Whether this kind of concern is understood as essentially about our personal identities (which last our whole lives), or whether it is essentially concerned with degrees of 'psychological unity', that is what it is concerned with. And therefore judgments of egoistic rationality will be made from the perspective of that concern. And from that perspective, discounting on the sheer basis of time or physical proximity will very likely be irrational.[232]

Therefore no matter what the prudential unity relations are, we can accept that the end-relational view can accommodate the possibility that temporal neutrality is a rational requirement. In fact, we can even accommodate the claim that it is a categorical rational requirement, but we will have to understand that nonrationalistically. We will have to understand that to mean that no matter what *other* motivated perspectives (desires) one has, acting in accordance with them will be irrational *from the point of view of one's egoistic concern*.

I submit that there is no rationalist account that handles the ostensible categorical rational requirement of temporal neutrality better than the view I have just outlined. My view can capture the relevant intuitions just as well as they can, and further, those intuitions themselves are totally unmysterious on my view, whereas it is far from clear how we could have reliable intuitions about what is categorically rationally required, where that means independent of anything we actually care about.[233]

[232] This will be true just in case temporal and physical proximity are not themselves among the things tracked by egoisic concern, which I think they are highly unlikely to be.

[233] It makes a very big difference that the rationality judgments at issue here are those of practical rationality, i.e., about what to *do*. Doing things, performing actions, requires motivation whereas believing something does not (not that motivation isn't often involved, it's just not necessary). Therefore I would not

The claim that my view can capture the intuitions just as well is stronger if we keep in mind that intuitions to the contrary can be explained in terms of what would be best overall for a person, as opposed to what it is rational for them to do. If they do not have the kind of egoistic concern that (most of) the rest of us do, then it might be that the only sense in which it would be 'rational' for them to be temporally neutral is that in making such a judgment, we take the perspective of their long(est)-term good. There will be more on this topic when we get to the question of specific problems of normative accommodation (such as what is rational for depressives to do) in section 3.3.3.[234]

3.3.2 Categorical Requirements of Morality

It is one of the central purposes of this essay to explain and undermine the belief in categorical imperatives, particularly with a mind toward changing the way we think and talk about 'morality'. So the general claim that we have categorical moral requirements is simply to be weighed against whatever I am able to do here (and other people have done elsewhere) to undermine such a claim.

In this context, I think that in order to generate a convincing argument to the effect that there is a problem of normative accommodation with respect to moral requirements, one would would want a *specific case* in which we believed strongly that we had reason to do something (moral or otherwise), but had no goal which involved

argue that it is equally strange to think that we could have reliable intuitions about what it is rational to *believe* that are independent of our motivations.

[234] Finally, it is worth noting that it is very far from a consensus view that temporal neutrality is required in the first place. Most people think some fairly mild form of exponential discounting is the rational ideal. If something like psychological unity is the ground of egoistic concern, then this might make sense on the assumption that time is a rough proxy for degree of psychological unity.

doing it. This is best conducted in the first-person or with someone we know well, since it is easier to say what their goals are and whether they have reason for doing things. This is presumably where we get our intuitions about novel cases (largely from imagining doing something and feeling a response). Absent such a case, it doesn't seem convincing (except perhaps to the choir) to insist that everyone has desire-independent reasons to be moral.[235]

3.3.3 Specific Problems of Accommodation

Depressives

The above discussions of categorical temporal neutrality and moral requirements might have left some readers dissatisfied. These are quite general, abstractly theoretical objections and responses. I said that I would like to see a case in which someone clearly has a moral requirement but no corresponding desire in order to be able to take that objection on. While Brink does not provide any such specific cases with respect to categorical moral requirements, he does give specific (hypothetical) cases in which it seems that people have reasons to behave in ways even without corresponding desires, and reasons not to behave in certain ways no matter what their desires.

These specific cases, wherein depression or other systematic neurological dysfunction inhibits motivation for both moral and prudential actions, might seem more

[235] The denial of the view that we all have desire-independent reasons to be moral is called 'anti-rationalism' about morality and will be discussed in subsequent chapters.

damaging to my view.[236] People in these circumstances ostensibly have reasons for action (moral or otherwise), though they don't (appear to) have desires to that effect. Since many of us have experienced being depressed or know people who have been, we can have a clearer picture of a case in which it seems as if people have good reasons for action but do not have accompanying desires. However, if we think of desires as broadly as goals, then it's not clear what the upshot of depressives and the like is. There seem to be two broad possibilities: either they have some goals or they don't, and either they are capable of generating some motivation or not. I'll focus on prudential goals, though the same arguments would presumably apply to moral ones (where these are not considered to have practical clout).

If they are *not capable* of generating motivation to act on prudential concerns, then it seems idle at best to say that they have a reason to do so, for having a reason implies that one can act on it. In Scanlon's (1998) terminology, such a depressive is not 'judgment-sensitive,' i.e., even if they judge that they should do something, it has no effect on their attitude or motivation, and to that extent they are not profitably considered (practically) rational creatures. Their inability to be motivated to act renders the ascription to them of a reason for action as idle as the ascription to me of a reason to fly to school every day, on grounds that it would be more fun than taking the bus. Of course, in general we don't know whether a person is truly incapable of being motivated by reasons. When we give normative advice, it is under the assumption that it could be taken. We don't say of someone in a coma that they have reason to rise. That would be

[236] Such cases form the heart of Smith's (1994) argument against Humean instrumentalism, so my arguments here may be considered a rebuttal of his primary considerations against instrumentalism.

ridiculous. When we do say of a depressive that they have reason to rise, it is because although we regard them as motivationally impaired in some way, we do not suppose that they are incapable of generating the requisite motivation. When we do suppose they are incapable, we do not ascribe to them a reason to do so.

Now, if they *are* capable of generating the motivation to act in accordance with prudential or moral considerations but nevertheless have neither moral nor prudential (or any other) goals, then they are a very strange bird indeed, for it's hard to know how to imagine a person with no goals at all of any kind (or to even imagine them as a person). They will soon die if not kept alive artificially, and once again I find it very odd to speak of what sort of reasons such a creature has. Reason-talk finds no home in such creatures; at least I find it as comfortable to say that that such a thing has no reasons as to say that it has some (I'm not even very comfortable, though a bit more so, speaking of its good).[237] And, if someone really and truly has no desires involving what we term the moral realm, then it seems question-begging to insist that they yet have a reason. Such an insistence is itself plausibly construed as a desire to give morality a boost (for which one might well have good reason!). I suspect that in most cases, it will be sensible to ascribe to a person the goal of getting out of the depression. In cases of severe and systematic neurological dysfunction, it might not be, but I think in just such extreme cases we are or should be reluctant to pronounce confidently on what it would be practically rational for them to do,

[237] I think Smith, though perhaps not Brink, moves away from his own conception of desires as goals in the case of depressives. A desire on the direction-of-fit model doens't imply a 'felt' desire, or a 'passion' to get up. That is a problem with depressives, they have had their motivational resources depleted. But that, even in severe depression, does not imply that they have no goal or desire *at all*. A test would be if you could give them the Staples 'Easy Button' to bring themselves out of it. Pushing it would be evidence that they have the desire, but the depression took away their ability to organize *sufficient* motivation in some cases, not their desires entirely.

not least because they might be incapable of it.

The lint-collector

It seems that the life of a full-time lint-collector is bad, even if the lint-collector is successful in her own terms and collecting lint is her deepest desire. We do not find *Brave New World* dystopic because the Epsilons' and Deltas' desires go unsatisfied; on the contrary, it is their desires themselves that are contemptible. That is exactly correct in my view; their desires are contemptible because we do not want to (we want not to) live in a world where our fellows' desires were like that and we wouldn't want our own desires to be like that. The emotion of contempt (or disgust) helps to 'push away' such debasements. So the first thing to note is that we are not like them; we are most certainly not people whose deepest (or most committed) desires *are* lint-collecting or anything remotely close to it. We would have to become quite different people in order for that to be the case.

This brings up the question of why we should not become such people. Of course, in the real world, we do not have that option. I can't decide to become someone whose deepest desire is counting grass even if I were to believe that would be best for me. However, that is not a good enough response since we (I at least) would certainly not do so even if we could, and we feel we would be right not to. And that seems to tell against unrestricted desire-satisfaction, since if satisfying desires is what's good, then there can seem to be no reason to avoid radical downward-adaptation of our desire to the point of being able to satisfy all my desires all the time were it possible. If the right drugs or neuroscientists come around, most of us won't sign up for this, and we'll feel strongly that we're right not to, and we'd like a theory of rationality to accommodate these feelings.

But on our end-relational view, 'that Φ satisfies my desire' is not ipso facto thought to 'provide me with a reason' to Φ. It seems that the basic desire-satisfaction view can't resist downward adaptation insofar as it claims that it's the *satisfaction of desires per se* that provides reasons. So unless there's a restriction on the content of those desires, the more desires satisfied the better, so we should all downward-adapt as much as possible. In order to resist such a noxious conclusion, it can seem that we need to invoke desire-independent beliefs about what is good for ourselves. But, according to the view of the will I've presented (though this theory is hardly necessary to generate this conclusion), those beliefs about our good are the product of our (deep and/or committed) desires, and act as commitment and stabilization devices for those desires. I've also suggested how they can employ emotions such as contempt, fear and loathing to do so, though this account needn't be correct either.[238] Central to (and a key advantage of) the end-relational view is that reasons are always given *from a motivated perspective*. We have no motivated perspective that endorses grass-counting, certainly not the motivated perspective that we associate with our 'deepest desires'. And further, we have powerful motivated perspectives that resist many forms of 'downward adaptation' of our desires.

It's also important to keep in mind that the complaint about lint-collecting as a

[238] Still, it might be thought that even if adaptation were not entailed by the end-relational view, then content-neutrality nevertheless entails that whatever nasty, 'immoral' desire someone actually has is an end relative to which she has a reason to do the nasty thing. That is correct, but doesn't strike me as a problem. We can be expected to condemn such desires, and react to them with negative emotions and unwitting propaganda, but I just don't see any (cognitive, non-strategic) reason to suppose that they are not reasons, relative to those ends. We will have reasons to oppose hers, relative to ours.

good life depends on the assumption that the lint-collector is capable of the ‘normal’ range of emotions. That is no small assumption to make about actual human psychology, that such a person *could* have lint-collecting as their deepest desire, and receive considerable emotional satisfaction from it. I think this is almost certainly false, but I am willing to grant the assumption. If the person receives significant emotional satisfaction from lint-collecting, then they are quite different, so far as I can tell, from any human with the normal range of emotional capacities who has ever existed. I would probably still find that activity absurd and/or contemptible.[239] Or a more likely objection is that although lint-collecting is the person’s deepest desire, it is not very satisfying, though they do have the normal range of *capacities* for emotional satisfaction. In that case, so long as they are concerned about the emotional quality of their lives, they have a very good reason to quit that assanine and unsatisfying lint-collecting and do something more rewarding, maybe even some philosophy.

So our end-relational view is both capable of fending off the charge that it is committed to downward-adaptation, by dint of the fact that it is one’s *actual* desires to which our reasons are related, as well as explaining the powerful emotion-backed commitments and beliefs we have that we should not adapt them downward.

3.3.4 Normative Authority

[239] I would probably just find it absurd. I think contempt serves to ‘push away’ persons and activities that one finds threatening (in a ‘lowering’ way) to the (emotion-backed) commitments associated with one’s practical identity. I find those who lie (including to themselves) for narrow personal gain contemptible. But I can understand the temptation. The contempt signals and reinforces my commitment to be above that. But I cannot imagine being tempted to count grass. So I don’t really find it contemptible exactly, just absurd. But I can imagine being tempted to play video games or watch tv all the time. And I do find that contemptible.

Let's recall Brink's claim that 'it is not at all clear why we should care about the satisfaction of desires independently of the way in which they were formed or of their content'. This reminds me of Emerson's (in)famous declaration: 'If I am the Devil's child, I will live then from the Devil.'[240] This expresses the idea that discovery of who he is is primary, and that this is all he needs to determine how he should live, broadly speaking. Now I don't suppose that quoting Emerson is a knockdown argument, but it's illustrative of a conception of life with which I and many others are sympathetic. There are people who if convinced that their love for their children or their spouses or themselves were 'merely' the result of evolution or even the playing out of a deterministic universe inevitably leading to a certain arrangement of physical stuff in their brains and bodies and so on, would find themselves unable to value those emotions in anything like the same way or to the same extent that they used to. Some philosophers are convinced that we can't believe such things (especially no libertarian free will)[241] and survive.

We can think of many examples of desires we could have with dubious content and origins, but our objections to those desires plausibly come from other, possibly more fundamental desires. If we found out that a neuroscientist or hypnotist had made us like goat cheese or tennis, we might or might not consider what to do about these desires insofar as altering them is possible. But that is plausibly the result of a desire not to have our desires manipulated or some such thing. More seriously, we have a very powerful

[240] From *Self-Reliance*, 1841.

[241] Cf. Van Inwagen (1983) and Smilansky (2000).

desire to feel in control of ourselves, and especially not to feel controlled by others.

This recalls the expression of ideas similar to Emerson's in Frederick Douglass's autobiography[242]. He wondered why he despised being a slave so intensely, even when he had things *relatively* cushy in a New England home. He didn't have the conceptual tools to wrestle with this question at the time, and wondered what it was about him that made him so much more miserable than others at the contemplation of his lack of freedom. The important thing, I take it, is that it just did not matter very much. That was who he was; he had a deep and abiding and powerful desire to escape slavery, and I expect it didn't matter to him whether it came from the devil either.[243]

Some desires seem to be 'at the core' as it were, and the story of how that came to be the case is potentially very interesting, but I don't think, and I haven't been shown, why it should have the capacity to reduce their significance to me, in the absence of some other concern of some kind with which it conflicts. Again, I can readily accept evolutionary arguments to the effect that my love for my (future) children is largely the result of adaptive forces geared toward increasing my inclusive fitness. And so what? I still love them all the same. Now that is not to say that no geneological explanation of desires can (rationally) have a 'debunking' or destabilizing effect on one's desires; I very much think they can. But the way that works, or rationally should work, is by pointing out conflicts between things that I care about. More on this issue shortly.

Brink notes that '[o]nce one recognizes the legitimacy of the question of normative authority, it can seem difficult to answer' (42). As the reader might have

[242] *Narrative of the Life of Frederick Douglass*, 1845.

[243] There's a scenario in which he's in jail and some slave-traders come by and threaten to 'take the devil out' of him for being disrespectful. I suspect the truth or falsity of whether the devil was in him or not ultimately didn't matter to him any more than it did to Emerson.

gleaned by now, there is an important sense in which I don't think this question is legitimate. As has been suggested, I don't think that emotionally and intellectually rich lives are *unconditionally* good. I think that our conception of them as (unconditionally) good is the result of our 'perspective,' i.e., our evaluative outlook which is dependent on contingent features of ourselves. Brink, rightly in my opinion, diagnoses the attraction some have to Kant as a result of just this seemingly difficult question of normative authority. For there seems nothing but a grounding of practical reason in our rational nature itself that could preclude further questioning as to the source of its authority.

I would like to draw upon a nice description of Schopenhaeur's rejection of Kant on this question of rational necessity, provided by Bernard Reginster (2008). The Kantian idea that justifications must end in sufficient reasons is taken to mean that any justification by its very nature requires a series of justifications for each ground until a ground is found that can be given no further justification. For Kant this ground must be rationally necessary; otherwise it is subject to question as to its justification. Schopenhaeur rejects this view, claiming that a sufficient reason is given at each stage of questioning.

This goes for theoretical as well as practical reasoning. If someone asks why it is raining, we can explain how rainclouds form and under what circumstances, and state that a subset of those circumstances obtains now. We can say that a low pressure system moved in yesterday without having to go indefinitely back in time, tracing the causal chain to the Big Bang, in order to have given them an explanation. Nor need we explain the lower-level physics of the higher-level explanation of raincloud-formation. If asked why a vase broke we can (but need not, depending on context) appeal to its molecular structure and the forces that acted on it without having to say why vases have their molecular structure or why the forces acted on it. Explanations give out in the theoretical realm, and it is not necessary to reach the unconditioned (God or an uncaused cause) in order to have a sufficient explanation.

Kant thought analogously about practical reason, and Schopenhaeur rejected a requirement for the unconditional here for the same reasons. Schopenhaeur seems to agree that sufficient reasons are necessary, in the sense of reasons whose 'normative force no longer depends on further considerations' (80), but that these don't require 'pure' reasons, in the sense of rationally necessary ones. This is to claim that a justification can be complete without appeal to unconditioned principles. Rather, when I deliberate from contingent inclinations, those inclinations have prima facie force. This means that absent some reason to call these reasons into question, they are *sufficient* reasons, and the justification is complete. The question whether that inclination is justified *might* arise, but it need not, in the absence of substantive reasons for doubt.

The example Reginster gives (81) is his deliberation on whether he should join an exercise program. He thinks he should because it would contribute to his health. If the question arises, which it can, whether he should be so committed to his health, this question most plausibly and intelligibly takes the form of inquiring about other commitments (his own term) that might be incompatible. In Reginster's case, he mentions commitments to family and achievement, which he then concludes would be served by his feeling good (which is why he wants to be healthy), which is itself a prima facie reason. 'In the absence of additional competing commitments, there is simply no reason to ask questions about *its* justification. It would not be a requirement of reason to do so...' (81). There is not, by the nature of justification itself, a need to question why he should desire feeling good, and only if such a substantive worry (not blanket skepticism) about this desire arises need it be questioned. Therefore, since it is a contingent matter whether an agent has competing commitments, it is contingent whether such an evaluation (that I should join the program) is fully justified. It's important to be clear that competition alone isn't enough to raise a serious worry, since I might have a competing desire to eat ice cream, but my desire to be in good shape is clearly and emphatically prioritized on reflection as (much) better for me overall, and is therefore the rational course of action.

Since these desires are contingent however, it always seems open to question whether one shouldn't change one's desires. This question can certainly be asked, but only from within the 'perspective' of one's entire range of desires. One can't step outside one's desires entirely to ask this question. I can't step outside *all* my desires/perspectives to ask if I should take any of them up. The picture of rationality that presupposes that we be able to do so threatens nihilism about practical reason. Instead, we should realize that our desires and commitments form part of our identity and we can only reason about them from within those perspectives/desires.

This is a crucial difference between the Schopenhauerian and Kantian conceptions of deliberation. For Kant, one uses one's practical reason to deliberate *about* one's desires. For Schopenhaeur, one's desires, or inclinations, are what one deliberates *from* (74). Therefore, rather than our inability to escape our inclinations when considering what to do being a limiting condition on rationality, is a condition of its possibility. Our perspectives 'provide the terms in which we think and reason.' If we give them up, we are not left with an ‘undistorted, unadulterated representation of the 'good as such,'’ but with 'no evaluative judgment at all.' (85).

Now, I’d like to briefly address one aspect of the perfectionist conception of practical reason and the good that Brink advocates. On p. 52, he cites Green as identifying the will with post-deliberative desire. It is put to the purpose of making a distinction between will and desire, where the former is taken to have normative significance whereas the latter does not. I can't see why this would be so. I see no reason why the sheer fact of having deliberated on something should give it intrinsic (that is, relative to no desire) normative significance.[244] Of course deliberation might, depending on the specifics of person and circumstance, fairly reliably lead to outcomes which are better for one overall, and/or better reflect one’s deep commitments.

And of course the fact of having willed an end is always significant from the perspective of the will, the nature of which is to attempt to control one’s future behavior. When it succeeds, its ability to do so tends to be strengthened, and when it fails, its ability to do so tends to be weakened. By the very fact of attempting to control one’s behavior then, it is attempting to strengthen itself. Relative to that inherent end of the will, acting in accordance with it always has normative significance.[245]

But deliberation per se does not seem to be able to confer intrinsic normative significance on action. Deliberation can lead astray as well as aright, depending on the adequacy of one's tools, information and starting point. Perhaps a committed Stalinist

[244] For the purpose of this discussion, I bracket general skepticism about intrinsic normative significance, and wonder why deliberation would confer it if there were such a thing.

[245] This is of course to distinguish ‘acting in accordance with one’s will’ from acting volitionally. I only employ the distinction here to illustrate a sense in which there is always a form of significance associated with ‘willing’, where that is understood as following through on a commitment, that does not attach to volitional behavior as such.

doesn't pre-deliberately want to starve the Ukranians to death, but after deliberation on the importance of subduing sentimentality to the service of the Five Year Plan, comes to want to (and perhaps want not to at the same time). Or a U.S. Air Force bomber doesn’t want to kill thousands of peasants, but after deliberation on the importance of stopping dominoes from falling, comes to want to. Examples can be multiplied ad libitum (there will be an extended one in the next section).

Of course Brink recognizes that there have to be constraints on the content of choice, i.e. that the will (or choice) is not the *only* source of normative significance. I want to insist that *in itself*, it seems to be of no more significance than pre-deliberative desire. Deliberation may take indefinitely many forms, one of the most common of which is the question how to best serve one’s god(s). Depending on one’s conception thereof, the deliberation may go in any direction. Any of various principles may serve one well or ill, and no principles at all might often serve as well as any principles to which someone has access.

I want to be clear that I think deliberation is generally very important and valuable. Not only does it tend to orient us toward our longer-term goals and commitments, but it gives us a chance to gather relevant facts and think clearly about what is likely to happen, and to reflect on how we are likely to feel, if we do x rather than y. Acting in accordance with our deliberations can make us (feel) powerful, in at least two different ways.

First, our deliberative capacities are a tremendous source of power in themselves, allowing us (inter alia) to predict fairly reliably what would happen in limitless possible scenarios, without having to actually try them to find out. And second, the results of our deliberations are often that we should do things that we don't really feel like doing. When we do these things anyway, especially when we can see the rewards of having done so, this combines the sense of our power to make a good judgment with that of the power to carry it out in action. When we make a decision to do something, this automatically engages our wills, which in and of itself provides a reason to act accordingly. These considerations seem to be enough to explain whatever apparent *intrinsic* significance deliberation has, without undermining its acknowledged centrality to the quality of our lives and our experience of the power of our agency.

Perhaps at this stage a rationalist would concede that I have outlined an alternative perspectival view, but that I have not yet shown that it is the correct view. That is, I have not yet shown that we can't step outside all our desires and ask if we should take any of them up. I have not yet shown that we cannot evaluate our entire desire-set from the standpoint of what it would be reasonable to treat as values or from the perspective, so to speak, of pure practical rationality. And that is correct; I have not shown it, certainly not in the strict sense. But I do think that I have provided good reasons to think that the 'Schopenhaurian' view that we reason from desires and not about them is more plausible, mostly by undermining reasons to think that we ever can or need to step outside our desires to make these judgments.[246]

The view that we can do so posits some capacity over and above what the Schopenhauerian view posits. That capacity is thought necessary to explain features of deliberation, justification and normative authority (perhaps normative accommodation as

[246] The idea that we need to but cannot is what I meant above when I said that the rationalist view threatens practical nihilism.

well) that the desiderative view cannot, which is presumably why the posit is thought warranted. I have argued that no such capacity is necessary to explain these features. The motivations of rationalism are in large part predicated on the putative failures of anti-rationalism. I have been showing that my view doesn't have these failures, in which case the rationalism begins to look unmotivated, or less motivated, such that we might wonder why we should think there is any standard to be found in practical rationality itself. All that said, I agree that at this point in the essay my perspectivalist view is not demonstrably correct, but I do think it is more than just a coherent alternative at this point. I take myself to be in the process of undermining the motivations and reasons for being a rationalist, a process which I grant is not yet complete. One reason it is not is that I have yet to substantively engage with some other prominent rationalists.

3.4 End-relationalism vs. the Varieties of Rationalism

The 'perspectival' or end-relational conception of practical rationality (and normativity generally), combined with and supported by an Ainslean model of the will are doing a lot of work on my account. If the rough relationships between rationality, normativity and the will as presented in my view are plausible, it promises to help reconceive neo-Humeanism in a way that allows it to respond effectively to most or all of its criticisms.

The rest of the chapter will continue a theme I began above when I argued that my view gets us all the good things about Korsgaard's view without the bad things. I argued that mine was to be preferred largely on the basis of the fact that the model of the will I am employing can capture the intuitively plausible things about Korsgaard's conception of the connections between practical rationality and the will. Further, it is a vastly more empirically-grounded, unifying, sophisticated and richly elaborated model of the will than Korsgaard has to offer, capturing vastly more 'data', in addition to giving more intuitive answers to what counts as 'irrational', at least in some cases. And this model has no executive rational agent residing beyond all desire or reward, who decides which actions get taken on the basis of their perceived rationality.

I dealt with Korsgaard above because I wanted to address in detail the objection that my view entailed skepticism about practical reason. But now I return to advertising the superiority of my end-relational view over rationalist alternatives, largely by attempting to show that several leading rationalist views are largely predicated on a misconception of the connections between rationality, evaluative belief, desire and the will. Specifically, I'll show how these rationalists make a (near-) identification of practical rationality with willpower. In the process, I'll also show how my end-relational view can overcome other rationalist objections to reductive desiderative views of practical rationality.

3.4.1 The Weakness of the Willpower Conception of Rationality

I mentioned above that many philosophers judge akrasia to be the paradigmatic form of practical irrationality. I argued that we could understand this in terms of the fact that the perspective from which judgments of irrationality are (typically) made is that represented by one's relatively stable and long-term commitments, the consciously available evaluative beliefs associated with them, and the experience of willing in accordance with these commitments and beliefs. This might not yet strike the reader as plausible. So I propose to use the view I'm developing to both criticize and explain this 'willpower conception of (practical) rationality' as I'll call it. That is, the conception of rationality that holds that acting (including desiring and/or thinking) contrary to one's evaluative beliefs and/or judgments are not only paradigmatically, but necessarily irrational. Our end-relational view can explain this commonly held idea, and show both what is right and wrong about it.

Smith's and Scanlon's Versions

I'll start with Smith's (1994) conception of (ir)rationality. Smith has an anti-Humean (anti-desire-satisfaction) theory of normative reasons. Smith analyzes 'our' belief that we have a (normative) reason to Φ as a belief that we would desire to Φ if we were fully rational. So, if we believe that we have a reason to Φ and do not come to desire to Φ, Smith holds that we 'certainly are' irrational, '[a]nd by our own lights ... if we believe that we would desire to [Φ] if we were fully rational then we rationally should desire to [Φ]' (177).

Now I can see why, especially on the model of the will I've provided, one would say that if someone believed that they would desire to do something if they were fully rational, then that gives them *a reason* to desire it. But to say that they are irrational for not doing so seems to go too far, if the charge of irrationality is supposed to retain its usual normative force. For there are many cases in which someone might not want to, even might want strongly not to do something that we would regard as heinous, but yet believe that they have reason to do it. I don’t think it is a good theory of rationality that holds that someone in such a position is irrational for not forming a desire to do so.

Now it could be that what Smith has in mind by ‘irrational’ is just that they should have *some* desire (motivation) to act accordingly, even if they should have far more to act contrarily. But this is not at all the sense one gets in reading him, and there is no suggestion that he’s conscious of the kind of problem I’m raising here. The fundamental problem as I see it arises from the basic premise that desires fundamentally ought to serve reasons and not the other way around. It is such a picture that, in my opinion, is in part responsible for our blindness to the possibility that willpower could have any downsides (or perhaps it’s the other way around, or they are mutually reinforcing).

Earlier in the book (69-70), Smith employed the ‘fact’ that a ‘good and strong-willed person’ is motivated to act differently (and accordingly) upon changing her moral judgment as an argument against motivational externalism. Externalists, he thinks, cannot explain this. Internalists however, naturally explain the fact that ‘good and strong-willed people’ change their motivations in light of new moral judgments. The reason they do so is that ‘genuine’ moral judgments result in corresponding motivation, in the absence of weakness of will.

I am not taking up the question of internalism vs. externalism here, but only want to point out that the 'good' in 'good and strong-willed' is worse than superfluous. Maybe they changed their judgment from Jews should be left in peace to Jews should be killed. Now perhaps a ‘strong-willed’ person would have a corresponding desire to kill Jews, but it's entirely unclear in what sense the person would be good over and above having a strong will. It's telling that Smith sees no need whatever to address this objection, and it helps to underscore the great potential for damage if we are ignorant of the downsides of will, especially in contexts of its exploitation.

Scanlon's (1998) notion of irrationality is representative, in the sense that most people would agree that the things he regards as (clearly) irrational are so. I like his conception because he is in fact concerned to *restrict* it to a 'narrow' sense, but nevertheless offers what would probably be generally agreed upon as 'clear cases of irrationality.' These are cases in which a person 'fails to form and act on an intention to do something even though he or she judges there to be overwhelmingly good reason to do it' (25). This case differs from Smith's primarily in that Smith required that a desire form while Scanlon requires an intention to form and that one act on that intention. This is much more nearly an identification of willpower with rationality.

Now I don't want to deny that there's *something* wrong with a situation in which one judges there to be overwhelming reason to do something and yet does not do it. What I want to highlight is that in both cases the authors assume that in a case where someone believes they have (overwhelming) reason to do something, that if they do not desire it (Smith) or form and act on an intention to do it (Scanlon), the fault must lie in the fact that they are not *doing* or *desiring* what their faculty of reason has commanded. Such cases suggest that it is simply part of what it *means* to be irrational that one is not guided by one's (subjective) reasons. But though this may be a 'failure' of willpower, it is neither necessarily an overall bad thing nor a fault of any reason but whatever, if any, reason led them to the belief in the first place.[247] Once again, the problem may be that they have an evaluative belief (as a commitment device) that is divorced from, and/or or in conflict with, their (underived) desires and is thereby hijacking their will.[248]

Huck Finn

Mark Twain gives us perhaps the most compelling (imagined) case of just such a phenomenon in his masterwork, *The Adventures of Huckleberry Finn*. I'll recapitulate the scenario briefly, illustrating the power of my view to analyze such cases. I'll also use it to help me make several points in favor of the view I'm offering and against those I'm opposing.

Jim is Huck's (adoptive) aunt's slave. Jim overhears that he is about to be sold and shipped to New Orleans, away from his wife and children, who are all owned by nearby people, and whom he plans to buy or steal. He happens to meet with Huck, who has also run away, and they begin their adventures with Jim telling Huck that he's run away and why. Huck promises not to tell anyone about it.

But soon, he begins to feel very guilty about helping Jim run away. He 'realizes' that he is helping Jim steal (himself) from his aunt, who 'tried to be good to [Huck] every way she knowed how'. His conscience made him feel like he would 'die of

[247] As we saw in Chapter 2, many evaluative beliefs are often provided with reasons post-hoc. We unreasoningly absorb very many of these beliefs from our peers and parents. And even if we do reason ourselves to them, that reasoning can be pathological, as exampled presently.

[248] Note that I do not think that one needs to experience one's will as having been 'hijacked' in order to make this judgment. Addicts often feel this way, but very often people identify with what appear to be pathologies of the will, most obviously compulsives.

miserableness' and he continually 'abus[ed] himself' for not having the moral character to turn in Jim.

He continues to be miserable, his conscience upbraiding him mercilessly, until the moment he resolves to turn Jim in at the next opportunity, at which instant he feels 'easy and happy and light as a feather'. As soon as he gets the opportunity, he shoves off in his boat to go tell somebody on shore that Jim's run away:

> …and as I shoved off, [Jim] says:
> "Pooty soon I'll be a-shout'n' for joy, en I'll say, it's all on accounts o' Huck; I's a free man, en I couldn't ever ben free ef it hadn' ben for Huck; Huck done it. Jim won't ever forgit you, Huck; you's de bes' fren' Jim's ever had; en you's de only fren' ole Jim's got now."
> I was paddling off, all in a sweat to tell on him; but when he says this, it seemed to kind of take the tuck all out of me. I went along slow then, and I warn't right down certain whether I was glad I started or whether I warn't. When I was fifty yards off, Jim says: "Dah you goes, de ole true Huck; de on'y white genlman dat ever kep' his promise to ole Jim."
> Well, I just felt sick. But I says, I *got* to do it – I can't get *out* of it.

He's quickly met by some men who announce they're looking for 5 runaways and ask him if there's anyone on the raft with him. He tells them there's one man, but when they ask whether he's white or black, Huck tries to tell them the truth, but he 'warn't man enough -- hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying.' And then he cleverly lies to them, saying it is his father on the raft and implying that he has smallpox. This sends them on their way.

> They went off and I got aboard the raft, feeling bad and low, because I knowed very well I had done wrong, and I see it warn't no use for me to try to learn to do right … Then I thought a minute, and says to myself, hold on; s'pose you'd a done right and give Jim up, would you felt better than what you do now? No, says I, I'd feel bad -- I'd feel just the same way I do now. Well, then, says I, what's the use you learning to do right when it's

> troublesome to do right and ain't no trouble to do wrong, and the wages is just the same? I was stuck. I couldn't answer that. So I reckoned I [would] after this always do whichever come handiest at the time.

There are several points I want to make. First, he only begins to feel badly about what he's doing when he conceptualizes what he's doing (or abetting) as *stealing*. Only when he connects the idea of helping someone take some of his aunt's property with helping Jim run off does he become, I think it's fair to say, disgusted with himself. Second, there are clearly two primary motivated perspectives here—a powerful (not at all 'calm'!), stable one associated with an evaluative belief, and a more fleeting one, not associated with any evaluative belief. Guilt is plausibly involved in motivating both perspectives**.** Third, he feels weak-willed by failing to turn Jim in. And fourth, any conception of rationality that conceives of akrasia of this sort as (paradigmatically) irrational is committed to the view that Huck's actions were (paradigmatically) irrational.

How does my view analyze this case? First, how does Huck come to feel so guilty in the first place? Huck comes to feel guilty when he thinks of what he's doing as (abetting) stealing. And worse, (abetting) stealing from his aunt. But of course there is a very large difference between helping a person steal himself so as not to be sold into brutal conditions far from his family and helping a person steal some furniture.

Why doesn't this difference seem to matter as far as how guilty Huck feels? One might say that Huck doesn't regard Jim as a person. But this claim would be wrong. It is belied many times over in the book, but nothing could be better evidence than that he shares equally between himself and Jim the forty dollars that the slave-hunters gave him as a consequence of their own feelings of guilt for not helping Huck's (supposedly) smallpox-stricken father. A plausible answer is that he feels (roughly) the same about helping Jim steal himself and helping him steal some furniture (at least in large part) because *he assimilates them by principle*. 'Stealing is stealing.'[249] By hypothesis, the *whole point* of principles is to lump together actions and contexts that might be quite different in important respects under a common category and treat them all the same.

Doing this makes Huck feel very guilty. This emotion is either causal to or (partly) constitutive of the evaluative belief that helping Jim is (very) wrong. This evaluative belief represents an emotion-backed commitment to turn in Jim. But we saw that there is another motivated perspective within Huck. He feels guilty at the thought of turning Jim in, especially when his attention is drawn to their friendship and Jim's trust in and affection for him. Why doesn't this perspective acquire the status of a commitment? Plausibly, there are no available norms (at least we don't see any in Huck's mind) with which he can *justify* these feelings of guilt. I argued in Chapter 2 that having access to norms that justify one's feelings have a stabilizing and committing function. Huck at no point tells himself a story by which his feelings of guilt at 'telling on' Jim are justified, or by which his turning him in would be condemnable. He doesn't even explicitly (though presumably he does implicitly) represent it as *betrayal*. If he had, he would have had a means to stabilize those feelings, for to *explicitly* categorize something as an instance of betrayal generally has significant emotional/motivational consequences. By hypothesis,

[249] Earlier, during his travels Huck eases his conscience by telling himself that he's 'borrowing' food from people's cornfields, as opposed to saying that he's stealing it but that it's ok under the circumstances. This is I hope an entirely familiar sort of behavior, and one that speaks to the importance of categories like 'stealing' on one's emotions and motivations. Huck is also shocked to hear Jim 'come out flatfooted and [say] he would steal his children'.

it has these consequences by providing the additional level of stabilization (or commitment) associated with evaluative beliefs.

Given that one of Jim's perspectives was supported by available norms and principles, it is the perspective that was stabilized and represented as an evaluative belief, an emotion-backed commitment which gave rise to the conscious experience of resolute willing in accordance with it. Then 'failure' to do so is experienced as weakness of will. Huck evaluates the morality (rightness) of what he does from the perspective of this belief.

This belief and the motivation that comes with it are especially pathological, because on an Ainsliean understanding, Huck has not only *motivation* but *reason* to turn in Jim, because failure to do so, so long as he perceives it as resulting from weakness, will weaken his resolve to do the right thing in the future. Which of course it does; in fact he gives up entirely on doing right since 'the wages are the same' as doing wrong but doing right is harder. In the story-world, since we find the norms abhorrent and Huck a good kid, this is an ironically charming result. In the real world, this can in some cases be a charming result, but it can also be a serious cost.

What are we to say of Huck's rationality? On Scanlon's view his actions are straightforwardly irrational. Suppose we change the story such that his friendship with Jim had had much more time to flower, and his commitments to Jim were much stronger and more stable. Then perhaps some slick talker could have convinced him, eventually, that turning in Jim would have been the rational thing to do, specifically convinced him that if he had full information and rationality he would desire to turn Jim in. But then suppose he yet formed no desire to do so. He is now irrational on Smith's view, 'and by

[his] own lights'. Again, we could see Smith as only maintaining that there must be some minimally faint, ineffectual desire, on 'pain' of irrationality. But again, that is not at all the sense one gets in reading him. That would be such a weak requirement that it would undermine his whole project, which is to show that morality consists of (true) categorical imperatives. He will not be satisfied with a view that ends up only requiring rational agents to have only the faintest of desires to be moral.

Note that this problem is not easily resolvable by appeal to subjective vs. objective rationality. To say that he was subjectively, though not objectively irrational is to say that relative to his epistemic situation, he should have (it was rational for him to have) betrayed Jim. But that seems wrong. It seems that he should have seen (attended to the fact) that Jim was his friend and that betraying his friend and dashing all Jim's hopes for his and his family's freedom would be (vastly) worse than failing to return Jim to Huck's aunt. At minimum, it seems he should have seen that there were competing considerations. But he acknowledged only one genuine consideration, and considered the contrary desire nothing more than a source of weakness at the time.

Scanlon might object in the following way. It's far from clear that Huck has a clear and unconflicted view of what he ought to do. He's torn between a moral code he's been explicitly taught and partially internalized, which treats slaves as property, and an opposing moral code or set of normative commitments that is inchoate and that he is discovering in himself, according to which slaves are free and equal people. He ends up spurning the code he inherited for the one that he has discovered based on his own experiences with Jim. Perhaps Scanlon can claim that it is more a case of moral conflict than traditional weakness. And since there is such conflict, he does not fit Scanlon's requirement of considering himself to have overwhelming reason to turn Jim in, and is therefore not irrational for not forming and acting on an intention to do so.

I think this reads a fair bit too much into Huck's psychology, especially the part about a 'moral code'. I agree that he is discovering the commitments (all practical commitments are normative in this sense) in himself. But to say that he's discovering a moral code seems too much. If he were to think about it he might *create* a moral code or fit his commitments into some pre-existing moral code. But he has not yet 'codified' these emotions in the way he has stealing.

Further, he doesn't spurn that code. It's really important that he doesn't. He regards himself as too weak to carry it out. The closest he gets to spurning it is to 'realize' that it (doing right) pays the 'same wages' as doing wrong, but the latter is easier. That is as close as he comes to a justification of what he's done. Later in the book he goes through a similar dilemma and berates himself again for not turning in Jim in a new circumstance. This shows that he has not spurned the prior code but still regards himself as weak, unable to do his moral duty.

Still, Scanlon might insist that this is not a paradigmatic case of akrasia. Now it's true that this is not a *completely* paradigmatic case of akrasia. But very many cases of akrasia, especially in the moral realm, are not paradigmatic. But what makes it unparadigmatic is not that there's ('moral') conflict. The vast majority of akratic cases (in the moral realm at least) involve conflict that can be construed as 'moral conflict'. I think the things that make it unparadigmatic are 1) that we approve of the passions that overcame him and 2) that after the fact he seems somewhat ambivalent. However, it does look like passions overcame him. That is paradigmatic. And he doesn't even rationalize at the time of action (though you could interpret it in a way that he did—it wouldn't make it less paradigmatic). After the fact, he reverts to the judgment (if he ever left it) that what he did was wrong, but simply concludes that he is too weak to do what is right. That looks very paradigmatic, except for the part about his giving up on doing right and being ambivalent about it.

Scanlon says that if one subjectively believes that one has overwhelming reason to do X, then if one does not form an intention to X and act on that intention then that is paradigmatically irrational. I submit that this is exactly what the Huck case looks like. In the case at hand, from Huck's perspective (and mine), he *was* overcome by passion! We like those passions and desires and also might think that his being overcome by them was good for Huck. We have norms that stabilize and otherwise support those passions. But it would be hopeless for Scanlon to point to the conflict he experiences (and the conflicts that people generally experience in cases such as these) and say that that means that he didn't really think he had overwhelming reason to turn in Jim. People experience themselves as having conflict in all kinds of cases where they have judged themselves to have overwhelming reason to do something. And then when and if they are overcome by passion, they tend to rationalize their actions to themselves such that *at the time* of action, they do not so judge. Often, like Huck, if they fail repeatedly or spectacularly, they quit trying and perhaps then must begin to change their judgments about what they should do, if only on pain of feeling utterly feckless.

If the standard for Scanlon is that people are paradigmatically irrational just if they do not form and act on intentions in response to judgments that they have overwhelming reason to X *at the time of action*, then that would be just to exclude almost all of the interesting cases. And further, it can't plausibly be what he means. The point, presumably, of including 'forming an intention to act' over and above just acting is that the forming of the intention is supposed to carry the agent until and through the time of action. The problem with akrasia is generally that the intention formed is not carried out, and that is often aided by the process of rationalization, the effect of which just is to undermine that prior sense of having overwhelming reason.

So conflict, whether considered moral or not, is the norm, as well as having one's sense of having an overwhelming reason undermined at the time of action. If Scanlon wanted to categorize all the subjectively akratic cases in the history of sex, revenge, war, etc. as those of moral conflict in a way that shielded him from the kind of critique I employ here, it seems that he would have to do so in a terribly ad hoc way that would render his view entirely uninteresting. And/or he would have to build into the idea of rationality not only that it is subjectively-held-reason-following, but also that the normativity of doing so is in principle immune to the entire class of criticism I raise here. No matter the extent to which someone says they think they have overwhelming reason to do X, if we approve of the conflicting motivations and rationalizations that cause them not to X, then we call 'moral conflict!' and deny that they really thought they had overwhelming reason in the first place.

Now that is uninteresting. What is interesting is the empirical/theoretical claim that willpower has potentially serious downsides, and therefore a conception of practical rationality that is (nearly) identified with willpower will have to accept that being practically rational has potentially serious downsides. I think the only feasible strategies are to either abandon this conception of rationality or to divorce the rational from the normative and accept that being practically rational can, even often, be the wrong thing to do. I favor the former strategy, since I see little reason to divorce practical rationality from normativity, and lots of reasons to favor an end-relational view of practical rationality.

Still, there is some truth in Scanlon's (imagined) response. That truth is that what we conceive of and experience as 'weakness of will' is not always (if ever) best interpreted as such. From our perspective, we don't have to see Huck as weak, though he sees himself that way. Given the concept of 'affect as information,' (from Chapter 2) we can see his guilt feelings at the thought of 'betraying'[250] Jim as important information, to which he did not have conscious access. Indeed, he realizes, only after thinking himself so weak that he could never learn to do the right thing, that he would have felt no better had he turned Jim in. His feelings of guilt at the time of contemplating this act were information to that effect, though he didn't conceive it as such.

Let me continue my imaginary objection from Scanlon. He might say that surely I am not denying that Huck had conflicting values. Now, if values are beliefs, then he had conflicting beliefs about what he should do. And if he had conflicting beliefs about what he should do, then it is implausible after all that the thought he had overwhelming reason to do one of them. To think that would in effect be just to abandon the other belief. But, as I admit, he does not abandon it. Therefore, our imaginary Scanlon continues, it is I and not Scanlon who is making Huck out to be irrational, by saddling him with the belief that he has overwhelming reason to do A, but also the belief that he

[250] The motivations associated with betrayal do not require it to be conceived as such in order to *exist*, but they are *stabilized* by so conceiving of it (and likely altered somewhat as well).

should not do A. While it is not irrational to have conflicting desires per se, having beliefs that conflict in the manner described above does seem irrational. Therefore, if I grant that Huck has conflicting values but is not (paradigmatically) irrational, I have to deny that values are beliefs. This I do deny, and doing so brings us to our next rationalist purveyor of a willpower-conception of rationality.

3.4.2 Values are Desires

Michael Smith (1994) argues that values are reducible to beliefs, and not desires. The most influential advocate of the reduction of valuing to desiring is David Lewis (1989), who analyzes valuing as desiring to desire. The view is intuitive, and a short exerpt should suffice to give the gist:

> The thoughtful addict may desire his euphoric daze, but not value it. *Even apart from all the costs and risks*, he may hate himself for desiring something he values not at all…He desires his high, but he does not desire to desire it. He does not desire an unaltered, mundane state of consciousness, but he does desire to desire it. We conclude that he does not value what he desires, but rather he values what he desires to desire. (1989, 115, my emphasis)[251]

This allows Lewis to distinguish between motivating and normative reasons, where the first could be had by first- or second-order desires, but a justifying reason is provided by a second-order desire.

Smith notes that Lewis wisely denies that an action *must* have some associated second-order desire in order to be rationally justifiable. For Smith, this seems to be for

[251] Taken from slightly longer exerpt on 142-3 of Smith (1994).

the following reasons. First, it appears that the following is a constraint on normative reasons. 'If an agent accepts that she has a normative reason to Φ then she rationally should desire to Φ' (143). Therefore if to accept a normative reason claim is to value, and to value is to desire to desire, then an agent who has a desire to Φ, and a desire to desire not to Φ, rationally should rid herself of the desire to Φ and acquire a desire not to Φ instead. But, Smith asks, are our first- and second-order desires normatively related in this way?

Not on the Humean maximizing conception of rationality, they aren't (says Smith). On this conception, whichever desire is stronger provides greater reason to act accordingly. So if the thoughtful addict's desire for an euphoric daze is stronger than his desire for an unaltered state of consciousness, then he rationally should frustrate the latter and satisfy the former. So then it is not the case that an agent rationally should desire what she accepts there to be a normative reason to do. This result seems unacceptable (to Smith and many others).

Smith has two more considerations to bring against the Lewisian view. The first is that this view has no basis upon which to claim that reason is on the side of altering our first-order desire so as to match our second-order desire. That is, even if we assume that 'reason is on the side of a *harmony* between our first-order and second-order desires' (145, emphasis his), which way to achieve that harmony seems left completely open, or worse, if we are 'Humean', is determined by whichever desire is 'stronger'. Smith chalks this up as a big problem for the project of analyzing values as desires, one that doesn't arise for analyzing them as beliefs. For in the latter case, 'if any change is to be made at all, there is a principled reason why in cases of disharmony an agent's desires must change so as to match her values: namely, the fact that an agent's contrary first-order desire *will not be in any way suggestive* of the fact that her evaluative belief is untrue' (145, emphasis mine).

Smith's final consideration, in fact 'a decisive objection' against the reduction of valuing to desiring, is that there is no principled way to locate values as *second-order* as opposed to some other-order desire. 'Those who seek to reduce valuing to higher-order desiring of some sort *must come clean* and identify valuing with higher-order desiring *at some particular level or other*' (146, first emphasis mine, second his). And Smith rightly contends that this cannot be done. Lewis does not claim the highest-order desires are identical with values because it does seem possible to desire to value differently than one actually does. And it must be at least second-order, for that is the source of its intuitive appeal. But, Smith argues, all we have to do is imagine someone with four orders of desire to see that to exclude the first and highest orders in no way entails that it is the second. More pointedly, choosing the second is 'simply arbitrary, given [Lewis's] premises'. Which just goes to show 'how formidable the original objection really is' (147). If we could choose the third with just as much justification as the second, fourth, fifth or any order desire, so long as there is one higher, then 'we cannot identify valuing with desiring at any level' (147). Thus ends Smith's case against the Lewisian reduction of value to desire.

I can easily handle these objections

Not only may all these objections be easily handled by my end-relational view, but in doing so I will make a powerful (perhaps even decisive) objection to Smith's account of valuing as believing. Further, the mistakes Smith (and others) makes can be explained by my view's apparatus, which includes the concept of affect as information. First, we've rejected the notion that 'stronger' desires give 'stronger' reasons. Judgments of rationality are made from some motivated perspective(s) or other; within a person, typically from the perspective associated with her evaluative beliefs and conscious willings.[252]

Second, Smith is both right and wrong when he says that 'reason' is not necessarily on the side of one's second-order desires. He is right in that if we think of (practical) reason as non-relational, non-perspectival, then there seems no reason to suppose that it must come down on the side of second-order desires. He supports this claim by assuming that the value-reductionist has to accept the conception of rationality on which stronger desires provide stronger reasons. But even if we reject this conception and go with some other non-relational conception, it might not be obvious why reason would always or even tend to favor the second-order desires.

But since judgments of rationality are, within a person, made from the perspective of one's consciously available commitments, evaluative beliefs and/or higher-order desires, Smith is wrong to say that there is no reason to assume that reason is on the side

[252] The attempt to have stronger desires entail stronger reasons is the result of a mistaken attempt to capture a non-relative fact of the matter as to which reasons are stronger than which, while at the same time relativizing them to desires. But it's relational all the way down.

of the harmony going in the direction of these second-order desires. Second-order desires are a lot like evaluative beliefs in important respects. They are consciously accessible, statable in propositional form and almost always the product of and/or supportable by consciously available reasoning. Second-order desires are very often if not always the result of, or at least supported/stabilized by, some sort of consciously available evaluation. When second-order desires conflict with first-order desires, the consciously experienced will is typically engaged (if it is engaged at all) in the service of the second-order desires, just as it is in the service of evaluative beliefs. This partly explains the attraction that I have argued Smith, Scanlon, Korsgaard and others have to a 'willpower conception of rationality'.

What of Smith's contention that a first-order desire can in no way be suggestive of the falsity of an evaluative belief, whereas first-order desires seem to compete with second-order desires without 'reason' having a clear stake in which way, if at all, harmony between them is achieved? Both of these claims are false. We just saw why, or at least that, reason does (at least tend to) side with second-order desires. Further, even if one doesn't accept my arguments to this conclusion, Smith's claim that first-order desires cannot be evidence of the falsity of an evaluative belief is problematic, in the light of my Huck Finn example and the concept of affect as information.

Huck Finn's situation is an unusually vivid and gripping rendering of a general phenomenon that is not at all unusual. Recall our Stalinist and our Nazi and our American bomber. Some Nazis considered their actions at the gas chambers so heroic in large part because of how strongly averse their first-order reactions could be. But they were 'good and strong-willed', and so successfully and effectively desired to do their duty as they saw it. Now we have seen in what sense, on my view, reason was on the side of their evaluative beliefs. I'm not clear about the sense in which reason was on their side on Smith's view. Had they not desired to gas the Jews, they might have been irrational *by their own lights*, but that does not entail that we have to make the same judgment.

We saw in chapter 2 that affect plays an important role in decision-making. The intuitions associated with this affect are the result of *cognitive processes*. Not 'cognitive' in the truth-functional sense typically used in metaethics, but in the sense of involving (highly complex) information-processing. Affective states and conscious reasoning seem to be subserved by distinct (not to say unconnected) systems in the brain. *But they are both important!* If George has a strong first-order desire not to torture someone, then that *can be in some way suggestive* that the evaluative belief that he should do so is false (especially if he knows about affect as information). Examples can be multiplied indefinitely, in both moral and prudential contexts.

Therefore Smith has a dilemma. If he maintains that his conception of values as beliefs *entails* that first-order desires can't be evidence against those beliefs, then that is, in my view, *very* strong evidence against his conception. Because they absolutely can. If he admits that desires can sometimes be evidence against the beliefs, then he loses one of his ostensibly key advantages over valuing as desiring. Moreover, my view can easily handle 1) why it is that desires are not *generally* good evidence against evaluative beliefs, 2) why they sometimes *can be* such evidence and 3) why it is that we would be likely to resist the truth of (2).

Smith might think he can forego this claimed advantage over values as desires, since reducing valuing to desiring is still vulnerable to his ‘decisive objection’. This objection is however not decisive, but rather just confused, mostly due to its unargued and unmotivated requirement to ‘come clean’ and identify some particular level at which desiring is identical to valuing. Lewis is exactly correct to resist claiming that this identity must be at the highest level, for it is not only possible but actual that people wish to value differently than they do. But there is just nothing dirty about not thinking that there is not one particular level at which desiring is valuing. In fact, once one has dismissed this requirement, it starts to look quite odd that someone would think it legitimate.

We should be able to see that we can and do have values at different ‘levels’, corresponding to different orders of desire, that can be in conflict with each other. I assume most of us can easily empathize with the following general sort of circumstance:

> I wish that I didn’t wish that I didn’t wish to eat cream cake. I wish to eat cream cake because I like it. I wish that I didn’t like it, because, as a moderately vain person, I think it is more important to remain slim. But I wish I was less vain. (Elster, 1989a, n37.)

We can (and will) use Ainslie’s model to help us analyze this case, but we don’t need it to see that conflicting values at different levels are possible. We can value our appearance but wish we didn’t (so much) because we think it interferes with a ‘higher-level’ value. We can think our vanity gets in the way of us living the most rewarding life we can live, not only because it denies us cream cake sometimes.

Note that this is not to say that we value our vanity but also have values inconsistent with it. We do not, in this case, value our vanity, but the fact that we are vain means that we value our appearance.[253] These values can be naturally understood as representing interests (desires) at different time-ranges.

> His long-range wish is not to be vain, which defines vanity as a temporary preference in the sellout range. His vanity is in turn threatened by an appetite for cream cake, a temporary preference in the addiction range. (Ainslie, 1992, p. 120).[254]

Also note that the first- and second-order desires in the original example given by Lewis are completely naturally understood in terms of desires at different time-ranges. We don't suppose that the addict wants to want a 'mundane state of consciousness' for its own sake, but rather (at least in large part) because the temporary euphoric daze is had at the cost of longer-term rewards. If he is a 'workaholic' and wants the mundane state so he can work more, but wants to want to work less so he could enjoy his family and/or other things in life, then as a workholic he values high productivity, but he does not value being a workaholic.

What both Lewis and Smith fail to see is that Lewis's reluctance to place valuing (exclusively) at the highest-order desire, since one can wish that one's values were different, is a simple consequence (or statement) of the fact that we can and do have values that conflict with our other values! Which means that not only is there no

[253] This entailment is only meant to be one-way. Valuing appearance doesn't entail vanity, but vanity does entail valuing appearance (perhaps in a particular way).

[254] 'Sellout' and 'addiction' are here just names for the time-ranges at which these temporary interests operate. It doesn't imply that liking cream cake entails that one is addicted to it in the normal sense of addiction. Importantly, without the vanity (in this example) the cream cake would not be a temporary preference at all. It is temporary just in case it conflicts with some desire at a longer time range.

requirement to pick some particular level at which to identify valuing as desiring, but rather *there is a requirement not to*! And what separates values from desires is that values are commitments. I don't think it would be felicitous to describe all commitments as values, but I do think all values are describable as commitments. And on the direction-of-fit model, which Smith endorses, commitments are desires.

Incidentally, note that on the direction-of-fit conception, to desire to desire something *for its own sake* is to desire that very thing. To want to want to be healthy is to want to be healthy. This is the explanation for why the person wants to want to be healthy. When we say that we want to want to stop smoking or doing heroin, we are thinking about desire in its more usual usage, where it implies a feeling of being pushed or pulled. We might not feel any real pull to stop smoking or shooting up, such is the state of our addiction and perceived (lack of) power to stop. But if we want to want to stop, and this is so we will be healthier, then we want to be healthier. People don't generally want to stop smoking or doing heroin 'for its own sake'. If smoking made people healthier and their teeth look nice and breath smell good, they generally wouldn't want to want to stop. If heroin didn't have its catastrophic consequences, people wouldn't generally want to want to stop that either (not entirely at least). It's because we care about our health (and our attractiveness) that we want to want to stop. Just as at another time smoking's ostensibly raising one's attractiveness made some people want to want to smoke. Wanting to want something for its own sake is already to want that very thing.

This is good for the end-relational view. If I have no desire to be healthy but only want to want to be healthy, then there would be no motivated perspective from which I have a reason *to be healthy*, but only a reason to *want* to be healthy. This seems like a serious cost, but it is avoided if we realize that anything we want to want for its own sake we already want for its own sake. We also presumably want the pleasure of the cigarette for its own sake, but acting on these desires (very often at least) is irrational, as always, from the perspective of one's longer-term/higher-ranked desires.

The end-relational view, along with its associated tools, has, in my view, convincingly rebutted every one of Smith's arguments against the reduction of valuing to (n-order) desire. That's important because that attempted reduction is close to my view, insofar as I think that desires to desire represent commitments (of widely varying strength and character) and that these commitments seem to map fairly closely onto what we call values. In addition, we've presented a powerful challenge to the attempted reduction of values to (evaluative) beliefs, *as Smith conceives of them*. For he thinks that first-order desires can never be evidence against such a belief, when it is plain that they can. However, if we conceive of evaluative beliefs as commitment devices which help stabilize (normally long(er)-term) desires, then it's straightforward that first-order desires can be 'evidence' against them.

Perhaps the most common way that such desires are evidence against an evaluative belief provides us with our final, and perhaps decisive, objection to Smith's account. That is the simple fact that evaluative beliefs are (at least) quite often motivated! Nothing can be more familiar to us than each party to a conflict believing that it is in the right, that it deserves more money, more respect, etc. People believe that they should kill or maim others for revenge, that they should punish themselves for having sexual thoughts, that they should kill civilians if ordered to.

I'm not saying that if these beliefs are motivated then they're not true. I'm saying that so very many of our evaluative beliefs are the result of rationalization of what we antecedently want, and that very very many of them are not true. As far as I know the majority of evaluative beliefs ever held are not true.[255] The fact that one is strongly motivated to reach a particular conclusion doesn't mean it's not true, but it does mean that one should be particularly careful to check the reasoning and facts that one is using (if one is using any at all) to support it. If desires often play a role in the formation of our evaluative beliefs, then there is no reason to suppose that contrary desires can't be evidence against those beliefs.

Now I'm happy to admit that the claim that evaluative beliefs are commitment devices has not been established beyond a reasonable doubt. But there cannot be any reasonable doubt that very many (I say all, but not in way that's inherently problematic) evaluative beliefs are motivated. But if this is right, the sheer fact of having an evaluative belief (a belief that one has a normative reason) provides *no prima facie reason at all* to suppose that 'reason' (in a rationalist sense) is on its side and against a competing first-order desire. If all we know about someone is that they believe that they have a normative reason to Φ but a (n especially strong and stable) first-order desire not to Φ, I don't think we know enough to hazard a guess as to what they ought to do.[256]

[255] That is, even if we ignore the possibility of an error theory.

[256] Just to be clear, I think that *in general* it makes sense to treat our evaluative beliefs as more normative than our first-order desires, just not for the reasons normally given. In fact, it's not clear that we can do otherwise. For should one come to think that one's first-order desire was good enough evidence against one's evaluative belief that one no longer treated that evaluative belief as normative, it seems that one would have (at least) suspended that belief for the time. To have an evaluative belief then just is to treat it as normative.

Of course Smith only argues that they ought to desire to Φ if they believe they have a normative reason to Φ. That is, they will have such a desire just insofar as they are rational. But this is easily captured on my hypothesized connections between evaluative belief, rationality and the will. To have an evaluative belief that one should Φ indicates that one has a commitment to Φ. It might be a very weak commitment; but Smith doesn't given us any requirement on the strength of the desire that a rational person will have upon forming such a belief. Which is as it should be. At this level of generality, covering all possible normative reasons, it would be silly to specify what strength of desire a rational person would have. But I think it's safe to say that if someone thought they had a very strong normative reason to Φ, then they would not necessarily have a strong 'desire' (in the 'feel attracted to the idea of doing it' sense) to Φ, but they would presumably have a fairly strong *commitment* to Φ, backed-up *at a minimum* by the threat of (the feeling of) weakness of will should one not act accordingly. As when Huck says, 'Well, I just felt sick [not attracted to the idea]. But I says, I *got* to do it – I can't get *out* of it' [nevertheless committed to it].

On the end-relational view, all evaluative beliefs are motivated, but where 'motivated' doesn't take on the implications of irrationality that normally accompany it, which implications only make sense against a background notion that truly rational judgments take place independently of our motivations. We've seen that there is an inherent tradeoff between the benefits of acting on principle and the costs of rigidity. Sometimes it would be best to relax the rules a bit, despite one's initial belief that they should not be. Cases in which one has a strong desire to violate the rules *can be* evidence that those cases would be a good time to do so. We can see this not only in 'moral' cases like Huck Finn, but it is an ubiquitous feature of our lives. That desires can be evidential in this way is readily acknowledged and makes easy sense within a view that regards all evaluative beliefs as motivated, but it is invisible and nonsensical on Smith's rationalist view.

3.4.3 End--relationalism vs. Scanlon's and Wallace's Rationalisms

Scanlon Against Desires as Motivational or Justificatory

Scanlon (1998) is closer than Smith to identifying weakness of will with practical irrationality, and argues against desires both as a source of reasons *and* (unlike Smith) as a source of motivation. It should be noted right away that Scanlon is operating on a 'directed-attention' sense of desire, a conception that I grant is closer in some ways to the everyday meaning of the term. On this conception, someone has a desire 'that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favor of P' (39). Therefore some of his challenges to desires being the source of reasons or justification are simply not relevant to the direction-of-fit, 'pro-attitude' conception of desire I'm using.

Nevertheless, many desires on the direction-of-fit model are also desires on the directed-attention model, and so there are important points of contention. I want to use these disagreements both to criticize his view of (ir)rationality, and to illustrate how the end-relational conception of reasons I'm offering is not vulnerable to his arguments against desires as motivations or as justifying reasons. I'll start with Scanlon on motivation, then turn to justification.

According to Scanlon, desires are not a 'special source of motivation, independent of our seeing things as reasons' (40). Two general considerations support this conclusion. First, we are often motivated to do something when we have no desire to do it. Second, even when we do have a desire to do something, and we act on it, the motivation is provided by 'the agent's perception of some consideration as a reason, not some additional element of "desire"' (40-41).

The first critique is not relevant to us, for we have addressed this at quite some length, using a different conception of desire. The second we can address as follows. It's true that when we desire something in the directed-attention sense, we generally see a 'consideration in favor' of it. But the perceptions of considerations in favor of actions are themselves (by hypothesis) from motivated perspectives. Look back to Elster's scenario. He wants to eat cream cake, but wants to not want to eat cream cake, but also wants to not want to not want to eat cream cake. Each of these are motivated perspectives, each seeming to provide a consideration in favor of action when attention is directed toward (or from) that perspective.

The idea that the *motivation* to eat a piece of cream cake is typically provided not by the desire to eat it, but by the perception on the part of the agent of a reason to eat the cream cake strikes me as so implausible[257] that I have a hard time taking it seriously. However, we can see how such a mistake would be made, in part, by Scanlon's notion

[257] Not least because it seems to imply a fundamental rupture between the motivational systems of people and animals—unless they too are motivated by the perception of reasons rather than the having of desires.

that desire constitutes an 'additional element' over and above the perception of reasons for action.[258]

But if we accept that desires are not to be thought of as some 'additional elements' of motivational force, but rather the multitudinous and various goal-directed states from the 'point of view' of which anything can be a reason for action in the first place, then we do not have to imagine the desiderative view as one whereon people may feel there to be considerations in favor of any of a variety of actions, with or without any associated desire, and then are only motivated to act on the considerations that contingently have some desire 'added' to those considerations. Rather, we have a desiderative view in which the desire to eat the the cream cake, and the desire to not desire that, and the desire not to desire *that*, all come with perceived considerations in favor of their respective actions, especially when the rewards associated with that perspective have our attention.

Scanlon of course not only denies that desires are typically the source of motivation for action, but also that they 'provide' justification for actions. In illustration, Scanlon offers the example of his recurring desires to buy new computers. As he describes it, he finds himself 'looking eagerly at the computer advertisements in each Tuesday's *New York Times*' (43). He takes the features of these computers to count in favor of buying the computers, and keeps thinking about buying one of them. He takes it that being in this state of desire, which although it has normative content (since he takes there to be reason to buy them), does not provide him with any reason to buy one. This is

[258] Scanlon is not at all alone in this way of thinking of the relationship between reason and desire.

straightforwardly cashed-out in terms of his considered judgment that these features would not be of any genuine benefit to him. It is important to see that Scanlon is not saying that the desire to buy a new machine is outweighed by other considerations or desires, but rather that his desire provides him with no reason at all.

Scanlon characterizes the idea that his desire provides him with a reason to buy the computer as the idea that he has a reason to do so 'because doing this would satisfy [his] desire' (43). We have already seen that this is the wrong way to think about it. This is to still conceptualize practical reason as proceeding from a standpoint independent of all desires, yet judging that it follows from the fact that something is a desire that one has a reason to act on it. But 'that it would satisfy my desire' is never, or maybe only rarely, a reason.

Scanlon recognizes this, but draws the wrong lesson from it. For he says that if he *were* to form the considered judgment that buying a computer were warranted, then that would not be because it would satisfy his desire, but rather because he would enjoy it, or it would help him with his work, or bring some benefit to him (44). He recognizes that this conclusion could easily be accepted by a desire-theorist, who claims that one's reasons depend on one's desires, without its being the case that 'because it will satisfy my desire' is itself a reason. For example, in the case of the computer, the desire-theorist would claim perhaps that it is his desire for whatever benefits the computer would bring him upon which his reasons for buying one depend. Scanlon therefore undertakes to analyze, and debunk, cases where this dependence seems to hold.

The first two types of case Scanlon addresses depend on the directed-attention sense of desire, and I won't address them here since we are not using that conception.

The third seems to be the most important type at any rate, according to Scanlon. This is that we take the pleasure we expect to have as a result of satisfying some desire as a reason to satisfy the desire. Suppose someone says they want to go to Chicago. Scanlon says that we interpret this as meaning not only that they think they would enjoy Chicago, or that they are longing to go, but rather that they take themselves to have reason to go. That is, they see the enjoyment they would have as a reason to take the trip, not only in the directed-attention sense discussed earlier, but at the level of a judgment that the pleasure to be had by going really would be a good reason to go. Scanlon says that this is indeed a state that can give someone reasons for action, but that it is misleading to call this state 'desire'. Rather, it is a combination of the agent's having identified some consideration(s) in favor of action and their 'decision to take them as grounds for action' (45).

The considerations in favor of action in this and the other cases Scanlon discusses have to do with the pleasure that someone expects to receive from some activity. If any consideration for action is made from a motivated perspective, surely it is the consideration that something would be pleasurable for the considering agent. Nothing could be more congenial to the end-relational view I'm espousing than that people take anticipated pleasures to count as reasons for action. Of course they do.[259] *Pace* Scanlon, it is not misleading to say that the person's reason for going to Chicago depends on her desire for the pleasure she expects to get from going there, combined with the claim that the consideration she sees in favor of that pleasure is *relative to the desire for that*

[259] Not always of course. There are sometimes conflicting commitments and norms, often moralizing ones, that deny that pleasure is any reason to act in certain cases.

pleasure, or pleasure in general.[260] She has a reason to go, at least in part, *in order to experience the pleasure* that she expects to enjoy. That she 'really thinks' it is a reason to go is plausibly a consequence of the fact that when she reflects, or imagines taking the trip, she predicts that she really will enjoy it, and that there are no strong norms or other commitments against taking the trip under the circumstances, or against taking one's pleasure in doing so to be a reason.[261]

I said just now that part of what would cause someone to think that there really is a reason to go to Chicago is that when they imagine going there, these imaginings or reflections act as predictions that they really will enjoy themselves. We can have many directed-attention desires that seem to offer enticing rewards, but when we imagine the scenario playing out, we sometimes predict not only that those rewards will be offset by greater harms, but that the rewards are themselves illusory. We judge that we will not in fact enjoy what had struck us, and may continue to strike us recurringly, as something that we would enjoy. This is happening in Scanlon's computer-advertisement example. The features of the computers seem attractive. They are shiny and new. It seems as if it would be nice and pleasant to have them. But really it would not be nice to have them, were one to actually buy the computer. And further, having recognized the uselessness of

[260] Again, that we have norms allowing for pleasure to be a reason is important. If we were in a culture that regarded one's own pleasure as counting for nothing, then to the extent that one accepted this norm, there would be a commitment against, and a consequent destabilizing effect upon, one's taking her own pleasure to count as a reason for action.

[261] I understand the temptation to ask why it is not the expected pleasure itself, but the desire for that pleasure, that gives the person a reason to go to Chicago. But I also understand that the way we conceive and talk about practical reasoning presupposes that there is an executive reasoner apart from our desires that is provided with reasons for action. And looking at it this way, it seems strange that the fact that we desire the pleasure should be what gives us a reason to pursue it, as opposed to the fact that we will get pleasure. For we could think that there would be something wrong with us if we didn't desire (certain kinds of) pleasure. But I want to say that reasons are not 'provided' to an extra-desiderative rational self, but that reasons are all from motivated perspectives, some but not others of which are identified (to greater or lesser extents) with the self.

the features has a strong tendency to eliminate, not just offset, whatever enjoyment one might have otherwise gotten out of them.

If somehow Scanlon were to buy that computer after judging that it would be of no real benefit to him, he plausibly would not be able to receive any enjoyment from it because he would feel like a fool for paying many hundreds of dollars for a computer that he recognized would not be of any real benefit to him. He would feel some combination of regret, guilt, shame, weakness, and/or some such things. These feelings do not simply outweigh the little pleasures of having a nice new machine, but can often preclude them. Any time Scanlon were 'tempted' to feel good about having some new bell or whistle, the feeling of guilty foolishness for having wasted so much money on it would likely squash that feeling and turn it into fodder for those negative emotions, which would plausibly persist beyond the range of potential excitement over the new machine. Here conflicting desires do not only 'outweigh', but actively undermine the potential rewards associated with others.

But now suppose that Scanlon's current computer were to break down and was under warranty, and the manufacturer offered Scanlon the choice between one of the new ones that Scanlon had felt a desire for, or to repair his current one so that it is as good as new. Imagine that everything can be very easily transferred from the old one to the new. Now does Scanlon have a reason to get the new one? It seems perfectly plausible to say that he does, whereas I agree that it is implausible to say that he had a reason in the case as he described it. What is the explanation for this? Most obviously, there is no conflict now. And since there is no conflict, whatever little pleasures are to be had from having a nice, brand new computer with a couple of new little toys, are free for the having. These pleasures may not amount to much, which explains the sense that there isn't any very strong reason to take the new one. However, *now* there is a reason to get it because *now* one can imagine experiencing some enjoyment from having the new computer.

In the original case, one could not imagine getting any pleasure because one has the resources to imaginatively predict that one will in fact receive no pleasure in the case that one actually buys the computer, keeping one's judgments constant. If Scanlon were trading long-term gain for short-term pleasure, then I think we would and should reject the idea that he has no reason to buy the computer; rather, we should say that those reasons are easily outweighed by the reasons associated with his long-term interests/desires. But our intuitions allow us to see that there will be no such trade.

'Really is a reason' amounts to 'really will give me something I'm motivated to get (or avoid what I am motivated to avoid)'. Sometimes when we reflect or imagine, we can predict that we won't get some imagined reward based on features of the situation or ourselves that have nothing to do with having conflicting desires (commitments), and other times we can predict that those very commitments will preclude the rewards that would exist were it not for the contrary commitments and their emotional backing.

Scanlon has done nothing to make us doubt that whatever justification there is for Scanlon to buy a computer depends on that computer being able to do something for him that he wants. One may say that the reasons to buy it depend not on desire, but on the benefits it will bring. But if one were strange enough to be motivationally indifferent to being benefited (if this even makes sense) then there's no reason I can see to insist that one would yet have a reason to buy the computer. The reasons to do what will benefit oneself, like any reasons, are made from motivated perspectives. The intuitive lack of a reason to buy the computer in Scanlon's formulation, the presence of one in my formulation, as well as the presence of a reason to go to Chicago given that one predicts that one will enjoy oneself, are all easily cashed out in terms of the respective desires in play, and not in play.

Wallace's Rationalist Response to the Teleological Argument

I want to address one final rationalist proposal. In doing so, I will make good on the promise I made in section 3.1.2. After describing Smith's teleological argument in favor of Humeanism about motivating reasons, I noted that that argument did not actually cinch the case for the Humean, or perhaps even support him. For Wallace, as I said above, rightly characterizes the 'central point between the Humean and the rationalist … [as concerning] the extent to which rational processes of thought—those which are governed by rational principles or norms—can contribute to the explanation of motivation' (2006, p. 20, n.12). The teleological argument's conclusion is that a state of belief alone is not sufficient to motivate; there must be some state of desiring 'in addition' to it. But in order for this to support the Humean, we must add the premise that the necessity of invoking desires to explain motivation rules out the explanation of motivation in terms of distinctively rational principles (22). Here is where Wallace wants to challenge the Humean.

In doing so, he draws on Thomas Nagel's (1970) distinction between motivated and unmotivated desires. Nagel left this distinction obscure, characterizing unmotivated desires as those which simply 'assail' us (like certain appetites and emotions), and motivated ones as those ‘*arrived at* by decision and after deliberation’ (Nagel 1970, p. 29, quoted on 22-3 of Wallace 2006). Wallace correctly notes that this cannot be the exhaustive typology that Nagel seems to have intended. For he has left out the crucially important class of long-term desires that are formed, at least in part, by one’s ‘moral upbringing’. These are the sorts of desires the possession and strength of which make for characterological assessment. But they neither assail us nor are they the result of decisions or deliberations (23).

Despite the apparent obscurity of the distinction as Nagel makes it, Wallace finds in it the seeds of a promising challenge to the Humean. The strategy will be to understand motivated desires as those which are always explicable by showing how they are ‘*rationalized* by other propositional attitudes that the person has’ (23, italics his). So the explanation of motivated desires, unlike that of unmotivated desires, will not be restricted to causal accounts involving psychological states that ‘trigger’ the emergence of a desire. Instead, motivated desires necessarily allow for explanation in terms of their propositional contents being rationalized or justified by the propositional contents of the agent’s other attitudes (24).[262]

It is not enough that the person’s desire *can be rationalized* by her other attitudes, but rather it must be the case that the person has formed and holds the desires *because* of these rationalizing attitudes. Otherwise these attitudes will not provide a rationalizing *explanation* for the desire. For example, if I have a motivated desire (in Wallace’s sense) to get up and go to the kitchen to get some food, then this desire can be explained and rationalized in terms of my desire to get some food plus my beliefs that there is food in

[262] This is not to deny that these rationalizing explanations can also be causal explanations.

the kitchen, that I can get there by myself, etc. It is (by hypothesis) because I have these rationalizing attitudes that I form the desire to get up and go to the kitchen, and these attitudes provide reasons for my motivated desire.

Wallace recognizes that the propositional contents of desires, such as 'that I get up and go get something to eat' do not seem to be suited to either rationalize or be rationalized by other propositions, if for no other reason than that such contents are not even in the indicative mood. But to see how desires can play a role in rationalizing explanations, Wallace draws upon the characteristic association between desires and evaluative beliefs. He notes that, aside from abnormal cases in which one desires what one values not at all, desires come with associated evaluative beliefs. So normally, if one knows that I want to get up to get some food from the kitchen, one can attribute to me the evaluative belief that doing so is desirable. And now this evaluative belief can be rationalized via the schema of a practical syllogism to the effect that I have an evaluative belief that eating is (prima facie) desirable and the belief that in order to eat it is 'necessary' that I get up and go to the kitchen (25). These beliefs justify the belief that getting up and going to the kitchen to get some food is desirable, and that belief might have been reached because of the rationalizing beliefs.

If all this is right, then we can distinguish between motivated and unmotivated desires in terms of the former's ability to enter into rationalizing explanations. Motivated desires have associated evaluative beliefs that allow for a rationalizing explanation. And importantly, when this is true, the reasons which justify the evaluative belief also justify the associated desire. The reasons for holding the belief are reasons for forming the desire. Now if people at least sometimes form desires because they accept the truth of the evaluative belief on the basis of the reasons that directly support it, then we will have explained these desires, and hence motivations, in terms of rationally justifiable (or not) beliefs. And if this can be maintainted, Smith's teleological argument as given is consistent with some motivated desires being explained purely in terms of evaluative beliefs, in addition to rational principles, and in these cases the motivations and asociated desires can be given a purely rational explanation (26).

But even if there were no other problems with this story, there is one further assumption upon which the entire strategy depends. As Wallace points out, in order for us to say that we are giving a rationalizing explanation for a desire, it must be the case that the desire formed *because* the person has endorsed the evaluative belief. And this seems to require the 'crucial assumption [that] the rational explanation for an evaluative belief may account for the formation of the motivated desire as well*, so that the reasons which explain the belief will equally be reasons for the motivated desire*. To say this, it seems, is to admit that *it is an independent principle or norm of rationality that one should desire in accordance with one'e evaluative beliefs*' (26, emphasis mine).

Wallace finds this a 'plausible minimal assumption to make about the content of the principles or norms of practical reason, since we do in fact try to adjust our desires to our evaluative beliefs, and take ourselves to be subject to rational criticism when we fail' (26). But Wallace recognizes that Hume (and presumably (neo-) Humeans—this one at least) will deny this assumption. However, Wallace believes Nagel and Korsgaard[263] to have shown that Hume appears to assume some irreducible principle of instrumental reasoning. For example, Hume claims that a person who no longer believes that some

[263] Nagel (1970, 33-4); Korsgaard (1986, sect. III).

course of action will lead to a desired end will immediately cease to desire that course of action (if that were the only reason for the desire). Wallace says that this fact about people is best understood as the dual claims that it is rational to conform one's desires to one's beliefs about what is instrumentally rational and that people are (generally? always?) rational in this respect. Therefore, Wallace concludes, Hume is 'poorly placed to reject out of hand the principle that rational agents adjust their desires to their evaluative beliefs' (27).

This assumption of an independent principle of rationality is absolutely crucial to Wallace's entire rationalist proposal in this essay, of which fact he is forthright enough to remind the reader repeatedly. But it is highly problematic, as we've already seen. First, and least importantly, Wallace characterizes the irreducible principle of rationality to which Hume is putatively committed as the 'principle that one should adjust one's desires to one's evaluative beliefs' (27). But this doesn't seem correct. If Hume is committed to some such principle, it is that one should adjust one's desires to one's beliefs about the means to one's own ends. These are not evaluative beliefs, but means-ends beliefs. But nevermind that. Perhaps the fact, if it is one, that Hume is committed to some kind of irreducible practical principle, namely the instrumental principle, leaves him 'poorly placed' to reject others out of hand.

But Hume is not so committed, and more importantly, there is no reason for a neo-Humean to be so committed. There is no categorical requirement of instrumental reason. Finlay (2008) has argued persuasively for this conclusion where 'ends' are understood as intentions, and I argued above for why such a requirement makes no sense if ends are understood as desires more broadly. Those a priori arguments were not meant to be sufficient in themselves, but combined with an analysis of the *relational* nature of judgments of rationality, we have no reason to regard failures to take the means to one's ends as *intrinsically* irrational. They are often, or perhaps even always irrational, but they are so relative to the motivated perspective(s) associated with our judgments of rationality.

I just said that even if it were always the case that not taking the means to one's ends were irrational, it would not be intrinsically irrational. But not taking the means to one's ends, however that gets cashed out, is not the same thing as not adjusting one's desires to one's evaluative beliefs. So even if we were inclined to grant the irreducible irrationality of the former, we would still have good reason for denying it as a free-standing, irreducible principle of rationality. And that is because the ubiquity of Huck-Finn type cases shows that a willpower conception of rationality, though attractive to many rationalists, is not tenable on rationalist conceptions of practical reason.[264] Returning to Scanlon for a moment can help us see this more clearly.

Scanlon combines his motivational and justificatory critique of desiderative views in claiming that when our desires

> … are 'contrary to our reason (that is to say, our judgment)…the motivational force of these states lies in a tendency to see some consideration as a *reason.* Akratic actions (and irrational thoughts) are cases in which a person's rational capacities have malfunctioned, not cases in which these capacities are overmastered by something else, called 'desire' (40, his emphasis).

[264] This is especially true if we consider practical reason the ultimate currency of normativity.

Is this true? Did Huck's rational capacities malfunction both when he saw Jim's trust and friendship as a reason (if he did) for not giving him away, and when he in fact did not give him away? In what sense exactly? Wallace says that the independent principle enjoins us to 'desire those ends and activities one takes to be desirable, to the extent one takes them to be desirable' (26). Did Huck violate an irreducible rational principle in not desiring to turn Jim in to the extent that he took it to be desirable? Do all the lovers and givers of mercy violate such a principle whenever they 'fail' to adjust their desires to their evaluative beliefs, which may well be the result of any manner or intensity of propaganda?

I don't know Scanlon's or Wallace's answer to these questions, or that they have considered that they ought to have an answer to questions like these. One common answer is that Huck exhibited a *form* of irrationality, call it 'executive rationality', which just is to 'fail to' (but this is perhaps to prejudice the case; we should simply say 'not') act on one's own beliefs about what one should do. Perhaps there are other forms that can conflict with this form, but it could be said that this is a form of irrationality, undergirded (in part at least) by the independent principle.

What I want to know is what exactly this 'failure' has to do with *reason*. Why is this mismatch rightly described as a form of *irrationality*? I have offered an answer. And that answer lies in the explanation for the fact that we do not tend to see these as simply mismatches, but rather as failures. Even if it's a 'lucky' or fortunate failure, it is a failure. Something was willed, but that thing didn't happen. The sense of trying or attempting is so familiar that nothing is more naturally described as a failure, and the source of the failure as a form of weakness. It's telling that Korsgaard's arguments to the effect that we can fail to do what is instrumentally rational by our own lights made essential reference to this and other forms of weakness. Practical rationality, we see over and over again, is opposed to this ‘weakness’.

The reason the willpower conception of rationality is so attractive is that executive ‘rationality’ just is executive *power*. Having made an evaluative judgment that there is overwhelmingly good reason to Φ, if one does not form and act on an intention to Φ, one will have the sense that the ‘rational executive’ is weak relative to one’s contrary desires, all of which it transcends in making such judgments. It is tempting to think that the will and rationality are connected in this intimate way because our evaluative judgments are based on reasons, or are at least reason-responsive. To the extent that these judgments are the product of pure practical reason, and they are nevertheless not carried out, then we can say that that failure is straightforwardly irrational.

But even if there were *some* cases in which pure practical reason issued in practical judgments, surely the overwhelming majority of cases are not like this. The judgments we come to in practical reasoning are very often influenced by our desires and commitments, where there is no reason for thinking that these desires and commitments are themselves the product of pure practical reason. We should remind ourselves from time to time that people are capable of evaluative beliefs of such extraordinary variety and extremity of ludicrousness that they would surely not be thought possible were they not actual. The more we keep this in mind, the more bizarre it will strike us that there would be ‘an independent principle of rationality’ that enjoins our desires to follow our evaluative beliefs as such.

But the explanation for this bizarre belief is the fact that judgments of rationality are typically made from the motivated perspective associated with the conscious experience of willing, and the essential nature of the will is to attempt to strengthen itself. Again, this is because the nature of the will is to control our future actions. To the extent that it is successful, it becomes more likely to be successful in the future, more powerful so to speak, so that by the nature of its operation it 'attempts' to become more powerful. Our conscious experience of ourselves as agents is largely constituted by conceiving of ourselves as extending into the future (and the past), and projecting our current will and intentions into that future. It is also tightly bound up with the conscious reasoning process, the cognitive activities that are both consciously available and, to a large extent, publicly statable and shared. We (especially rationalists!) identify ourselves with our 'rational' selves in this sense, though there is very much (more) cognition taking place outside conscious awareness.

This rough sketch of an explanation for why a willpower conception of rationality would be attractive to rationalists, who after all have a burden of showing how rational considerations necessarily motivate (rational) agents, need not be believed in order to reject the putative independent principle of rationality. However, it is a strength of my competing view to be able to offer a non ad-hoc explanation of it. In fact, finding the best explanation of the attraction and crucial argumentative appeal to various versions of the willpower conception of rationality common to rationalists is plausibly the best way to adjudicate between the neo-Humean and these rationalists.

If the attraction to this conception, i.e., if the belief in such independent principles of rationality is best explained by some desiderative account as I have offered, then there will seem to be no question-begging reason to suppose that there are truly such independent principles. That is, there will be no reason in the absence of an argument that does not appeal to the intuitions that are themselves explained in my or some other neo-Humean terms. So far, to my knowledge, there is no such argument, and worse, there are quite serious Huck-Finn-style counterexamples, complete with alternative, anti-rationalist accounts of the sense in which Huck was (ir)rational. And worst of all, neither Smith nor Scanlon nor Wallace see that there is even any potential objection here.[265] I'm not saying that they certainly could not come up with an answer to these objections. However, as it stands, my end-relational view has a good answer to the question of how to reconcile our 'willpower'-related judgments of (ir)rationality with competing judgments in Huck Finn-type cases. The rationalist authors I've canvassed not only don't seem to have any such answer, they don't seem to be aware the question exists, though it strikes at the heart of all their proposals.

The Desire-Out, Desire-In Principle

To have rejected Wallace's independent principle of rationality is not to have directly defended the neo-Humean view. In order to do that, we would have to defend what Wallace calls the 'desire-out, desire-in principle'. This principle is precisely contrary to Wallace's proposal, which is to try to show that one could rationally explain a desire without having to make (ultimate) reference to another desire. On his proposal,

[265] Brink sees that there would have to be restrictions on the content of (rational) products of deliberation, but as I said, I don't think these restrictions help show that deliberation is *intrinsically* rational.

desires could be explained purely in terms of evaluative beliefs and principles or norms of rationality. By contrast, the desire-out, desire-in principle says that any rationalizing explanation of a desire will ultimately bottom out in an intrinsic desire for which there can be given no rationalizing explanation. To secure the neo-Humean view of practical reason, Wallace says we need to have an argument for this thesis.

My argument for this thesis just is the end-relational view I've been defending in this chapter. On that view, reasons are always provided from a motivated perspective. My view entails this desire-out, desire-in principle, and I submit that (something like) this view is likely to be correct. I'll quickly summarize its strengths. It makes for a more plausible Humeanism insofar as it 1) shows why acting on certain desires can seem to be, but never is, intrinsically (ir)rational, 2) jibes very well with an independently motivated and compelling model of the will 3) and with many other models of the self as 'collective', which is in keeping with Hume's wish to avoid positing a 'self within the self', while still 4) explaining the attractions of the internalist requirement. Further, it 5) explains what's right about Korsgaard's intuitions about the will and rationality, while avoiding implausible implications about (ir)rational actions, 6) explains the ambivalence of our judgments about Huck's rationality (and indefinitely many other similar cases), and 7) explains what's both right and wrong about the willpower conceptions of rationality shared by several prominent rationalists.

The willpower conception is closely related to, and perhaps dependent on, the idea that desires can't be evidence against evaluative beliefs. But we have both good reasons to deny this claim, and a good explanation for why it would be held, from within the end-relational view. That is that evaluative beliefs are in the business of committing us to some particular motivated perspective, normally one that would otherwise be vulnerable to being undermined by competing desires. Crucial to the operation of this technology is that mismatches between the evaluative beliefs and competing desires are not to be investigated to determine how the tension should be rationally resolved. It is part of the nature of the committing technology that we feel that the tension should be resolved in the direction of accommodating the desire to the belief. However, when we recognize that very many evaluative beliefs are not in any robust sense the products of reasoning, but are rather very often simply culturally inherited or are the rationalizations of antecedent desires, we have even more reason to reject the notion that this direction of accommodation is to be explained or understood as an 'independent principle of rationality'. Instead, it seems to be (at least partially, and better) explained by the fact that if we have recourse to the thought that desires can be such evidence, then this thought is likely to be seditious in important ways, especially when temptations are powerful. So rather than evaluate felt desires to see what sort of evidence they might contain against our evaluative beliefs, 'we' (rationalists at least) hold that they simply cannot be such evidence.

There is one last piece of business. My first claimed advantage over rationalism above was that my view explains why certain desires could seem to be, but are not, intrinsically irrational. It is at the heart of my view, and all neo-Humean views as I understand them, to deny that desires can be intrinsically irrational. This is important because even if the desire out, desire in principle is true, that would leave open the possibility that some of our desires are intrinsically irrational. From the fact that all motivated behavior bottoms out in desire(s) does not entail that this behavior cannot be intrinsically irrational.[266] And I have not directly engaged arguments to the effect that some desires would be intrinsically irrational (if anyone actually had them).

This is the last challenge I will meet. I will in fact admit that, in a sense, desires can be intrinsically irrational, but the sense in which this is true does not undermine, but rather vindicates the claim that all judgments of (ir)rationality are from motivated perspectives.

Intrinsically irrational desires

Let's examine again the nature of the irrationality in paradigmatic cases such as smoking or drinking where one knows better and has resolved to do good. It seems that the irrationality of acting on these desires is much better explained with respect to one's other, long-term desires than in terms of their being intrinsically irrational to act on. For a smoker or alcoholic or other addict, it can be quite uncomfortable not to indulge. The nearer the opportunity, the more insistent one's attention is drawn to the reward, the greater the anxiety involved in resisting it. This anxiety can build recursively—anxiety predicts violation of resolve, which generates more anxiety, which makes violation seem more probable, and so on.

The desire to relieve anxiety and uncomfortability is not intrinsically irrational. The desire for pleasure is not intrinsically irrational. But in specific cases they are irrational with respect to the the long(er)-term goals one has, which, it shouldn't be

[266] Combining these views might threaten at least a limited version of practical nihilism. If there are intrisically irrational desires that motivate some of our behaviors, then perhaps there are things that we do that are intrinsically irrational, but which our rational capacities are also impotent to correct.

forgotten, are not simply ‘from a different perspective’, but also promise more and richer satisfactions. When an addicted smoker wants a cigarette, we might imagine this desire appraised by our rational agency as to whether or not such a desire constitutes any reason to smoke. We might be inclined to answer that it does not. If we do so, we will be answering from the perspective of our long(er)-term ends, and from this perspective, there is nothing to be said for smoking. Saying that we have no reason can be thought of as benign propaganda on the part of this perspective. But relative to (that is, in order to achieve) the end which we inherently have of ridding ourselves of anxiety, or of ceasing to have our attention distracted by cravings, or for a bit of pleasure, we ought to smoke. But these short-term ends are often subordinated to our wills, which are paradigmatically concerned with our longer-term goals.

One might admit that the desires I’ve covered here aren’t intrinsically irrational, but claim that others are. Parfit (1984) canvasses several putative examples. One of them is Future Tuesday Indifference. Someone with this pattern of concern cares normally about what happens on present Tuesdays and all other days, but does not care now about what happens to him on future Tuesdays. He has no false beliefs about personal identity or time or Tuesdays or any other relevant thing. He just doesn’t care what happens to him on future Tuesdays. This fact will often lead to great regret once those future Tuesdays become present Tuesdays, since he might have committed himself to having a much more painful experience on what has now become a present Tuesday rather than a trivially painful experience the next day, but this does not affect his continued indifference toward future Tuesdays. When asked for his reasons for his indifference toward future Tuesdays, his reason is that they are (future) Tuesdays.

Parfit plausibly insists that [*t*]*hi*s *is no reason* (124, emphasis his). That seems right. What *could* be reasons for undergoing worse rather than lesser pains? Parfit lists the following. Suffering more over less might not be irrational if one thinks one ought to suffer penance, or to make oneself tougher so as to better endure future pains, or to maintain a resolve only so as to strengthen the will, or even to bring wisdom (123).

All these have obvious long-term benefits and/or comprise long-term goals or commitments—even penance, which presumably involves some commitment to punish oneself for certain kinds of violations. This may or may not be conceived of as having the function of reducing the offending behavior in the future, but especially when it is not so conceived, there is an emotion-backed commitment to do so.

Future Tuesday indifference will lead to considerable amounts of regret. And importantly, there is nothing to be said for it. What does this mean, other than that there is no reason for it? According to my account of what reasons are, the (explanatory) reason we say there is no reason for it is that *we have no motivated perspective from which we can identify with this pattern of concern*. We cannot imaginatively take on this person's perspective in a way that allows us to empathize whatsoever with this person's motivations. The intuition behind motivational internalism is that reasons have to have the capacity to motivate, and this consideration, from our point of view, does not even slightly allow us to be or even imagine being motivated by it.

If the indifference gave the person pleasure of some kind in the present or future, or anything that we could imagine motivating us, that would be different, but it does not. It is however radically at odds with both this person's long-term desires and with our own, should we have it. Therefore it appears intrinsically irrational. There is no perspective we can take up from which such a consideration can move us *at all*. Parfit says that '[i]t is [intrinsically] irrational to desire something that is in no respect worth desiring. It is even more [intrinsically] irrational to desire something that is worth *not* desiring—worth avoiding' (122). If I'm right, we can understand 'worth desiring' as a judgment made from at least *some* motivated perspective, though typically from a longer-term perspective. And 'worth avoiding' is even more reliably made from a long-term perspective. Future Tuesday indifference fails to find any motivated perspective to endorse it, and is very strongly counterindicated by one's other, especially 'rational' perspective(s). It is in these terms that we can understand how a desire can be 'intrinsically irrational.'

This sense in which desires can be 'intrinsically irrational' is not a bullet for my view to bite, but fodder for it. For we are understanding the irrational desires that real people actually have as irrational in the sense that these desires *conflict* with others that are associated with our evaluative beliefs, long(er)-erm commitments, etc. None of these are intrinsically irrational, because their irrationality consists in their conflict with these other motivated perspectives. But a(n) (imaginary) desire that is *actually* 'intrinsically irrational' is just one that we can't see as in *conflict* with other motivations, because the 'intrisically irrational' desire can find no (imaginary) motivational purchase. *Where we cannot imagine any motivational purchase, we cannot imagine any reason.*

Again, if someone with this 'desire' received anything that we perceived as attractive, even a moment's pleasure from the satisfaction of the desire, then we can see a conflict between this pleasure and his long-term goals. We can imagine being pulled by the attraction of a moment's pleasure, and though we don't get it from Future Tuesday indifference, we can abstract away from the specific objects that give another person pleasure and empathize with the attractiveness of pleasure. But in the absence of anything like this, we can feel no conflict, *nothing in relation to which* our other motivations stand in conflict, and as such, the desire is 'intrinsically irrational'.

Just in case the reader is suspicious that this analysis is peculiar to Future Tuesday Indifference, let me show that it fits Parfit's other examples, such as '*bias towards the next year'* and '*Within-a-Mile-Altruism'*. The former involves caring equally about one's life for the upcoming year, then half as much after that. The latter involves caring greatly about those within a mile and little about those farther away. Parfit says that bias toward the next year is 'more open to rational criticism' (125) than simple bias toward the near future, and within-a-mile altruism also more open to rational criticism than caring about everyone equally or caring much more about one's own community.[267]

As evidence and justification for these claims, Parfit asks rhetorically how these things could be reasons. How could it matter (so much) that pain will be felt 53 rather than 52 weeks from now? How can someone being just over a mile away be a reason for so much less concern? Parfit does not answer these questions; they are rhetorical after all, meant to suggest that there can be no reason for this pattern of concern. But these questions don't have to be only rhetorical. One could really want to know why it can't matter so much that some pain is 53 weeks away rather than 52. The answer is that it

[267] Parfit claims, without argument, that neither of these attitudes is irrational. I find the idea of caring about everyone equally as irrational as anything under the sun. The notion that I should care equally about my brother and a random stranger is as absurd a proposition as can be mustered. In fact, it would only be possible to care about everyone equally, if it is possible at all, by radically reducing our level of concern for those we already care about significantly. Our emotional capacities are simply not equipped to care about the thousands of people killed and raped every day as if they were our children. Unless of course we were to care very much less about our children, which is undesirable if not impossible.

can't matter because of the way we're built. Specifically, our motivational systems, which are of course connected to our cognitions and representations, are not 'digital' in this way. Our concern doesn't change in large abrupt ways at discrete distances or times. There is a descriptive (probably somewhat evolutionary and perhaps derivatively cultural) psychological story to be told about why this is so, no doubt.[268] On the other hand, we *do* often seem to care more about the near future, and we generally care more about our own communities. Many of us are motivated to believe that we ought to care about everyone equally, though it's unlikely that any of us can imagine actually being that way.[269]

The view I'm offering not only accommodates the intuition that such desires are intrinsically irrational, but goes beyond assertion and rhetorical questioning to provide an answer to the question of *why* such 'reasons' cannot matter, an answer grounded in presumably natural facts about our motivational systems. This same reductive strategy can plausibly answer the question why we both can empathize with patterns of concern that (steeply and/or hyperbolically) discount the future, as well as consider this very discounting irrational. The explanation proceeds from the physiologically deep fact that we tend to be more motivated by near than far rewards, in addition to the fact that we

[268] Presumably *no animals* have motivational systems that work like this. Of course this has to be in part due to the fact that they can't represent time and distance in discrete units in the first place. But there is no reason to suppose that adding such capacities to any animal would result in an ability to empathize with these patterns of concern.

[269] It's just the idea that 'arbitrariness' can be characterized independently of what we actually care about that is partly responsible for this preposterous belief.

have, and have devised technologies for strengthening and stabilizing, our desires for longer-term (and larger) rewards.[270]

I conclude that my end-relational view is more plausible than rationalist alternatives, and it entails the desire-out, desire-in principle. If the view is correct, then I have fulfilled my promissory note, and shown that evaluative beliefs can only provide motivation as extensions of antecedent ends, where those ends include increasing one's (will) power to achieve one's ends.

3.5 Conclusion

Neo-Humeanism has a lot of theoretical and intuitive attractions. However, many rationalists have argued that it also has a lot of theoretical and intuitive costs. I take it that the attractions of rationalism are largely due to the perceived weaknesses of neo-Humeanism. I have, I think, shown that my version of neo-Humeanism can meet all of the challenges that I canvassed from the most prominent rationalists. Given the important attractions of neo-Humeanism, a version that could meet these challenges should make it the default champion over both rationalist views and versions of neo-Humeanism that cannot.

Moreover, I have argued that the willpower conceptions of rationality that Smith, Scanlon, Wallace and Korsgaard rely on are vulnerable to Huck-Finn-style objections on pain of opening up a potentially large space between normativity and rationality. That

[270] Compare what I say here to how I propose to conceive of the kind of project that Parfit and McMahon (and others) undertake in discovering what grounds rational egoistic concern. They are investigating the nature of our actual concerns. In identifying 'intrinsically irrational' desires, Parfit is noticing facts about how those concerns do and do not actually work.

none of them have seen clear to even address this sort of objection, though all their views depend on doing so, strikes me as a very serious problem for them. I argued that Korsgaard's conception of the will and its relationship to rationality is Ainsliean in important ways, and that where my view and hers differ, mine is supported by a well-developed view of the will, as well as giving more intuitive answers to particular questions, as I argued when I compared them head-to-head.

I have tried in this chapter to both address the strongest arguments against neo-Humeanism—which arguments in large part motivate the attraction to rationalism—as well as directly attack the most prominent versions of rationalism. I have diagnosed their fundamental mistake in terms of a misconstrual of the relationship(s) between practical rationality and the will. If the thrust of those arguments is correct, then there can be no rational authority for performing actions, moral or otherwise, that is independent of our (existing, nonidealized) motivations. Therefore, there is no practical clout to morality or anything else. What (not) to do about that is the subject of the remaining chapters.

Chapter 4: A Pretend Solution

4.0 Summary and Introduction

This chapter begins the third and final part of the dissertation, which is concerned with arguing against the continued employment of moral discourse and in favor of paying more and better attention to our actual concerns as a central part of ethical inquiry. These are the negative and positive faces of the project as I described them in the Introduction. Though I believe the accounts of moral judgment and practical reason I have offered up to now are of intellectual interest in their own right, for the remainder of this dissertation they may be viewed as a two-part foundation for the positive and negative faces of this normative project. This section will both summarize the most important aspects of what has come before and give the reader an overview of what I intend to argue—and not argue—from here on out. In essence, I will be laying out an error theory for moral discourse that, unlike any of the error theories so far expressed in the metaethical literature, is fundamentally practical, not truth-theoretic. That is, I will not be primarily concerned with moral claims are systematically *false*, but rather with the question of their *value* (having already addressed the question of their nature). Along the way, I will try to show how a misguided focus on the question of their truth or falsity has resulted in a serious distraction from the much more important question of their value. In the remaining chapters I'll do what I can to remedy that situation.

Chapter 3 argued that all reasons and values are relative to motivated perspectives. I said in the introduction to that chapter that I think that such a conception of reasons and values fundamentally threatens moral discourse. I also said that normally this threat is couched in terms of providing a crucial premise in an error theory about morality. The error-theoretic strategy proceeds by attempting to show that the concepts employed in moral discourse are essentially committed to some features or propositions that are not true, and therefore the entire discourse is systematically flawed.

In Chapter 2 I said that Joyce defined practical clout as both rational inescapability and authority. That is, judgments involving practical clout are understood to both apply to you and have rational authority over you independent of your desire-set.[271] Chapter 3 argued that all reasons are from motivated perspectives, i.e., relative to desires, and from that it follows that there can be no rational authority outside of any such perspective. Therefore there is no practical clout. Therefore if moral discourse is essentially committed to the existence of practical clout, then moral discourse is systematically in error. Schematically, we can construct an error-theoretic argument that goes like this.[272]

1) (Conceptual claim) Moral discourse (involving first-order claims or judgments) is essentially committed to the existence of practical clout.[273]
2) (Ontological Claim) There is no such thing as practical clout.
3) Therefore all (first-order)[274] moral discourse is essentially committed to the existence of that which does not.

[271] Joyce characterizes it as independent of desires or interests, but I will just say desires, or desire-set.

[272] This is Joyce's (2001) strategy for a moral error theory.

[273] I will continue to use 'clout' in the way I have been, but it could also be thought of as a placeholder name for whatever, perhaps unalyzable, feature(s) of moral discourse that a number of authors have attempted to pin down. If the error theorist has not adequately analyzed this feature, that is not necessarily her fault since the problem with moral discourse in the first place (so she says) is that it is deeply mysterious, like ideology in general (Cf. Hussain 2004, Joyce 2007).

[274] First-order moral judgments involve claims about what is morally right, wrong, good or bad. Second-order judgments involve claims about responsibility, praise- and blameworthinesss for having done these things.

Many philosophers accept that moral discourse is committed to (something like) practical clout, but others do not. Joyce argues for this claim at length in his (2001), largely responding to Phillippa Foot's (1972) rejection of it in (though she later changed her mind). I gave a brief argument to that effect in Chapter 2, but did not make efforts to engage with philosophers who have disagreed that it is essentially committed to practical clout, especially categorical rational authority. The reason I didn't is that although I think this claim is plausible, it is not very important for my project.

As I said in Chapter 3's introduction, I think that the sense in which (peculiarly) moral discourse is committed to practical clout is not to be best understood as a conceptual commitment, but rather as a commitment device, the operation of which relies on a lack of awareness of one's motivations, and to that extent a lack of awareness that moral judgments are made from motivated perspectives. In other words, I think that the conception of values and reasons that I put forward in the last chapter, if generally accepted, threaten moral discourse because of the motivational role that a lack of awareness of our motivations plays in peculiarly moral discourse.

Though most philosophers seem to accept the conceptual claim and reject the ontological one, I have not done much to convince someone who rejects the conceptual claim that morality is systematically in error. Nevertheless, this chapter will address arguments for moral fictionalism, which is conceived as a response to the truth of the (or any successful) moral error theory. I will argue against moral fictionalism. But one might think that until I do more to show that the error theory is correct, there is little reason for one who rejects the conceptual claim to follow my arguments regarding what (not) to do *given* the truth of the error theory. I will first address this concern by saying more about why I am not (primarily) interested in an error theory as such.

While I think that it is plausible that (first-order) moral judgments are systematically and irretrievably false, I am not really convinced, and more importantly, I don't think it matters, at least not nearly as much as Joyce and other error-theorists think it does. Most if not all of what I take to be the important claims in this essay will be little- or not-at-all affected by the question whether the conceptual claim is true.

This is a strength for my project for several reasons, perhaps the least of which is that the conceptual question might be indeterminate or undecidable. There is not any philosophical consensus on how to decide the question whether some discourse is committed to some feature or concept. Joyce thinks that the answer is determined by what the relevant population of users of the discourse would decide if forced to make a choice (2007, p. 10; also in 2001, 2006). If this is right, then an assertion to the effect that a discourse is committed to some feature is in effect a prediction of what the population of competent users of the discourse would collectively decide. I agree with him when he says that whatever they were to decide, there is no a priori reason to suppose that they must be consulting some 'hidden principle' in order to do so. Their decision might vary according to circumstances, even perhaps trivial differences in circumstances, such as the presence of 'advertising jingles' (2007, p. 10).

If this is indeed the right way to think about what determines essential features of a discourse, then there might be no interesting matter of fact about what any discourse is essentially committed to. And even if there were, conflicting claims about what is or isn't essential to a discourse amount to predictions about practically untestable counterfactuals. This should make one despair of establishing a moral error theory, due to the in-principle or in-practice impossibility of settling the question whether the discourse is essentially committed to practical clout, or any feature that could be called seriously into question. Finally, given that people are seemingly committed to keeping their moral discourse, they will be motivated to judge that it is not essentially committed to the existence of any feature that they have been given reason to doubt.

In the face of these grim prospects, Joyce suggests that we have a 'decent chance of getting at the answer' if we investigate the uses to which we put the discourse (2006, p. 201). That is, what is this discourse *in the business of*? We can ask what it is used for, what practices it supports, and then see if we can judge whether it could continue to stay (at least roughly) in business as before if the offending feature were removed. In the terms introduced in Chapter 2, we would be inquiring into the instrumental function(s) of moral discourse. For example, our pre-relativistic motion judgments had the function of describing what moved relative to what, even if we conceived of them as describing what moved absolutely. When we discovered that all motion is relative to an inertial frame, this did not affect the ability of our typical motion-discourse to keep doing its job just as well as before. On the contrary, when we discovered that there are no women with supernatural powers, the purposes to which we had been putting our witch-discourse were fatally undermined.[275]

[275] Note that this remains true even if it is the case that there were certain natural properties that were instantiated by all and only those women convicted of being witches (such as, e.g., 'women feared by the authorities'). This is an important point to be kept in mind with respect to attempts to find natural properties that all and only the things we call (im)moral have in common. Even if we were able to find such properties, this would not automatically rescue moral discourse from its ostensible commitment to mysterious and/or nonexistent properties, if the discourse could not continue on roughly as before without belief in them (Joyce 2006, p. 202).

Therefore the task for an error theorist is to argue that moral discourse is relevantly like witch and not motion discourse. Motion discourse went on just fine with the lack of belief in absolutism, while witch-discourse did not and presumably could not survive the lack of belief in women with supernatural powers. Again, the best way to approach the question of which of these discourses is most analogous with moral discourse seems to be to inquire into the point(s) or function(s) of moral discourse, and to ask whether the function(s) can plausibly survive the lack of belief in practical clout.

In Chapter 2 I argued that an important instrumental function of moral discourse is to commit us to certain sorts of actions or standards of conduct, largely by deflecting attention away from our 'momentary affect and desire', as Nietzsche put it. More perspicuously, it is in the business of motivating us without consideration or reference to our felt or introspected preferences, desires, or even values. This is not to say that the (kinds of) actions or attitudes we regard as moral or immoral are not (highly) informed by our preferences, but it is to say that insofar as we regard (kinds of) actions or attitudes as morally required, then we will both tend to be motivated to perform them irrespective of whether we regard them as our preferences, and believe (or accept) that others should also be so motivated irrespective of their preferences.

Moral discourse, by hypothesis, deflects attention away from and precludes conscious introspection on our own preferences and values (as such)[276] as means of arriving at decisions about how to act (morally). 'Practical clout' is a name for the supposed feature that moral judgments have that comprises or entails the idea that they

[276] That is, of course we introspect our preferences and values, but not as such. Rather, they are conceived of as intuitions about what is morally right/wrong, which is itself conceived of as independent of those intuitions.

remain legitimate, correct, appropriate, authoritative and/or inescapable, whatever one might (most or most deeply) want to do. Given the broader point of moral discourse, which has to do with coordinating and regulating actions and attitudes among members of social groups with often conflicting desires, it shouldn't be surprising that the discourse impugns the relevance of one's own desires in considering what to do and instead demands allegiance to desire-independent injunctions. Employing Ainsliean resources, I agreed with Joyce (and Frank) that the value of this discourse (and associated emotions) had a lot to do with committing us to certain courses of action.

For Joyce, the question whether moral discourse could stay in business if its users were fully aware that there is no practical clout, as we've both argued there isn't, is evidence for the question whether moral discourse is essentially committed to the existence of practical clout. If it could continue on much as usual, then it appears that Joyce would admit that his error theory is likely false, and if it could not, then he would think the error theory is likely true. And if the error theory is true, then moral discourse is systematically in error, and all first-order moral judgments are false. That certainly seems to be an extreme and radical conclusion, and most people regard it as obviously absurd, deeply dangerous and/or at a minimum, to be accepted as a last resort, only after having explored all remotely promising avenues by which we might save the discourse from such a fate.

I find both this error-theoretic approach and the common, instinctively defensive reaction to it understandable but ill-guided. First, the approach to the problem. As I mentioned, Joyce understands the question whether the discourse can get along without its users believing in practical clout as providing us with a ‘decent chance’ of getting at the question whether the discourse is in fact committed to practical clout.

Viewing the matter this way makes the question of whether there is such a thing as practical clout and whether moral discourse could survive its being discredited interesting and valuable questions largely if not entirely due to their roles in confirming or refuting (this particular) moral error theory. It makes the question of whether moral discourse can survive an awareness of the best account of practical reason worth asking to the extent that it is a good proxy for the untestable hypothesis that a majority of competent users of moral concepts would vote that the discourse is in fact committed to practical clout. *This gets our priorities very badly confused*.

I think it is of very little consequence how this counterfactual vote would turn out, *even assuming that it is criterial* for settling the question of essential commitment (which I doubt). This is especially true since I agree with Joyce that the outcome could be due to highly contingent features of the voting circumstances. I think it is of incalculably greater consequence whether the discourse can continue in its instrumental function—especially if this function is a very important one—should practical clout be widely thought discreditable. One might think that by saying this I’m advocating that the criterion for essential commitment is *really* whether the discourse can continue on roughly as before, rather than how people would vote. Understood one way, this is what I’m saying, but it’s crucial to emphasize that in saying this I am not primarily concerned with the question whether people are uttering falsehoods when they assert first-order moral claims. For me, the question whether the discourse can survive disbelief in or pervasive doubt about practical clout—and whether we should want it to—*just is* what we should be concerned about. The question whether people are uttering falsehoods in their first-order moral discourse is, at best, a proxy for *these* questions, not the other way around.

I agree with Joyce that the discourse is unlikely to be able to survive and serve the kind of function(s) it has traditionally served in the presence of pervasive doubt about practical clout. I think that the peculiarly (academic) philosophical obsession with truth and falsity and (factual or conceptual) error is obscuring the importance of this issue. For example, as we will see in the next chapter, Finlay (2008) argues against Joyce's error theory in a way which is largely consistent with the truth of the claim that the traditional, important functions that Joyce and I have argued that moral discourse serves could not survive disbelief in practical clout. Arguments like his do not directly address what I take to be of crucial concern, but can he helpful nevertheless. In chapter 5, I will use Finlay's critique of the error theory to show how it could be that his critique is essentially correct and yet not only leaves it open whether we should be abolitionists, but even helps to suggest that we should.

Given my critique of the error-theoretic approach, one can see why I find the instinctive defensive reaction understandable. If we focus on what is plausibly the *real* locus of our concern, i.e. the motivational, social and generally *practical* importance of moral concepts, then the defensive reaction makes much more sense than if we had to explain the reaction purely or primarily in terms of people's fear of uttering falsehoods as such. And I do think the defensive reaction is a perception of a real and serious threat.[277]

[277] Believing that there is a real, serious threat here is one of the things that most distinguishes me from Greene (2002, unpublished dissertation), who also argues against using moral discourse.

However, in many ways I think the reaction is mistaken, or at a minimum it is ill-considered (it is almost always *not-at-all-considered*, by design). I think that there are serious costs associated with moral discourse, costs that are as unlikely to be perceived by its typical users, especially its fervent users, as are the costs of god-discourse by its (fervent) users. I have described some of these costs, mostly associated with the Ainsliean 'downsides of the will'. Though Joyce's error-theoretic approach does (radically, in my view) overemphasize the importance of true and false beliefs as such, he does see the dangers inherent in no longer believing in practical clout. He sees this as a basic threat to moral discourse, and given the ostensible importance of that discourse, seeks to find a way to preserve its benefits. His is not a simple, instinctive defensive posture, but one rather grounded in an effort to preserve the very benefits and function of the discourse that he and I broadly agree that it has. His proposal is that we continue to employ the discourse as before, with the exception that we pretend, or accept but not believe, that there is practical clout. I will argue that his proposal suffers from an apparent total lack of awareness that not only are there benefits to be 'recouped' from moral discourse, but costs to be shed. When we factor in the costs as well as the benefits of the specifically committing function(s) of moral discourse, there is no clear answer to whether it is on balance worth preserving, even if we could do so at no cost. This conclusion, combined with the psychological, social and (to a lesser extent) conceptual obstacles involved in deploying a successful fictionalism, render the project distinctly unattractive. Or so I will argue.

In this chapter I want to show the reader that fictionalism about morality is not a worthwhile practical option. Some of the primary reasons for which it is not will carry over to Chapter 5, where I argue that antirationalism itself is not a good *practical* idea, whether or it is defensible as an account of current moral discourse.[278] As in this chapter's critique of fictionalism, my arguments against antirationalism in the next chapter will be accompanied by considerations to the effect that the coming, or perhaps ongoing, demise of morality can be as much an opportunity to be cherished and seized as a calamity to be avoided or mitigated.

As I said, the fictionalist proposals I address below are made by error-theorists who, despite the systematic falsity involved in using moral discourse as we currently do, find that the costs of abandoning it might be so high that we are well-advised to keep conducting ourselves within it in a fictional way, perhaps by 'accepting' rather than believing moral judgments. To motivate her case, the fictionalist needs to show us first and foremost what sort of costs we are likely to incur by abandoning the discourse, and also how we can avoid those costs and keep the benefits by keeping the discourse in a fictionalist way. I will be primarily interested in addressing the claims about the (net!) costs and benefits of moral discourse as it is, and secondarily in the daunting obstacles that any fictionalist enterprise will have to overcome in order to maintain those benefits. I'll argue that even if all the benefits of moral discourse could be salvaged by a fictionalist, it is far from clear that it would be a good idea to do so. That the cost-benefit analysis of moral discourse is far from clearly in favor of retaining it (independent of any

[278] The same will go for 'reforming antirationalism', i.e., an overtly practical proposal to reconceive morality along antirationalist lines.

falsity involved), added to the severe practical challenges to a successful attempt to salvage the discourse, make the fictionalist program distinctly unattractive.

In section 4.1 I'll give a brief introduction to normative fictionalism in general and address the arguments in favor of moral fictionalism by Daniel Nolan et al. (2005). These arguments are all about costs and benefits and do not attempt to raise issues about how (psychologically or otherwise) feasible it would be to implement fictionalism. I will conclude that Nolan et al.'s arguments are almost entirely without merit and in fact provide good (if implicit) arguments in favor of abolitionism.[279] In section 4.2, I turn to Joyce's arguments in favor of moral fictionalism, which are much better, and widely considered to be the best, most sustained defense thereof to date. There I will take issue both with the psychological feasibility and the ostensible (net) costs and benefits of moral discourse. While I think Joyce is entirely correct that moral discourse provides substantial motivational benefits, and even that this is its primary *raison d'etre*, he fails to recognize that inflexible commitment to moral principles can bring with it serious downsides. These downsides are just the ones we would expect if Joyce and I are right about the committing function of morality in the first place, i.e., they are the downsides of willpower. And while willpower might be most important of all in the moral sphere, there also lies its greatest potential for pathology.

[279] I address them despite finding them meritless because I think that some of the the mistakes they make are, unfortunately, quite common.

4.1 Motivating Fictionalism

Hermeneutic vs. Normative Fictionalism

If a discourse is or seems committed to the existence of that which does not, there are two apparent options: get rid of the discourse or keep it and find a way around the ontological worries. Hussain (2004, p. 1) remarks that '[f]ictionalism has made a comeback over the last two decades as one of the standard responses to ontologically problematic domains'. While normally employed as a solution to ontological issues surrounding numbers, properties, etc., it has in the last few years been offered by a few authors as a way to keep moral discourse, despite what they take to be its fundamental flaw(s). These authors take there to be various and serious costs to abandoning moral discourse, but also serious costs (perhaps for some, the impossibility) of continuing to believe in things the discourse presupposes.

It's important to distinguish between two types of fictionalism, namely hermeneutic and normative fictionalism. A hermeneutic fictionalist is someone who thinks that fictionalism (or something much like it) is the best analysis of our actual practice. That is, they do not think that we are really committed to the literal reality of the objects that we might appear to be committed to in using the discourse. Kalderon (2005) has put forth an argument in favor of hermeneutic fictionalism. His account is that the best understanding of our current practice is a fictionalist account. On his view, we don't really believe these judgments or propositions in the first place (we in fact assert no propositions in making moral judgments). Thus there is no presumption that any ‘debunking’ needs to be done; we never were making the error that the error theorist supposes. However, there is of course a serious error of another sort in the failure to know that one’s fictions are fictions. At any rate to discover, as Nietzsche said, that ‘we act once more as we have always acted—*mythologically*’[280] is of course not yet to take any normative stand on the matter. In this chapter I am concerned to argue against normative fictionalism, whether hermeneutic fictionalism is true or not.[281] However, my arguments will be directed specifically at the normative fictionalists who are also error theorists.

Joyce (2001, 2007), as well as Nolan et al. (2005), are explicit in that they are not offering fictionalism as an hermeneutic interpretation of existing practice, but rather as a normative suggestion for how to keep the discourse despite the fact that as it stands it is committed to nonexistent entities.[282] Therefore, the discourse must change while it remains the same. What I mean by that is that the discourse itself, what gets said by whom and under what circumstances, need not change at all, but the understanding of what we’re doing is to change dramatically. It is as if religious people were to continue to read their religious texts, but understand them as mythology rather than history.

[280] BGE 21, italics in original.

[281] Let me briefly explain, in case this strikes the reader as strange. First, given what I take to be the downsides, I think hermeneutic fictionalism has many if not all of the costs of non-fictional moral discourse. I also think that it has all the problems I make for Joyce associated with the awareness that one is within a fiction. It, like at least most noncognitivisms, could be a correct analysis of moral discourse but one that would tend toward the undermining of its value. For simplicity’s sake however, in this chapter I will discuss normative fictionalism as a response to an error theory.

[282] Both authors, though Joyce more than Nolan et al, are noncommittal about the ultimate advisability of fictionalism. Both take themselves to be mapping out fictionalism as an alternative to abolitionism. (p. 180 of Joyce (2001), p. 8 of Nolan).

An Analogy

Normative fictionalists would seem to have an immediate task before them. Having discredited a discourse, or, having found out that something which we took to exist does not in fact exist, the default response would be to stop employing that discourse and believing in that thing. Much like if a person or society became convinced that there is no god, over time, if not right away, one would expect them to stop reflecting on what to do, in public or private, in terms of what any god wants or commands. Of course many factors would affect the trajectory of this process, most obviously the availability and nature of the alternatives, and how 'psychologically entrenched' the god-discourse was in the relevant people. Nevertheless, if the idea that there were a god who wants or commands things were widely thought discreditable, then all else equal, we would expect people to find another way of conducting their normative discourse.

For a fictionalist, all is not equal. She might be quick to point out that the god-discourse served important functions in the society and that to simply toss it out would be to toss out something very valuable to the people in the society, whether or not it has false presuppositions. She might argue that the discourse is so important that it should be kept in spite of its systematic falsity, and offer the people a way of keeping it without even having to utter falsehoods! This might be exciting. But for it to be exciting, they (at least some of them) would probably want to know what was so great about the discourse that they could not or should not do something else instead. Perhaps, some might think, they could talk about what is moral and immoral without having to make reference to any god. Some people might find this even more exciting, especially if in the course of examining what was so great about god-discourse, they also came to think that it had been holding them back in ways they had not supposed.

Likewise, the moral fictionalist, before worrying about (other) problems with fictionalism as a substantive proposal, must convince us that we have any need for what they are selling. They must show us what is so (relatively) great about moral discourse compared to the alternatives that we should be interested in attempting a fictionalist reconceptualization of it, especially since doing so is sure to involve some costs. Thus, their first task is to show that abolitionism has serious costs that fictionalism can at least hope to avoid. I will focus first on four arguments against abolitionism mustered by Nolan et. al, before turning to a lengthier discussion of Joyce's argument.

Four (pseudo-) Costs of Abolitionism

The first advantage Nolan et al claim for fictionalism over abolitionism is 'psychological convenience.' They list 'right and wrong and ... duties, virtues, rights, justice and obligations' as examples of elements of our common discourse that it would be difficult if not impossible to give up (310-11). As an abolitionist, I'm not recommending getting rid of any of these terms. Right and wrong have perfectly good normative uses outside the moral domain. Duties, rights and obligations also have perfectly acceptable uses outside the moral domain, understood as elements of various institutions, both formal and informal. Signing up for the military imposes duties and obligations on you, and we all have rights in virtue of being citizens of a country which has certain institutions, most notably legal ones, which prohibit various forms of infringements on our freedom on the part of government or other individuals. It is only the usages of these terms which rely on practical clout which would be abolished on my proposal.

Finally, I see nothing peculiarly or necessarily moral about the notion of a virtue. Anscombe (1958, p. 1) thought that modern conceptions of morality did not fit into Aristotle's virtue ethics. In general, I don't think that virtue ethics is a *moral* theory per se, but is best thought of as one kind of answer to the fundamental question (which is often called 'ethical') of how to live.[283] However, I don't need to take a stand on whether virtue ethics is best thought of as a moral theory (in the modern sense of 'moral'); it is enough to note that while it is true that people often speak of moral virtues, there are nonmoral virtues as well, and therefore not even a *prima facie* reason to get rid of all or even most talk of virtues.

Nolan et al. compare giving up moral discourse to giving up folk psychology (and include 'goodness and badness' as terms we'd have to give up!). Well, if we indeed had to stop saying good, bad, right and wrong and all the others, then maybe the comparison would be apt. The primary reason that giving up folk psychology would be tremendously difficult is that we would have nothing of equal or greater value to replace it with at present. As much as I think 'belief' and 'desire' are imperfect concepts, and might even someday be replaced by others, at least in some scientific enterprises (Churchland, 1979), we have little choice but to engage in the talk now, inserting caveats here and there where we think they might be misleading, because we simply have no alternative that allows us

[283] There is a large and interesting literature on whether virtues must benefit their possessors. If they must, then that alone puts heavy strain on calling them moral virtues, insofar insofar as the reason to cultivate the moral virtues is because of these benefits.

to do the tremendous amount of communicative, predictive and explanatory work that those concepts allow for. The case is not remotely the same with morality. We do have terms, some of which are roughly as imprecise and problematic as folk psychological terms, like 'values,' 'interests' and ‘goals’, as well as the folk psychological terms themselves, like belief and desire, which figure prominently in accounts of prudence, and practical reason more broadly. We also have emotion- and emotion-derived terms such as ‘shameful’, ‘contemptible’, ‘prideful’, ‘laudable’, and ‘honorable’, as well as Williams’s ‘thick terms’ such as ‘courageous’, ‘cowardly’ and many others. These terms all seem to play important roles in regulating our attitudes and dispositions to act in various ways.

So if the point of normative discourse is ultimately to figure out how to live, what to do and perhaps why to do so, we *already have* quite a rich source of concepts to draw on. In addition, I think there are good reasons to believe that these resources would improve and grow richer if they were the focus of our attention rather than the moral concepts. It shouldn't be too surprising that we don't know our true interests or values very well if we rarely consult them directly in making 'moral' decisions.

I am tempted to save the next rebuttal for last, for it is the one I take to be most revealingly wrongheaded—the argument from impracticality in applied ethics. This is the claimed unavoidability of getting 'side-tracked' on practical ethical matters. The authors suppose that an abolitionist, if asked by an anguished friend whether to have an abortion or not, must respond in the following manner: ‘[W]ell, I don’t really believe that you should do anything since I don’t really believe that there’s anything that ever should or should not be done; nor do I believe that it would be wrong for you to have an abortion, or right for you to have one for that matter ... ' Fictionalism, they maintain, lets us 'get to the point of the discussion right away' (311).

If I ever were to believe that anything like such an absurd response to a friend in need of advice were remotely entailed by abolitionism, I would recant in horror. Such an entailment, if it were one, would be sufficient to render the proposal completely ludicrous. But rather the notion that there is such an entailment is ludicrous.[284] The mistake is two-fold, and both aspects are important. First, an abolitionist clearly has no requirement, nor any apparent reason, to start in on an error theory, even if the person were asking for moral advice, which is not necessarily the case. Second, it has not even been a page since the authors have used 'should' in an explicitly pragmatic, as opposed to moral way. That is presumably because they don't believe in the moral way, and wanted to be clear to the reader that they were not advocating keeping moral discourse despite its flaws for moral reasons, but rather for pragmatic reasons (310). There is no suggestion of why an abolitionist would or should not be able to give advice to a friend on what could be considered pragmatic grounds. Of course they needn't be considered pragmatic. They could be *any grounds other than moral ones.* Perhaps one particular abolitionist thinks that the best thing to do is whatever strikes one as most beautiful. That, I maintain, need have nothing to do with morality, nor is it clearly pragmatic. Any sort of non-moral practical reason is in principal available to the abolitionist.

[284] I know this is strong language. It might be supposed that I should not spend time on views that I find absurd. But I'm afraid I cannot afford not to, as it is very far from obvious to me that the concerns that Nolan et al. express would not be shared by a large portion of my audience. But then it might be thought that my language should be softer. I am sensitive to this point. But sometimes strong language is warranted. I would like the reader to get a sense of what suggestions such as these look like from my perspective, and, I hope, to begin to see them that way herself. After all, if even *moral error theorists* such as Nolan et al. cannot separate narrowly moral concepts from broader normative discourse any better than they do, it stands to reason I might have to make some of these points quite strongly for an audience that has not given as much consideration to the possibility of abandoning that discourse as they have.

Not only do I claim that an abolitionist can ‘get to the point right away’, but *the fictionalist cannot* insofar as she retains peculiarly moral discourse. Let’s suppose that the fictionalist wants to help her friend as much as possible. That is to say that the fictionalist wants to help her friend do whatever is best for her.[285] But engaging in moral talk would seem an odd strategy to use, as it is not well-suited to this end.[286]

Presumably, whether getting an abortion is the right thing to do will depend on quite a few factors, most conspicuously financial and psychological (including moral and/or other religious beliefs), whether and how much the potential mother wants a child, the status of her relationship with the father, and so on. As her friend, the fictionalist might already know some of these things. But it seems to me in helping someone decide a question as important as this, the *first* thing one would want to do would be to ask questions along these lines; try to do what many counselors do, and simply get the person talking out loud about their own values, desires, perceptions and so on, and let them be guided to the best decision they can make in the circumstances, while the friend helps in a variety of ways, such as pointing out if the friend is perhaps deceiving herself, ignoring important information, etc. That would be what I would consider getting to the point of the discussion right away. I would think the *last* thing a fictionalist would want to do is start in on what are, by her own lights, fanciful rights, duties, permissibility and

[285] I am supposing that the fictionalist doesn’t want to help her friend do what is moral, since she doesn’t believe in morality. At first, this seems a safe assumption, but perhaps the fictionalist is really committed to the fiction. Since the fiction requires that one be concerned not only with what is best for one’s friend, but rather with what is most moral (according to whatever moral theory about which one is fictionalist), then maybe the fictionalist will give advice not suited to best help her friend (except by accident), but rather in keeping with whatever moral theory about which she is a fictionalist. With fictionalist friends like that, who needs enemies?

[286] Not that it can’t be used this way, but ‘duty’ and ‘obligation’ are not the tools I’d use to explore the normative landscape if I wanted to discover what was in my or someone else’s interest; they are arguably concepts which have as a core feature abstraction (or distraction) from just such concerns.

obligation, which at best would likely be useless and at worst mislead the friend into doing something which is worse for her than an available alternative, due to their placing obstacles in the way of her reflecting on her own values and emotional dispositions.

One thing to note especially is that the claimed advantage for fictionalism in this regard only begins to get traction on the assumption that the person was wondering what she *morally* should do. If she were wondering something else, like, perhaps, what would be the best thing *for her* to do, then there is no reason an abolitionist would have any more difficulty than a fictionalist. In most cases, people do not have it clear in their heads whether they are asking for specifically moral or non-moral advice. They just want to know what to do, and they are variably moved by what we would call moral concerns. But we certainly may have a friend who wants to know whether having an abortion is morally ok. She believes in morality, specifically that there are things that she should and should not do, regardless of her desires and interests. How is the fictionalist going to help her?

The first thing to note is that just because one's friend 'believes in' morality, that in no way need alter the fictionalist's goal, which by hypothesis is to help her friend do what is best for her. So even if we know that our friend believes in or cares about morality, that needn't prevent us from embarking on just the same program as before, namely inquiring into facts and values and helping our friend think along lines that are best suited to helping her do what is best for her. But suppose our friend can only think in moral terms, or, to return to a helpful analogy, suppose she can only think in terms of what God wants her to do.

Many atheists, including myself, have religious friends. Must we be fictionalists about their religion in order not to have to explain our atheism to them if we are asked for advice? True, if they are our friends, they will probably know we are atheists, but so would they likely know we are fictionalists. So it seems no more or less likely that a friend would ask for advice about what God wants from an atheist friend than she would ask for specifically moral advice from a friend known to only be pretending to believe in morality. The interesting question is what sort of strategy an atheist should employ in helping religious friends; those will presumably be analogous to the resources an abolitionist has in helping moral friends.

Let's take it as a premise, as it seems a fictionalist should, that moral beliefs, like religious ones, reflect at least in large part the values of the holder of those beliefs. So if a religious friend asked me if she should steal some money, where that involves wanting to know what God would think about it, it is entirely open to me to ask her what *she* thinks God would think about it. I can even say things like, 'From the things I've heard you say in the past about God ... ' it either does or doesn't seem like He'd mind. In the case of abortion, I could ask, 'Does God take into account your financial situation, the relationship you have with the father, whether you feel you are ready for children?' and so on. One can equally ask about a friend's moral beliefs, and help the friend to do whatever is most consistent with those beliefs. Of course, we can even help to guide the friend's thinking away from those beliefs if they seem harmful (perhaps our friend has the moral/religious belief that the husband should always have the final say in such matters, or that they should attempt to maximize utility).

I have taken a lot of space dealing with this point because I think it is instructive. Proposing to move beyond moral discourse leaves us not only perfectly capable of giving normative advice, but often in a better position to do so, especially with respect to friends (and ourselves). It is an instructive (and surprising) mistake to suppose that an abolitionist would wish to say, 'I don't really believe that you should do anything since I don't really believe that there's anything that ever should or should not be done.' An abolitionist has no reason whatever to think that there aren't things that should or should not be done. Of course, it won't be a moral should, and presumably will be related to some person or other entity's values, interests, goals, etc. Presumably Nolan et al. had in mind a moral should in saying that there's nothing that ever should or should not be done. And of course in this an abolitionist (who is also an error theorist) would agree. But this 'theoretical nihilism' need not give rise to 'practical nihilism' in the way they suggest. It is in fact the mission of this abolitionist to free us of the need to fall back on such externalist, moralistic notions in order to give purpose to our lives or guide our conduct.

A third advantage they claim for fictionalism over abolitionism is 'expressive power'. They adduce examples involving numbers and properties to show how much easier and more powerful it is to say certain things using a number- or property-fiction than it would be to say the same things without using numerical expressions or talk of properties.[287] As things stand, I have no objection to fictionalism in the domain of numbers and properties; it even seems like the right way to go so far as I can tell. One

[287] This of course assumes that there is some problem with number- or property-discourse in the first place, an issue I am not taking up here.

reason for this is that I agree with the authors that a fictionalist can ‘commit themselves to something true in virtue of expressing the falsehoods’ (312).[288]

The authors think that moral expressions are relevantly similar to number and property talk. Moral expressions allow us to say things such as, ‘the duty to stop and attend to the victims of a car accident is more important than our duty to keep our commitment to meet our friend for lunch’ (312). The ostensible benefit of using such expressions is that we can use sentences with such moral vocabulary ‘to imply things about non-moral features of the world, where it seems difficult to identify those features in non-moral terms’ (312). Nolan et al ask rhetorically what non-moral features are captured by ‘duty’ in the quoted sentence, citing the difficulty of coming up with an ‘appropriate non-moral literal paraphrase’ of the sentence.

It is unclear to me why I should be looking for a literal paraphrase any more than I should be looking for a literal paraphrase of the statement, ‘God commands that we stop to help car-accident victims rather than keep lunch appointments.’ I see neither more nor less reason for supposing that this statement can be used to imply non-moral things about the world; in fact, it seems to me that it can likely imply precisely the same things, if any. Now presumably someone who does not believe in God (and even maybe someone who does) will think that such a sentence reflects a (commitment to a) value-system in which helping those in need takes priority over keeping lunch appointments. And that is also the non-moral feature of the world that talk of ‘duty’ stands in for (unless perhaps it is a legal duty referred to). At its least problematic, saying that it is our duty to do A rather

[288] Though my understanding is that the expressions are not to be understood as falsehoods but rather ‘true-in-the-fiction’. It is a point irrelevant to my concerns, at any rate.

than B reflects the fact that we have a system of values that prioritizes A over B. And given that system of values, the statement implies that we *should* or *must* do A rather than B.[289]

If the fictionalist believes that we should attend to victims rather than lunches, why not just say so? That sentence is at least as short; what does talk of duty add? It adds just the sense of inescapable and authoritative obligation we've been talking about, i.e, it adds practical clout. That is plausibly why a literal paraphrase in non-moral terms is not possible, if it is not. Now there might be some non-moral description of the world into which (moral) duty-talk could be literally paraphrased. Let's suppose there is. It seems like it would be interesting to find out what that is, since the truths are presumably largely about ourselves and our values. But a fictionalist, just as a moral realist, directs attention away from what could be interesting truths about ourselves. Of course it is possible for a fictionalist to employ the moral discourse in certain contexts, and in other contexts reflect upon and investigate her values and other commitments. That is not necessarily objectionable in itself, and the primary advantage I see in adopting this strategy is to exploit the motivational advantages that I have agreed moral discourse can bring. I'll be discussing the claimed virtues of fictionalism with respect to these motivational benefits below in response to Joyce's arguments, so I won't take up the issue here.

Now suppose there is not any description of the world in non-moral terms into which the language of moral duty can be paraphrased. Then I take it that there *is no truth*

[289] Sometimes the dutues we think we have do not reflect our actual values. This is the much more problematic kind of case. Also, none of this is to suggest that a 'system of values' is or need be arbitrary. There might be (and I believe there are) systems of values that are much better than others. The notion of betterness will of course not be moral.

to which a fictionalist commits herself in virtue of expressing the falsehood. Nolan et al ask us to consider the claim, ‘The property rights of some farmers have outweighed the rights of the environment in this case’ (312). This ‘common sort of moral claim’ is supposed to imply difficult-to-identify non-moral features of the world. They claim that using fictionalism, she can maintain that it is ‘*literally* true that if there are rights then the property rights of the farmer are more important than the property rights of the environment’ (312, their emphasis). OK, but by hypothesis, there are no moral rights, so I’m unclear on how that is helpful. I don’t see how it helps us much to claim that it is literally true that if there were Allah, then His judgment would be that women are to be covered, unless we’re holding out some nontrivial possibility that there is Allah. But once we are quite convinced error-theorists, either about gods or morality, these conditional truths (if they are such) don’t seem helpful.

We are trying to figure out how best to live, I take it. Either there are some good (by the error theorist’s lights) reasons to prioritize the farmers’ property rights (which are clearly institutional and thereby as real as the institution) over considerations of protecting the environment, or there aren’t. If there are, why not try to spell them out? If there are not, then the *prima facie* absurdity of supposing that ‘the environment’ has (moral) rights doesn’t even have a mitigating benefit, but only serves to deflect our attention away from the fact that we value our environment onto some fanciful duties we have towards it.

The last argument against abolitionism by these authors that I’ll address is that moral discourse is important ‘in maintaining any social relationships’ (312). This argument essentially points to the fact that realist moral discourse is how we engage each other in normative questions and has provided the framework within which we decide questions of right and wrong. This framework is the framework of rights, duties and obligations. The authors cite it as a virtue that the framework 'tells us that there are correct decisions about what the morally appropriate decisions and outcomes are, and when there is genuine disagreement about what ought to be done, at most one of the parties can be correct' (312).

While I want to leave it an open question to what extent and in what sense and contexts the feature of moral discourse that entails that there is one correct answer to moral disputes is a virtue, I take it that there are plenty of discourses that can accomplish this. A moral abolitionist can love this aspect of moral discourse, and it might likely be in whatever replacement she proposes. The aspect of moral discourse she presumably won't like is that of (non-institutional) rights, duties and obligations. Supposing that these peculiarly moral concepts are crucial for social functioning is reminiscent of Locke's famous exception in his Essay on Toleration:

> Promises, covenants, and oaths, which are the bonds of human society, can have no hold upon an atheist. The taking away of God, though but even in thought, dissolves all; besides also, those that by their atheism undermine and destroy all religion, can have no pretence of religion whereupon to challenge the privilege of a toleration.[290]

A fictionalist response to the (spirit of the) above quotation would be to claim that while God-discourse is fundamentally flawed (due to ontological issues), we should keep the discourse. That is because the discourse involving commands and prohibitions of

[290] *A Letter Concerning Toleration*, 1689.

God provide a common and well-established framework within which normative discussions take place, a framework that implies that at most one disputant in a disagreement is correct.

My response is that for those of us who feel up to the challenge of moving beyond a fictional God-discourse (whether hermeneutic fictionalism is true of that discourse or not), a discourse that makes reference to our values and interests holds the promise of helping us to direct attention toward *those things that really matter to us*, rather than toward substitutes for those things, substitutes which have the demonstrated potential to lead us seriously astray.

The first thing I want to make clear is that I no more think that abandoning a god-discourse is at present a live or good option for all people than I think this is the case for abandoning moral discourse. What I am offering is the suggestion that for those people who have come to think there are serious problems with moral discourse, whether for ontological or other reasons, the discourse not only *can* be abandoned, but doing so holds out analogous benefits and risks of abandoning a god-discourse. But before I discuss those benefits and risks directly, I want to examine another rationale for keeping moral discourse despite its being systematically flawed.

4.2 Motivational Fictionalism

Joyce (2001)[291] argues for continuing to employ moral discourse for its anti-akratic benefits. We have seen in previous chapters the reasons for thinking that moral judgments have a committing function and how in general commitments serve as bulwarks against weakness of will. That Joyce and I agree both that, and roughly how, moral judgments serve to protect us from weakness of will provides important common ground between us. I agree with Joyce not only that moral judgments have this function, but that they have it in virtue of their playing a specifically committing role, and even further that they fulfill this role (at least in important part) by subjectively closing off options in (at least often) in a linguistically-mediated fashion. My objections to his fictionalist proposal[292] happily grant these crucial assumptions and take issue with his fictionalism largely on its own terms. Indeed, in making these premises central to what is widely considered the most sustained and careful argument yet to be made for normative moral fictionalism, he has provided me with a valuable opportunity. If I can show that the strongest proposal for normative fictionalism is deeply, multiply and likely fatally flawed (largely) on its own terms, then I will have done as much as can be hoped for to close the door on this proposal. The closing of that door, and the reasons why it should be closed will, I hope, help us direct our energies toward (further) cultivation of a post-moral discourse.

[291] All page numbers will be from this work unless otherwise indicated.

[292] By saying this I do not mean to imply that he is a committed fictionalist. He is merely making a case for it, as an alternative to abolitionism.

Joyce and I are in agreement not only that moral concepts do have motivational benefits but that that is their (primary) *raison d'etre*. Joyce's proposal is to capture as much of the motivational power of moral concepts as possible without literally believing in them. The idea is that we might be able to derive many of the motivational benefits of moral discourse by pretending to believe in the concepts without actually believing in them.

The fictionalist option is of course for those who have come to believe, for whatever reason(s), that moral discourse is fundamentally flawed. While most will presumably continue believing, those of us who know or believe the discourse to be systematically flawed don't have this as a realistic option. Even if we could somehow manage it psychologically, Joyce argues that allowing ourselves to believe that which we have good reason to disbelieve is likely to have long-term practical problems. Without elaborating, the reasons for this are 1) that it's (practically) better on average and in general to have true beliefs than false ones, 2) that useful false beliefs require compensating false beliefs in order to cohere with the rest of one's beliefs and 3) that having a policy of aiming at the truth 'is the best doxastic policy around', anything less leading to, in Charles Pierce's words, '"a rapid deterioration of intellectual vigor"' (179). I won't discuss the merits of Joyce's brief arguments to the effect that straightforwarldly believing as we did before would be practically sub-optimal. I think that this is a non-starter psychologically (short of very elaborate and possibly science-fiction-esque methods), and even if it were possible, I would be opposed to it, for reasons that will, I hope, only get clearer.

Given that we cannot go on as before, Joyce sees the principal options as between abolishing the discourse—in the way we've done with phlogiston, witches and vitalistic life force—and continuing to employ the discourse fictionally. Joyce (2001) acknowledges that for all he knows the best thing to do is to abolish it. But given the ostensible benefits that moral discourse provides us (which are presumably absent in the others mentioned), we should not be so quick to do so. For by hypothesis it plays an important, perhaps even crucial role in fighting akrasia. When we consider that the akratic episodes it helps us avoid involve some of the most important and far-reaching values and interests we have, we can see that Joyce is right to warn us not to peremptorily toss aside such a valuable tool.

4.2.1 Belief vs. Acceptance

Fictionalism promises to let us keep at least much of what moral discourse does for us without many or any of the costs involved in believing it. Joyce argues that in order to pull this off we have to 'accept' rather than believe the moral judgments we make. Thus the notion of acceptance is crucial to Joyce's proposal. He has to make this concept do two essential things that stand in tension with one another, i.e., make it capable of doing (much) the same sort of work as belief does while being clearly distinct from belief.

Joyce spends most of chapter 7 of his 2001 book countering the objection that the distinction between acceptance and belief can't be maintained (Newman, 1981; Putnam, 1971). He attempts to mark this distinction by appeal to a 'critical context': "Context *n* is more critical than context *m* if and only if *n* is characterized by a tendency to scrutinize and challenge the presuppositions of *m*, but not vice versa" (2007, p. 21). The purpose of making such a distinction is to provide an analysis of the two kinds of context that actors, novel-readers, movie-watchers and the like can inhabit. Crucially, all of these people can and do 'step out' of their fictional context, and when they so are disposed, deny the literal truth or reality of what was going on in the fiction. The moral fictionalist must likewise have a context in which she is disposed to deny the literal truth of moral claims. These contexts are not distinguished in terms of the amount or percentage of time one spends in each context, but at least in the fictionalist's case, she is disposed to step out 'if pressed in an appropriately serious manner' (2007, p. 18).

Joyce spends several pages defending the basic distinction, and describing what it means to be in one context rather than another. I will grant that the distinction can be maintained *in general*; however, as we'll see it will be more difficult to maintain it as a moral fictionalist, so I will briefly review the important elements of the distinction. They are: 1) the relationship between more and less critical contexts is asymmetric, i.e., from the more critical context one evaluates the legitimacy or justification of one's more ordinary, everyday thinking but not vice versa, 2) which context is more critical has nothing to do with the amount of time spent in the respective contexts and 3) someone has to have actully inhabited a context for it to be a critical context. If and only if someone S is disposed to assent to some statement T from within one context C1 and dissent from it in a different context C2 (both of which she has occupied), and if C2 is more critical (in the asymmetric, scrutinizing and challenging sense) than C1, then S is making a fiction of T (193).

The class of fictive judgments can be very broad and variable. The most important dimension along which fictive judgments can vary, for our purposes at least, is the extent to which one is 'immersed' in the fiction. At one extreme we can and do make 'fictive judgments' of a sort when we read stories about, say, dragons. If someone asked us whether we believe what we're saying we would find that silly—we have easy and instant access to our more critical context and that should be obvious in this case. At the other end of this spectrum are 'the more fully immersing kind of fictive judgment for which our disbelief will only be admitted in a very critical context, such as a philosophy classroom' (194). Joyce does not claim to know whether anyone in fact makes fictive judgments at this extreme end of the spectrum, but he 'see[s] no reason why we could not' (194). If we are to make a fiction of morality, it will have to be of this latter sort if it is to 'engage our emotions, guide our actions, influence our decisions [and] bring the pragmatic benefits of morality that we might expect—all without belief' (194).

As Joyce notes, some of his analysis invites more analysis, such as what it means to 'scrutinize and challenge' (2007, p. 21). I want to allow that these aspects of the analysis are not importantly problematic. Doing so can only make the fictionalist case stronger, and I want it to be as strong as possible. I think the best way to illustrate the distinction Joyce has in mind, and the attitude he's recommending, is to follow him in giving some examples. The most obvious place to look for illustrations of how to make a fiction, is fiction.[293]

[293] Joyce acknowledges important differences between morality as a fiction and literature as a fiction, most obviously that literature is created *as a fiction* whereas that's doubtful for morality, and that literary fictions have fleshed-out stories, whereas morality does not. This latter difference seems especially important since disputes about what happened in a novel can be generally settled by turning to the text while morality has no such text. Nevertheless, Joyce does maintain that the relationships that hold between morality's posits

The first examples Joyce gives are negative. A tourist showing up at Baker Street to see where the *real* Sherlock Holmes lived is not make-believing. He ignorantly believes in what is only fictional, but is not making a fictive judgment. Now Joyce asks us to imagine that a different tourist 'at some level' knows that Holmes is fictional but even when asked in all seriousness whether Holmes is fictional, he says no, he's real. He *seems to* sincerely believe in Holmes' reality and 'perhaps only sessions of psychotherapy will bring him to admit the *lie* he lives' (195, my emphasis). This person is not making a fictive judgment either, but is rather self-deceived; he has 'embarked on a self-induced course of error-acceptance' (195).

Now a third Holmes fan loves the books, tours the sites that Holmes visited in the novels, vividly imagines Holmes pursuing Moriarty from the spot he's standing on to another in view. He 'gives in to the fiction' (196) for the day and has fun not thinking about the fact that Holmes is Conan Doyle's creation. This person is making fictive judgments. Joyce say that 'the moral fictionalist is in important respects like this third Holmes fan. 'He is not self-deceived, *since it is within his grasp to enter the 'critical mode' should he care to*' (196, emphasis mine).

Joyce borrows from Kent Bach (1981) the idea that a person can believe not-*p* but think and act as if *p* while not believing it. While Bach developed this idea as an account of *self-deception*, Joyce favors it as an analysis of a fictive judgment.[294] Specifically, in answer to the question what is going on in our heads when we make a fictive judgment, Joyce answers, 'thoughts' (197). We can think propositions such as 'Holmes lived on

(such as those between rights and obligations) and conceptual constraints such as that causing pain is relevant to what is morally right to do, provides it with 'rough internal rules of disputation' (195).

[294] It is striking that this does not bother Joyce, a topic I will return to below.

Baker Street' or 'I'm a bear' without believing them. And importantly, thinking such thoughts can have significant effects on our emotions and motivations.

With the examples above I am content to grant that Joyce has adequately described and illustrated his conception of acceptance as distinct from belief. Having done this, it's important that he show that the thoughts involved in such acceptances can perform roughly the same kind of motivational role as beliefs. Here Joyce is much briefer, providing a few examples of how 'seemingly trivial aspects of a person's mental life' can make significant differences in their behaviors. The first is that some people regularly take 'insurance' in blackjack, which on average they will ('obviously', says Joyce) lose money by doing. 'It is hypothesized' (217) that such behavior is induced by the simple labeling of the strategy as 'insurance'. The second is that people playing Prisoner's Dilemma games can be affected in their playing strategies by exposure to emotional stories of sacrifice as well as stories about acts of violence. The former class are much more likely to cooperate than the latter. People are also affected in iterated prisoner's dilemmas by whether information regarding how the other 'prisoner' acted in the previous round is delivered by a hand-written note passed through a slot or whether by one of two lights being illuminated, though the content of the messages is constant. Lastly, what would seem the trivial matter of whether one is wearing a white lab coat (or many other trivial differences) can make a difference in how much painful and ostensibly dangerous electric shock one will be willing to deliver to another whom one believes to be a participant in a learning experiment.[295]

[295] Insurance hypothesis in (Wagenaar, 1988). Prisoners' Dilemma studies: exposure to stories (Hornstein et al., 1975); method of information-delivery (Enzle et al., 1975); electric shock (Zimbardo, 1970).

Though none of these examples involve a fictive judgment, they are meant to suggest that what might seem to be motivationally ineffective, such as thinking that some action is ‘just wrong’, can be effective even without the person believing it. I think that Joyce’s first example is suspect, but it doesn’t matter to me, because I am in general fully convinced that thoughts that fall short of belief can be motivational. However, I think that this in itself is far from enough to render the fictionalist proposal on offer feasible. The most obvious problem is that none of these effects have been demonstrated to work in the context of the subject’s being *aware* of these effects. Plausibly many will not work at all and others will work much less stably and powerfully. And even if this problem could be overcome, Joyce is far from showing that fictionalism is a (likely) better option than any of a number of possible forms of abolitionism.

4.2.2 Could We Do This Even if We Wanted To?

Joyce is offering fictionalism as a practical proposal, specifically as a competitor to abolitionism. Therefore in order for us to think that we should attempt to undertake the project of becoming fictionalists about morality, fictionalism has to be not only feasible but promise more net benefits than abolitionism. This section will challenge the idea that the fictionalism Joyce proposes is feasible. The next section will grant for the sake of argument that the program is workable but will maintain that its hoped-for benefits are, even if attainable, likely to be outweighed by its costs.

The problem of groups

First, I am going to make what could be seen as a practically fatal objection to the entire proposal, and then set it aside. The objection stems from the fact that according to Joyce, fictionalism can only work in groups, i.e., there can be no 'lone fictionalist', or at least not one who isn't lying when engaging with true believers. This is because Joyce believes that whether one makes an assertion or not, whether or not one expresses a belief, is fixed by linguistic conventions. It's important to note here that to 'express a belief' does not require that one actually hold the belief. I can express the belief that Santa Claus is real without actually believing it, just as I can express regret without actually feeling it (202).

Since the expression relation is conventional, not causal, if a fictionalist behaves in such a way that conventionally would be understood to express a moral belief, without prefixing his speech in such a way that serves to withdraw assertoric force, then he has in fact expressed that belief. Since he doesn't in fact believe what he is saying, he would either be lying, or would be operating in violation of the conventions which guide language usage and meaning to such an extent that he will have 'in a sense, ceased to speak' (203). The conventional understandings will have broken down to the extent that there is not genuine communication occurring between the would-be fictionalist and his interlocuters.

Joyce thinks that such attempts to be lone fictionalists '[encourage] a fractured and miscommunicating society' and that the costs of doing so are likely very serious (203). As we just saw, *according to Joyce*, such a person is failing to communicate or lying (204). Hence he is explicit that in order for fictionalism to be viable, a group must adopt the fictionalist attitude. The group needn't be homogenous in this respect, but there must be enough members to underwrite a convention to the effect that expressing moral judgments are to be understood as lacking assertoric force (204).

It is surprising that Joyce sees no need to address the problems this generates if we are to take seriously the idea of adopting fictionalism as a genuine practical proposal. In order to be fictionalists, we will have to conduct all our moral discourse with the rest of society in such a way as to make clear that we don't really believe what we're saying until such time as a convention is established that prevents our having to do so. This runs us smack into what I said was the most wrongheaded objection from Nolan et al. Their objection involved the high cost of having to do metaethics prior to giving advice to a friend. One of the ways this problem could presumably be avoided, as I mentioned, is that our friends are likely to know that we are fictionalists. But if they do not, then according to Joyce we will have to either explain that to them before having moral discourse with them or lie to them. As bad as this is, it is worse when it comes to those who don't know us well, since we will almost certainly not have had a chance to explain our metaethical beliefs to them prior to having substantive interaction with them and so will almost certainly be in the postion of having to choose between lying and prefacing our discussion with a highly distracting and likely confusing prologue explaining that we are just pretending to believe what we say so long as we are engaging with them in typical moral discourse. And as I said above, any such entailment should constitute a reductio of the entailing proposal.

It might be thought that one could engage in the fictional discourse only with those who already know about one's fictionalism, and either avoid moral discourse outside that circle, effectively being an abolitionist in those contexts, or explicitly state that one is only making-believe. Both aspects of this strategy are highly problematic. First, Joyce has had to spend a lot of time explaining what's going on with belief vs. acceptance, why we should be error-theorists in the first place, and why be fictionalists given the truth of error theory. Is it remotely plausible that we will be able to explain satisfactorily to even those close to us what we're doing and why we're doing it to the extent that it is not a serious source of confusion and distraction?

The prospect of telling others with whom one might happen to want to engage in substantive moral discourse that one is only pretending to believe the judgments is, I hope, so absurd on its face that it requires nothing further from me. That leaves us with the options of lying or avoiding moral discourse with those to whom we cannot feasibly explain our metaethical stance prior to entering into the discourse. The first is unacceptable by Joyce's own lights.[296] The second is already a huge concession to my proposal, which for the moment I'll be content merely to note.

Suppose now that the fictionalist responds thusly. His is a long-term mission. At first he might be forced to have to explain his fictionalizing ways to those with whom he most frequently engages in moral discourse. With others, he can avoid moral judgments until and unless such time as he has been able to do the same with them. Perhaps fictionalism will sweep the civilized world over the course of the next few decades and he will have much less work to do than he might otherwise have had to. But, barring this happy dream, he will repeatedly face the prospect of engaging lyingly, avoiding the

[296] Though I would probably not be as concerned as he is about this if I agreed that the benefits were significant.

discourse, or engaging in metaethics with would-be moral interlocutors. Over time, he might suppose, the fraction of the people with whom he can engage without preamble in honest fictional discourse will rise, and fictional moral discourse will gradually come to be more common than avoidance thereof, so there will be a long-term move toward fictionalism, though he might have to be practically abolitionist with most people in the beginning.

Even if we ignore everything else that might strike one as unenviable about such a task, it runs counter to what is peculiarly important to *moral* fictionalism. Moral fictionalism will very likely require a very high level of immersion. The critical context is to be inhabited as rarely as possible, at the risk of losing the benefits one is going to such lengths to secure. It has to be, in Joyce's words, a 'life-strategy' (219). What I outlined above is a *kind* of life-strategy, but the wrong kind. As I hope will become clearer below, the kind of life-strategy required for a successful moral fictionalist is the sort, just as Joyce says, in which the critical context is accessed only (very) rarely. Being a successful moral fictionalist is, at best, in strong tension with repeatedly going over the arguments for the error theory, explaining that and why one is a fictionalist and what fictionalism amounts to. This is because in order to be a successful *moral* fictionalist, it is particularly important that one not be too *aware* of it, especially at the times when one really needs it. For it to do its work when it needs to, it's crucial that one's attention not be drawn to the fact that one is only pretending (218). One needs to 'forget' this as effectively as possible when it matters, and the more time and energy one spends with one's attention on why moral beliefs are bunk, and why and that one is a fictionalist, the more unlikely it is that one will forget what one needs to when one needs to forget it.

The problem of immersion

Now, in keeping with my goal of making the strongest possible fictionalist case, I want to grant that the concerns I've raised above that relate specifically to the problems that arise for fictionalism due to its ostensible requirement of a sufficiently large fictionalist group to even *begin* to live the fictionalism in the way it needs to be lived in order to be worth the trouble, can be resolved. Instead I want to focus on the aspect of fictionalism that plausibly requires that it be a 'life-strategy.' In this section I will argue that 1) it will be very difficult to achieve the level of immersion required in order for the fictionalism to provide its promised benefits and 2) one achieves this level of immersion at the price of what looks *very* much like self-deception.

Suppose that Joyce convinced us that there is a tenable distinction between belief and acceptance, where the latter involves thoughts that can movitate us without our being disposed to assent to those thoughts in more critical contexts (and therefore the thoughts are not beliefs). We saw that it is possible to pretend, without apparent self-deception, that Holmes existed for real and did his deeds in the real London streets. The reason Joyce adduced for thinking that such a person is not self-deceived is that he can occupy the critical context in which he denies that Holmes existed 'should he care to'. It is just this contrast with the previous hypothetical tourist who in some sense knows that Holmes isn't real but will not or cannot occupy the context within which he admits that fact, and rather persists in the fictional context to what appear irrational and/or pathological extremes, that intuitively shows that the one is not self-deceived while the other is.

While this sufficed to show that it is possible *in general* to make fictive judgments without self-deception, it did not address the crucial question of whether cases of this sort could be extended to the fictive judgments required of the successful moral fictionalist.

Joyce recognizes that successful fictive moral judgments will require a deeper level of immersion than those made when pretending for a day that Holmes existed, or when reading novels, or likely any kind of fictive judgments that we have ever made. However, he does not seem to recognize not only that this will this be very difficult, but that to the extent that one is successful, one looks dangerously like the tourist who needs psychotherapy and not like the one with ready access to her more critical context.

Arguably, the crucial obstacle to be overcome in attempting to be a moral fictionalist is not peculiarly philosophical or theoretical, but rather avoiding the awareness that one is pretending.[297] To the extent that one is aware of this fact, the fiction will not plausibly be able to serve its anti-akratic function; at least it will do so no better than a 'straight-talker' who is familiar with the importance of commitments.[298] As Joyce recognizes, one cannot enter into the fictive context on a case-by-case basis, moving from straight-talk to fictionalism and back whenever it seems suitable. This will undermine the whole point of the fiction, which, as Joyce knows, is to commit one to certain kinds of (non)actions, even if they appear irrational, or would appear irrational if one were disposed to evaluate them rationally. But Joyce's argument for moral fictionalism hangs on the ability of moral discourse to preclude such rational evaluation by means of 'conversation-stoppers' like 'It's just wrong.' The entire point of the

[297] Joyce recognizes the importance of not being aware of one's own pretense (218), but spends no time addressing the problems that arise in this vein.

[298] Below I will compare the benefits of fictionalism to those of abolitionism in the context of such an awareness.

enterprise is to keep us from acting on (what might otherwise be) our subjective preferences in a range of circumstances. Therefore we can already see how importantly different moral fictionalism will be from Holmes-fictionalism, where we were told that the Holmes-fictionalist was not self-deceived just because they could access their more critical context ‘should they care to’. This standard cannot be maintained for the moral fictionalist, who, for the fiction to do its job, must not have access to the more critical context whenever she cares to, or indeed, whenever it seems *rational* to inhabit it.

In arguing for the (possible) superiority of moral fictionalism over abolitionism, Joyce has us imagine three different people who might be tempted to steal while shopping (where the chances of being caught seem small, but in reality stealing would be a ‘practical mistake’ for all three people). One is a prudential calculator, who weighs the probable outcomes of stealing, another is a moral fictionalist and another is one who just doesn’t consider stealing because of his upbringing. I am not comparing fictionalism to abolitionism yet, so I won’t take up the details of this comparison, but rather address Joyce’s response to the following objection.

The objection is that since the fictionalist doesn’t really believe what she’s saying, the benefits Joyce is claiming for her are unlikely to be obtainable since there is nothing stopping her from entering her ‘critical mode’ while in the shop. Joyce’s answer is that there is ‘no particular reason why she might not confront [her belief that morality is a ‘load of rubbish’] while shopping’ (227). If she does, Joyce says she can fall back on the prudential calculations that are available to the first agent, which will, we assume, counsel her not to steal in at least most situations. Joyce acknowledges that having genuine moral beliefs will likely be more effective than mere fictionalist thoughts, but reminds us that even having moral beliefs is no guarantee.

I find this response far from adequate. This issue is one of the biggest problems facing his proposal, and his answer is that there is nothing to prevent us from becoming aware that morality is a load of rubbish whenever we're tempted to act (fictionally) immorally. That the fictionalist can fall back on prudential calculation[299] in such cases is no help at all, since we're wondering why we should go to the trouble of attempting to be fictionalists in the first place.[300] I think it's hard to deny that adopting a life-strategy of fictionalism involves a very significant and difficult commitment. In order to embark on it, it would be nice to think that with practice we can do better than we'd have done had we simply stopped using the concepts.

I acknowledge that it is too much to expect from Joyce that he be able to give us compelling evidence that we are likely to sufficiently often succeed in keeping ourselves from occupying our more critical context, and most importantly, when we need most not to occupy it. Still, he does not acknowledge just how difficult this is likely to be, especially if we consider that we're likely to have to explain ourselves repeatedly and convince others to be fictionalists if we're to get on honestly. But put that aside for now, and let's just imagine that there could be a lone fictionalist. Even for him, it will presumably take training and practice, requiring no small discipline, to keep his attention from the fact that he is pretending to make moral judgments. While this will be true in general, it will be much more difficult to do in the context of serious temptations,

[299] I think there's a lot more left over than 'prudential calculation' when we get rid of moral discourse, and so there is a very misleading false dichotomy here, but I won't address it now.

[300] And explain it to and convince others to be as well if we're to talk to them.

temptations of just the sort that Joyce claims we would be unlikely to overcome as 'calculators' but would be more likely to do as fictionalists.

It is worth emphasizing that the fictionalism on offer is only interesting if it can do better than nonmoralized values and rational evaluation (in someone who understands the importance of commitments). It is only interesting if it is more effective in fighting akrasia than these methods. Therefore it is only interesting where there are fairly serious temptations, for we can expect that our second-order desires, as well as practice in understanding the pervasiveness of hyperbolic discounting and the importance of commitment, even when those commitments might seem irrational at the time, can deal fairly well with mild temptations, and without the additional bothers that come with fictionalism. So to be attractive, a fictionalist *has* to be able to promise to (tend to) fend off the awareness of what he's doing when it matters. And in order to do it when it matters, one has to do it at least almost all the time. One cannot be aware of this often and then expect not to be aware of it when it is most difficult. And it will be most difficult just when one is strongly motivated to transgress, since that motivation will in effect be a motivation to realize that morality is bunk.

Controlling one's awareness and attention in the context of strong motivations that run counter to the direction of this attempted control is very difficult. In order to even begin the life-strategy of fictionalism it would be nice to think that we could expect *some* particular reason why we could expect not to confront our true beliefs just when we need not to. But by his own admission, Joyce thinks there is none. This is a very bad problem insofar as we are to take this seriously as a practical proposal. However, in the absence of any other strategy which promises the benefits that fictionalism does, we might be willing to try and see how well we can do.

Since I am not yet comparing the benefits of fictionalism to other strategies, I will allow that this seemingly daunting task can be achieved. That is, let us now assume that with practice at least most of us could come to the point at which our critical mode is in some robust sense inaccessible to us when we need it to be. That is, we have trained ourselves to the point that even significant motivation to occupy this context can expect to at least often be overcome by our having trained ourselves somehow to keep our attention and awareness away from the fictional nature of our judgments.

Now we have a kind of answer to why the fictionalist can expect not to enter the critical mode, though it is so far circular. The answer is that the fictionalist can expect not to enter the critical mode just to the extent that she is 'immersed' in the fiction, where we understand immersion in terms of the disposition not to enter the critical mode even when motivated to do so. The circularity isn't a problem per se; it simply illustrates the fact that in order to gain the benefits of fictionalism we will have to somehow become disposed not to enter the critical mode except in not only rare, but particular kinds of circumstances, such as when explicitly doing metaethics (where we have not been motivated to do metaethics so as to enter the critical mode so as to bypass morality!). This will in my opinion likely be very difficult, but we're letting that pass for now and turning to what such success would look like. I submit that it looks a lot like self-deception, so much so as to render obscure the (importance of) the distinction between belief and acceptance that, in less immersed circumstances, was fairly clear.

4.2.3 Accepting Self-Deception

I mentioned above that Joyce had borrowed from Kent Bach the idea that one could believe not-*p* while thinking and acting as if *p*. Joyce modeled the attitude that a fictionalist is to develop on Bach's idea that thinking that *p* is neither necessary nor sufficient for believing that *p*. Curiously, Joyce was not bothered by the fact that Bach developed these ideas as part of an analysis of self-deception. Bach argues that self-deception does not require conflicting beliefs, but rather that one thinks what one does not believe, or does not think what one believes, where these thoughts or lack thereof are motivated by desires that conflict with one's beliefs (Bach, 1981).

Central to Bach's account is the phenomenon of inattention. Bach holds that beliefs are dispositions, not occurrent thoughts. Occurrent thoughts can be in accord or discord with one's beliefs. When one is motivated by some desire to avoid the occurrent thoughts that would accord with one's beliefs, or to have occurrent thoughts that are not in accord with one's beliefs, to the extent that those motivations are successful, one is self-deceived. I find this account of self-deception compelling, and am puzzled by the criticism that the account is 'too unparadoxical'[301] to capture our sense of self-deception. Interestingly, Joyce does not claim that it is too unparadoxical or that it is for any other reason inadequate as an account of self-deception. He only says that while it *might be* too unparadoxical to be correct, it could 'serve perfectly well as a description of the agent participating in a fictive judgment' (197).

[301] This unfortunate criticism comes from Darwall (1988).

This is surprising. Joyce is happily conceding that if the account is not too unparadoxical—if the account is essentially correct—then fictionalists are to embark on a project of deliberate self-deception. It is surprising because just the page before he seemed intent to convince us that the Holmes-fictionalist was not self-deceived, and that the point of doing so was just to show us that fictive judgments could reliably avoid self-deception. However, here he seems to be advocating making fictive judgments on Bach's model of self-deception while leaving it open that this is a good, even correct model of self-deception. This is bad enough when restricted to Holmes-like cases, but it becomes much more serious an issue when we turn to more fully immersed fictionalists.

Joyce insisted that the second tourist, who at some level knows that Holmes wasn't real, but nevertheless would not admit as much without psychoanalysis or some such treatment was not making fictive judgments but was rather self-deceived. I want to first point out that this seems to be in tension with his own analysis of fictive judgments. That analysis was one in which someone had to have occupied a context in which they had denied some thesis and would continue to be disposed to do so were they to occupy it again, and that context is asymmetrical to the less critical context in the way described above. There was, importantly, no requirement on the relative amount of time spent in the two contexts or how reluctant one had to be to occupy the more critical context (indeed, as we've seen, building in such requirements would seriously threaten a viable Joycean *moral* fictionalism). But we were given no reason to think that this self-deluded Holmes fan had never occupied his critical context, and Joyce is explicit that *if* he were to inhabit it that he *would* be disposed to deny Holmes' reality. The problem is that he just *would not* inhabit it even though it seemed irrational or otherwise crazy not to do so.

Therefore there seems to be no clear reason to deny that he was making a fictive judgment, other than the intuition that he was (pathologically) self-deceived, and that is not the picture of a fictive judgment that Joyce wants to promote.

My second point is that the more that the moral fictionalist is successful in overcoming the problem of over-awareness described above, the more she looks like this sort of person, i.e., (pathogically) self-deceived. Joyce, after listing Bach's proposed mechanisms by which inattention is achieved in the self-deceiver, claims that the details of Bach's mechanisms are unimportant, since Bach is specifically attempting to analyze self-deception, and 'since the fictive judgment has little to do with self-deception, many of [Bach's] points, though interesting, are oblique to [Joyce's] purposes' (p. 196, fn. 26). This is of course begging the question, especially since this footnote appears squarely in the context of arguments to the effect that the fictive judgment *doesn't* require self-deception! In effect, Joyce acknowledges that some authors don't think that conflicting beliefs are required for self-deception, and that he wants to model the fictive attitude on one of these authors' analyses of self-deception, and then claims without argument that the fictive judgments he has in mind have 'little to do with self-deception'.

He might be able to get away with this if the fictive judgments he were recommending were such that in making them one had ready access to one's more critical context. It is after all just this ready access that seems to distinguish the third from the second Holmes fan, and make the one appear self-deceived while the other is not. But we've seen that for moral fictionalism to be worth attempting, we cannot have ready access to our critical context whenever we want to. In fact, we cannot have access to it *whenever we (would) think it rational to occupy it*. If we could, we would be no better off than the rational evaluator who appreciates the value of commitments. Again, the *whole point* of the fictionalism is to *preclude* rational evaluation of certain categories of actions. We cannot then step out of our fictional context whenever it looks like rational calculation would counsel us to do differently than would our moralizing.

I said above that Joyce could give a (non-viciously) circular answer to the question how the fictionalist could expect her critical context not to pop up at inopportune times. But Joyce can actually do better than that. He can employ Bach's mechanisms of self-deception, especially those of evasion and jamming.[302] Evasion involves the simple avoidance of the thought that accords with the belief that one is motivated not to believe. It is the '[t]urning of attention away from some touchy subject' (1981, p. 360). Normally such turning of attention is not self-deceptive, and is often used to 'reduce awareness of pain or nausea' (360). But in the self-deceiver, unlike someone merely avoiding an unpleasant thought, avoiding the thought *of p* is motivated specifically so as to avoid the thought *that p* (360). 'Whereas rationalization lends itself to deceiving oneself in the first place … evasion best serves to preserve the state of being self-deceived' (361).

Jamming is the technique of having the thought that not-*p* as a means of avoiding the thought that *p.* If the thought that *p* happens to occur to him, he can have the thought that not-*p*. If the unwanted thought of *p* arises, out of a motivation that not-*p*, 'he focuses his attention on what it would be like if not-*p* ... [and may] perhaps even go so far as to [act] as if not-*p* were the case … Jamming is particularly effective in self-deception about

[302] Bach's other mechanism, rationalization, doesn't seem well-suited to self-deception that one has *deliberately* undertaken.

one's feelings or motives' (361 – 2). The technique of jamming seems particularly suited to a fictionalist.

Bach also distinguishes between the different strategies that will be required to be or remain self-deceived depending on whether one's claims are challenged by others. If one is not challenged, then one merely requires a lack of awareness, which can be accomplished by simply being motivated to believe not-*p* while avoiding the thought that *p*. If one is challenged, then one also requires the motivation to avoid the awareness that one is motivated to avoid the thought that not-*p*. Since the vast majority of cases of self-deception are not deliberately undertaken, there are normally important differences in strategy depending on whether one's thoughts are challenged, either by oneself or others. Most obvious is that rationalization is only needed in the cases where one is challenged.

However, the fictionalist does deliberately undertake her project. While employing moral discourse, she has to have the motivation (perhaps eventually, just the habit) to avoid the thought that morality is bunk, as well as motivation to avoid the awareness that she has that motivation. We can expect that she will rarely be challenged at the metaethical level, i.e, challenged about whether she really believes in morality at all, but she can expect to be challenged once in a while on some of her specific moral claims, like for example why stealing is inherently wrong. Since by hypothesis she has no justification that she believes for this claim, she will have to invent one, or borrow justifications from ethicists who have defended claims such as these. Perhaps she will plump for some kind of moral intuitionism. All this will have to be done without becoming (too) aware that in reality she has chosen to fictionalize certain moral injunctions for their probabilistic contribution to her long-term preferences.[303] So the fictionalist will have to employ rationalization as well, though not with respect to her belief in morality in general. She will however have to give rationalizations as to why some particular moral judgments are correct, all without being (very) aware of the fact that it is all a pretense.[304]

This points us back toward the psychological implausibility of the program, but I want to say that even if somehow the evasion, jamming and rationalization can be accomplished, then it will have been accomplished at the price of employing precisely the mechanisms that Bach argues are constitutive of—or at least crucial to—self-deception! So we can see that on Joyce's own account of fictive judgments, fictive attitudes and belief vs. acceptance, it will not do to glibly assert that fictive judgments have little to do with self-deception. In fact, the more we look into the account, taking his examples and inspirations seriously, the more they look like they have quite a lot to do with it. This is especially evident when we reflect on the fact that unlike the seemingly harmless examples of fictive judgments we were presented with, our moral fictionalist cannot, on pain of undermining the entire project, occupy her critical context even when she would judge it rational to do so were she to engage in such an evaluation.

This is a tremendous and fundamental difference between the fictionalism that one employs in order to spend a jolly day pretending that Holmes existed and the sort that is employed in order to reap significant anti-akratic benefits by means of committing oneself to certain modes of conduct as a life-strategy. First, the one is *relatively* easy to

[303] I'll say more on the problems of justification that confront the fictionalist below.

[304] It might be thought that the moral intuitionism route avoids the need for (much) rationalization. But many people find this position untenable these days, and there is little reason to think that anyone to whom one would have to bring up intuitionism in the first place would let the matter rest there.

pull off while the other might be nigh-impossible. Second, to the extent that one is able to pull it off, one appears self-deceived in so doing, and this self-deception can potentially be quite a bit more pathological than in the Holmes case, an issue I'll address below.

Earlier I mentioned that Joyce thought there was a range of extents to which one might be immersed in the fiction, starting at reading fairy tales for a short time and ending in the kind of immersion that will be required of the moral fictionalist. What he failed to see is that his analysis of belief vs. acceptance, as well as that of self-deception, makes an implicit appeal to what we might call the availability of one's critical context. The second Holmes fan could not or would not inhabit it without extreme measures being taken, while the third had easy access to it and simply chose not to inhabit in order to have fun. The spectrum of possible levels of immersion then appears to correspond at least roughly to a spectrum between pure belief and pure acceptance. The latter involves immediate and effortless access to one's more critical context, while the former involves a lack of that critical context altogether. However, if one's access to the critical context is sufficiently restricted, then the *practical* difference between this level of 'acceptance' and belief is far from clear. It is far from clear that is, that the person at this level of immersion is not for all practical purposes systematically self-deceived.

To be clear, I don't think that someone undertaking the fictionalist project is likely to be systematically self-deceived. I find it vastly more plausible that they will fail to reach the level of immersion at which this would be true, but in doing so will be largely if not entirely wasting their time (in comparison with other options, which I will discuss below). Further, I am not claiming that when the fictionalist does enter their critical context that they somehow remain self-deceived or that they will not be able to recognize that they have been while immersed in the fiction. I think they will be able to do so and that this will tend toward their not being so in the future, thereby underminding the project. I think that they will be especially likely to judge that they have deceived themselves when and if in making their fictive moral judgments they do something that they later judge to have been a serious mistake. If and when this happens, there will be particularly strong pressure *not* to avoid a rational evaluation in like circumstances in the future. As I've said, I think the fictionalist proposal is highly problematic on a number of fronts, and is generally psychologically unworkable.

But we are still accepting for argument's sake that it is workable at the psychological level, i.e, that one can generally avoid awareness that one is pretending. Now let's further suppose for the sake of argument, i.e., for the continued sake of making the strongest possible fictionalist proposal, that we're not worried about the self-deception problem per se. Joyce might be able to make a case that even if one is self-deceived, this kind of deliberate, rationally calculated self-deception is worth the benefits it promises. Or he might be able to argue that they are not 'really' self-deceived (they just really look like it). Now I want to evaluate how the successful moral fictionalist, whether self-deceived or not, fares in comparison to an enlightened abolitionist.

4.2.4 Fictionalism vs. Abolitionism

A Real Opponent

I think the considerations I've provided above expose very serious problems with normative moral fictionalism. As it happens however, I think that Joyce's motivations for fictionalism, and his recognition that it will have to be a life-strategy, are well-grounded, and so I think that the problems will not be avoidable for any plausible moral fictionalism. But now I want to grant that all of these problems can be overcome. I won't grant that it will be easy—there will be a cost to implementing the program successfully—but I will grant that it can be done for the sake of argument in order to bring out yet further problems with it even if all of the above problems can be solved. These further problems together are to my mind far more than serious enough such that we are better off avoiding them by simply getting rid of moral discourse and developing the alternative discourses we've already got.

For simplicity's sake, for much of the following discussion I will treat the fictionalist as, for all practical purposes, believing her moral judgments. This is for two reasons. The first is that, as I have argued above, in order to get the motivational benefits promised, the thoughts will have to be stable enough not to be undermined by her critical mode popping up when she's (sorely) tempted. Thus her more critical context has to be *practically* unavailable (or close to it). It will be more available than in a believer (since they don't have one), but close enough to unavailable for my purposes. The second reason is that many of the arguments I develop below about the downsides of fictionalism will carry over directly to the true believer(s) in morality. In presenting the downsides of successful fictionalism, I will also be presenting the downsides of straightforward moralizing. Indeed, the weakest aspect of Joyce's fictionalist proposal is that he sees no downsides of moral thinking aside from their problematic metaphysical commitment(s).[305]

One of the biggest problems with Joyce's comparison of fictionalism with abolitionism is that the only way he knows how to 'step out' of moral thinking is to ask what is instrumentally (most) profitable (181). Thus if we are trying to find out whether it was right or wrong for the British to commit genocide on the Tasmanians, we need to know whether it was good or bad *for the British.* There are two very serious problems here. The first problem is the idea that the only way to understand our (nonmoral) commitments or values are in instrumental terms. But the commitment(s) I have that would prevent me from accepting five dollars for pushing a button that would randomly kill some child across the globe aren't (only) instrumental in securing me benefits specifiable independently of the emotions I would be disposed to feel if I were to do so, which emotions form the foundation of the commitment itself. Many of our values are like this.

The second problem with Joyce's analysis of the Tasmanians and British is that it is an example of the fact that Joyce often analyzes instrumental benefits in self-interested terms. Though he elsewhere describes the benefits of moral discourse in terms of their

[305] To be fair, he does acknowledge Hinckfuss's criticism of morality, which (for present purposes) we can understand as the criticism that if the sense of 'must-be-doneness' is attached to the wrong kinds of actions, then that will be very bad (and has been very bad in the past). Joyce's response is that this shows that it's important to moralize only the 'already useful actions' (181). I'll address the problems with this response below, beginning with the idea of 'usefulness'.

conduciveness to our long-term preferences, in the Tasmanian genocide example and elsewhere he asks not (consistently) about British long-term preferences but rather about what is good or useful to the British. Though he alternates between interests and (considered, fully-informed) preferences, the focus on interests seriously weakens his (fictional) abolitionist. Having made both of these moves, Joyce makes the case that it was likely bad for the British because *in general* harming others' interests tends to harm our own. That is, the British *policy* of initiating violent hostility was a bad policy since this policy 'lowered the chances of being sought out by others with offers of mutually useful ventures' (182).

Joyce acknowledges that this appears a 'heartless logic', but it's the only way he says he knows how to step out of morality. He does not claim that this sort of thinking needs to figure in anyone's deliberations in order for it to nevertheless be the rational justification for the policy of not initiating violent hostility. In fact, he thinks it is plausible that the best strategy for pursuing a cooperative policy is to 'cultivate a concern directly for one's fellows' (182).

Of course most of us *already have* concern for our fellows, though it has been cultivated and pruned in us to very different extents and in very different ways. Whatever instrumental benefits are likely to come from future cooperative ventures have no special rational status that prioritizes them over the desire (commitment) that most of us have not to be a party to genocide--independent of the possibility that committing genocide could reduce future profits. It is an important mistake to suppose that the desire not to murder people—especially an entire people--depends for its rational justification on whether that desire is part of a ‘cooperative’ or other strategy that is likely to bring us some independent benefits in the future.

I want to be clear about what’s happening here. Joyce is appealing to the strategic value of the moral sentiments because he wants to be able to explain why those who don't have them should get them and why those who do have them should maintain them. I have no problem with that kind of project in general; there might be many such reasons. What I am objecting to is that what we call a direct concern for our fellows *requires* the kind of justification Joyce provides it. That is, if it turned out that no such strategic justification could be maintained (which in many cases seems likely), it would not follow that we have no reason not to commit genocide--even after we’ve rejected morality.

Joyce seems to view stepping outside of morality as synonymous with stepping outside all concerns but self-interested or instrumental ones, where these latter concerns can be specified independently of one’s emotion-backed commitments. I see stepping outside morality as stepping outside the conceptual framework involving categorical obligations, and/or duties and nonrelational values. On Joyce’s view, once we have stepped outside morality, should some such justification in terms of the independent usefulness of sentiments against genocide fail, then we would have no real reasons to hang on to those sentiments. I am saying that those sentiments are just as much a source of reasons as one’s desires for the strategic benefits that such sentiments might or might not conduce to. Wanting the strategic benefits that (might) come from a disposition to want not to murder people has no intrinsic rational priority over wanting not to murder people.

I will not speculate as to what sort of values the Victorian British had (though I suspect that their values ran contrary to murdering people by the thousands), but I do feel confident that my readers have strong commitments not to actively support genocide at least in part out of concern for the people involved. On the conception of practical rationality and justification I endorse, this commitment only requires justification if there are other comparably strong values with which it is in conflict. We might indeed have values that conflict with not killing innocent people in certain cases. Presumably the Victorian British valued expanding their empire and/or valued other goods they might acquire by killing the Tasmanians. Although killing them *might* threaten these general sorts of goods in the future, i.e., those they could acquire by means of cooperative strategies that might be undermined by initiating hostility, it also violated and threatened whatever values they had not to murder people. And though one can *bolster* this latter value in terms of its contribution to the former, one might just as well bolster one's valuing certain kinds of material goods in terms of its contribution to not killing people. Both sorts of projects are fine so far as they go, but *giving up categorical reasons doesn't render either value dependent on a justification in terms of the other*.[306]

For this and other reasons, I think that Joyce has pitted an abolitionist straw-man against his fictionalist. The abolitionist need be no 'prudential calculator', and he needn't cultivate concern directly for his fellows *solely* on the basis of a judgment to the effect that this will likely bring him instrumental benefits on average and in the long run. We *already* have patterns of concern, and some of these concerns might be in need and

[306] If I emphasize and put this point multiple ways, it is because I think it is of central importance, and the kind of mistake Joyce makes here is extremely common.

susceptible of justifications, including of the sort that Joyce provides. But our desire for instrumental benefits of the sort to be gotten from future cooperation are rationally on all fours with our desire not to murder people.[307] We need no general justification of the latter in terms of the former.

The abolitionist A I want to pit against Joyce's fictionalist F has not only interests but values. He not only has values but he understands the great importance of commitments in securing those values against specious rewards. He also understands that commitments, important as they are, sometimes warrant exceptions. He knows that there is no decision procedure for how to decide when to violate a commitment; he knows that employing good judgment and developing good habits and (self-) perceptual abilities and being vigilant against self-deception are all important but not sufficient for living a good life.

If we start with a preference-satisfaction view of rationality[308], then Joyce and I are agreed that whether we should place our bets with A or F depends (ceteris paribus) on which of these strategies is more likely to conduce to their satisfaction. F's task will be to fictionally moralize certain of her existing commitments (as well as perhaps create new ones to moralize) so as to render them more impervious to specious rewards. F understands all the things I listed above that A understands about the importance of commitment and so on. But F (we are liberally granting) successfully keeps this understanding and its role in her having decided to become a fictionalist out of her awareness except in 'philosophy seminar' contexts, while A can in principle call upon

[307] On all fours in terms of their intrinsic requirement for justification that is, not in terms of the strength of our commitments thereto.

[308] I think it is not very important for my arguments in this section whether this is of Joyce's preferred 'fully-informed' variety or my 'simple' view or some other.

any of this understanding and/or justificatory style at any time. Let's now examine the relative advantages that F can expect to have over A.

Anti-akratic[309] *benefits of moralizing*

As we've seen, F's strategy is to moralize her values in order to render them more immune to specious rewards. To do this, according to Joyce, is to put them in a category of things that 'must be done'. But we've seen that F doesn't believe that they're in this category, but rather only 'thinks' it. Left at this level of description, A (here for 'Ainsliean' as well as 'Abolitionist') can (but need not) do the same thing. A recognizes that his values are threatened by specious rewards and so can consciously train himself, or consciously prevent himself from becoming untrained, to gather the relevant action-types together under principles, and to act on those principles rather than evaluating candidate actions on a case-by-case basis. A can train (or not untrain) himself to not genuinely consider them options, to rule them out as unavailable since they violate the relevant principle.

The crucial differences between F and (this version of) A is that F is much more committed to deflecting her attention away from, or otherwise precluding awareness of the fact that the reasons for acting on her commitments are grounded in the preservation and/or protecting *her* (as opposed to intrinsic) values. The specific manner in which F does this is to systematically think and talk in moral terms, as well as engage both self-

[309] I will for simplicity's sake consider 'anti-akratic' mechanisms those that tend to cause actions in accord with one's long(er)-term preferences, and 'akratic' those which tend to cause actions in accord with one's short(er)-term preferences.

and other-directed moral emotions. On the other hand, while A might practice some evasion and jamming here and there, if called upon to explain what he means in saying he can't do something, he will not give moralistic rationalizations or explanations. He does not generally think or talk this way. He has embarked on no training program to keep himself from being aware that they are *his* values that he's protecting, and so in his case there really is no particular reason to think that he won't enter what we could call his 'critical mode' when tempted. But having entered it, he might quickly deflect his attention again from the tempting action, as he has trained himself to employ this as an effective anti-akratic strategy.

In portraying A as a prudential calculator (though perhaps one who recognizes the importance of commitments and therefore might have learned to cultivate fellow feeling), Joyce assumed that such a person could not have access to the thought that something must (not) be done. But this is not correct. The difference between F and A is not (mostly) a matter of F having access to thoughts that A does not have access to, but rather A having access to thoughts that F does not.[310] A can even think something is immoral if he wants to or thinks it would be helpful. But since this is not his typical strategy and he has not undergone a process of training himself to keep his awareness from the fact that it *is* a strategy, the motivations and emotions that are associated with this thought will not be as *stable* as they will be for F.

[310] Putting this in terms of 'having access' can be misleading. First, saying that F does not have access to the thoughts suggests that she is or would be consciously attempting to access them but is unable to. Understood this way, it is natural to think of A as the one who does not have access to the thoughts, but only if we understand that as having access to them in a stable way, i.e. one that is not undermined by the co- or quickly succeeding presence of thoughts to the contrary. Thus F has *stable* access to the moralistic thoughts because she 'lacks access' to (has trained herself to evade and/or jam) the thoughts that would undermine them.

Joyce says that F will have the motivational benefits of thinking that she is (morally) *reprehensible*, that she (morally) *deserves* punishment if she does wrong (217). A can think these things too, as we've seen, but they are unlikely to 'stick'.[311] His mind will tend to (sooner or later) think of his mistakes in terms of violating his own commitments and values, which he might care about in their own right, or because they conduce to other values or desires. It would however be a serious mistake (which Joyce makes on p. 226 and elsewhere) to suppose that A can't feel intense, long-lasting guilt at having violated his values, e.g. at having done something foolish and harmful to himself or others, especially those he cares about deeply.

We have seen that one source of F's advantage over A is that F (by design) does not have access to her metaethical beliefs in the way that A does, which can be understood as A not having access to the kind of stability of certain thoughts that F does. But while Joyce doesn't neatly distinguish between them, there is another and different source of F's supposed advantage, and that is the peculiarly *moral* character of the thoughts and emotions that F is disposed to (stably) possess. For, as Joyce acknowledges, the 'must be done' thought needn't be moralized. It can be the thought that one 'must do fifty sit-ups' every day (in order to get or stay in shape) even though one knows that missing a day or two here and there won't be a big deal (215). As we've seen, A can think this and derive some motivational benefits from it, though he won't likely be disposed to make great efforts to hide from himself or others the value of such

[311] Except in rare circumstances, he is also unlikely to engage in public moralistic justifications or condemnations of his or others' actions. Since they tend not to stick, and for other reasons, A is unlikely to think them often if at all. My point is just that he can if he wants to. Also, Joyce does not add the word 'morally' before these terms as I do in parentheses, but I think it's important to do so since I think there are perfectly good and nonmoral ways of understanding them (though admittedly, most people understand them morally).

thoughts. But even if A did go to such efforts, this wouldn't make him a *moral* fictionalist insofar as he does not moralize the sense of 'must be done'.

So F's advantages (and, as we'll see, disadvantages) are not only a matter of successful maintenance of the thought that something must be done, but of the moralizing character of those thoughts, with the corresponding moral emotions. When Joyce says that F can (stably) say and think to herself and others that something is *immoral*, whereas A can 'only' say that it violates his values, we should understand at least part of whatever extra oomph there is in the former as coming from the moral emotions, and not the sheer thought of 'must-be-done-ness'. And while I have said that there is nonmoral guilt, I grant at least for the sake of argument that susceptibility to the moral emotions provides an additional and powerful source of motivation.

Akratic costs of moralizing

The anti-akratic mechanism associated with practical clout (thoughts of the sort that x must/must not be done) as well as the one(s) at play in being susceptible to the moral emotions each have a corresponding akratic mechanism as well, which Joyce does not mention. First, it's worth reminding ourself about how the former works. The clout mechanism provides anti-akratic benefits by considering all actions categorized in a certain way as either required or prohibited. Once an action is so categorized, the mechanism works by precluding rational evaluation about whether to do it. Performing an action that one has categorized as forbidden weakens the ability to organize willpower for related actions in the future, and it is just this ability that is the 'side-bet' involved in each choice of whether or not to violate a rule. The function of the side bet is to raise the stakes for each individual decision such that at the limit, all the benefits of future 'cooperation of intertemporal selves' rests on each individual decision.[312]

Now the moral emotions seem to involve a distinct mechanism, even if the concept of practical clout contributes to their intensity, stability and/or character. For plausibly it is the relative intensity and stability of the moral emotions that explains at least much of their motivational power. That we can anticipate feeling terrible moral guilt lasting for years as a consequence of certain actions surely has much to do with our motivation not to do the things that would engender those feelings. The action-tendencies associated with guilt can motivate us to go to great lengths to repair damages we think ourselves morally guilty of having caused.

Joyce has argued that F is likely to be less akratic by being less (or not at all) susceptible to rationalizing self-deception, due to the 'conversation-stopping' aspect of moral discourse, and further that F is more stably motivated to do the (fictionally) moral thing due to her fictive engagement of the moral emotions associated with her fictive moral judgments.[313] So far this looks good for F. We have identified another mechanism to underwrite the claim that (pretending) to believe an action *morally* obligatory is more motivationally effective than the kind of principled obligatoriness that A has access to. The moral-emotional mechanism adverted to here would work by virtue of the fact that F

[312] I am obviously assuming that Ainslie has roughly the right analysis of this phenomenon. This strikes me as warranted because 1) it is to my knowledge the best analysis out there and 2) it supports Joyce's view by providing him with a vastly more developed model of the importance of such principled thinking to the working of the will than the intuitions on which he relies. It's just that the model doesn't *only* support his normative views regarding willpower.

[313] Though the moral-emotional mechanism is not clearly distinguished from the one involving clout, or rather the former is thought to follow directly from the latter, which is far from clear to me.

would expect to feel especially bad and/or for an especially long time when she perceived herself to have done (moral) wrong[314] in comparison with A who will not anticipate such bad feelings lasting for such a long time. Something along these lines seems to be the best way of making out why we might expect F to do better than A.

However, the increased painfulness of the moralized emotions is a double-edged sword. Feeling morally guilty because one has violated one's moral values requires a perception or judgment that one has done so.[315] We can suppose that to the extent that a moralizer has had this perception and consequent guilt in the past that she will be motivated to avoid this perception more strongly than if she were restricted to nonmoral guilt. As should be familiar to Joyce and everyone else, there are two ways to avoid the perception. One is not to do what one perceives as wrong, and the other is not to perceive what one does as being wrong. For some people, the greater the painfulness of judging that one has done wrong, the more they will be motivated to avoid developing any better self-perception than they already have, and in fact they will be motivated to have less. The more self-aware one is, the more one will be unable to avoid the recognition that one has failed to meet one's moral standards, which will cause pain and motivations to repair damage and not to do wrong again. The less self-aware one is, the more one will be able to avoid this recognition, and those who can avoid it will (tend to) do so, since the recognition is painful and people are in general motivated to avoid pain, in rough proportion to its intensity and imminence.

Indeed, the *immediacy* of emotion upon the relevant perception is also double-

[314] A and F can both do wrong and think that they do wrong. F can stably think it is morally wrong whereas A cannot. It is very common for writers to simply assimilate 'wrong' into 'morally wrong' but there are lots of nonmoral ways of being wrong, including violating one's (nonmoral) values.
[315] The perception needn't necessarily be consciously available.

edged, i.e., akratic as well as anti-akratic. A central pillar of Frank's argument for the evolution of moral sentiments is that the sentiment is present at or before the time of the specious reward with which it competes. Beginning to cheat or contemplating cheating (broadly construed) causes aversive feelings right away, and these can compete with the rewards of cheating much more effectively than if one only anticipated those feelings in the future. But this same feature makes for very effective motivations to avoid the self-perceptions that would cause those very feelings. So the increase in effective commitment F gets by her susceptibility to moral emotions depends very heavily on her prior commitment or disposition to be relevantly self-aware. Failures at self-awareness not only result in failure to receive anti-akratic benefits, but tend toward *more akratic* behavior than would be expected without the moral emotions.

Now let's return to the mechanism involving conversation-stopping principles. We have already seen in Section 1.2.2 that such a strategy motivates self-misperception by magnifying the importance of lapses. If one has a lapse and perceives it, there is a potentially great loss of the ability to organize willpower in the future. Therefore one has a motivation not to perceive the lapse, as well as motivation not to catch oneself ignoring it, for to do so damages one's credibility with oneself. In the intrapersonal realm, such processes can lead to failure to save money or lose weight despite 'strict' budgets and diets. In the interpersonal or political realm it can lead to systematic deception despite a strict policy of never lying, or systematic usage of aggression and terrorism despite strong beliefs which condemn it in the harshest terms.

I venture that the operation of all four of these mechanisms is both intuitively very common, especially in 'seriously' religious and moralistic people, and based in sound principles of motivational psychology. Those basic principles predict that strong disincentives to judge oneself to have acted wrongly will tend to motivate people not to do so. Borrowing some terminology from Freud, those in whom the ‘reality principle’ is relatively strong, i.e., those who are disposed to perceive themselves accurately in the relevant respects, will have to go the route of avoiding the behavior that causes the judgments, while those in whom the ‘pleasure principle’ is relatively strong (which could be just to say that their ‘reality principle’ is relatively weak) will go the more direct route, unaware (whether blissfully or not) of having chosen any route at all, which strictly speaking they might not have. The same phenomenon can be expected to occur--and I think it should be a truism that it does occur--with respect to other-directed moral emotions. That is, the more (immediately) painful you make it for someone to realize or admit that they have done wrong, the more you incentivize them to not realize or admit it. For some, this results in more effective motivation to stop doing wrong, while for others it results in more effective motivation to not realize it. In general, it is almost certain that both processes operate in almost everyone, though to varying relative extents.

Self-awareness as solution to akrasia threatens fictionalism

If the would-be fictionalist acknowledges these competing mechanisms, and further agrees that they tend to operate differentially depending on a person’s level of self-awareness, then perhaps she will only recommend fictionalism to those who are fairly self-aware. Perhaps she has the conceit that the sort of folks who are likely to entertain fictionalism as an option are already error theorists and therefore philosophers and for some combination of these reasons are likely to be (much) more self-aware than average, and so it should work for them by and large.

Suppose we allow them this conceit; if it is true, it threatens to spoil everything. For in what does this sort of self-awareness consist? It seems to be something like the tendency to perceive, whether one wants to (at the moment) or not, what one is 'really' up to. It is a tendency to adopt a critical stance toward oneself and to evaluate what one is up to, specifically to ask oneself questions of the general sort, 'Am I really just kidding myself here?' and 'What am I really doing and why?' and so on. You can see where I'm going, and perhaps the fictionalist will think it unfair. Joyce might respond that this kind of critical stance is not the kind he's saying that one adopts when one steps outside the fiction. He would likely say that he was careful to distinguish between that critical stance and the kind that one could adopt within the fiction, wherein one could ask the sorts of questions above, but while remaining completely within the fiction, asking them very much like the way that a true believer would, though with perhaps not as much commitment. After all, the believer cannot step out at all, while the fictionalist can, so the latter cannot expect to get quite the motivational benefits of the former, but while she is in it, she can be critical of her own motives in the way that I've suggested is associated with or constitutive of self-awareness in much the same way that a genuine moralizer can. Her critical tendencies needn't result in her being always bounced out of the pretense.

I find this response dubious. I see no particular reason to suspect that these kinds of critical tendencies, the one associated with self-awareness and the one associated with belief as opposed to acceptance, are really distinct in the way they would need to be.

And even if they are neatly distinct, I don't see why one would think that one can deploy the first to a large degree while keeping the other one quiet except when one consciously decides to put on the philosopher's hat. Self-awareness of either kind plausibly ain't like that. You don't get to choose so easily when and in what sense you perceive what you're really up to (and that's a good thing!). I think it's quite plausible that the more self-aware one is, the harder it will be to stay in the fiction when it really matters.

And when it matters is when it's hard. When it matters is when temptation is strong, and when the anticipated or actually felt moralized guilt is intense and persistent. These are the times when all the trouble of becoming and being F is supposed to pay off. If when the going gets tough, F has access to the thought 'It's only a movie' (so to speak), she's not getting her money's worth. So I think that if we agree that it's important for F to be generally committed to being self-aware for the project to have a good chance of success, we're making it awfully difficult to grant what we're trying so hard to grant for the sake of argument, which is that the project is psychologically realistic.

But let's redouble our efforts in favor of the fictionalist yet again. Let's grant for the sake of argument that F is self-aware in the right way(s) but not in the wrong ways and so can remain quite within the fiction when temptations are not only strong but seemingly rationally warranted (that is, they would seem so if one could step outside the fiction) and when strong moralistic guilt and anger are weighing on her. This marvelously managed self-awareness makes the moral emotional mechanism and the principle-based mechanism work in her favor rather than against her. Now I want to focus on the latter of these.

The costs of the benefits of willpower

As we've seen, principles (rules) motivate misperception, since their perceived violation threatens loss of future willpower. But F is quite self-aware, and so is less disposed to (be able to) ignore her own lapses. To the extent that she does in fact lapse, then given that she's made a large side-bet, she stands to lose a lot of willpower in the future. In this case, she will likely be worse off than had she not bothered with becoming F in the first place. So she will have to lapse very rarely. She will have to do a very good job at abiding by her principles given that she will tend to perceive lapses when they do occur and that will be very bad. In this circumstance, has F finally arrived?

No. For this just runs her up against the *other* downsides of willpower. The downsides she has escaped (by stipulation) are those associated with the magnification of lapses and consequent motivation for misperception. These had akratic tendencies insofar as perception of lapses tends to weaken willpower and the motivation not to perceive them tends toward the undermining of the benefits willpower is to procure, even if we don't feel our wills have been weak. For simplicity's sake, we can think of these downsides as those that undermine the effectiveness of the will. But there are other, at least equally serious downsides, which are the result (or constitutive) of the will being *too* effective. And in protecting F from the former we have made her more vulnerable to the latter.

These downsides are that principles serve 'compulsion range' preferences at the expense of longer-term preferences and overshadow 'goods-in-themselves'. In order for

F to avoid the interpretation that she has lapsed, i.e., violated her moral principles, she is prone to being over-cautious. At the extreme, she may resemble Kant, who, as a consequence of his (in)famous view that the will should operate according to an absolutely inviolate moral law, proscribed lying in any situation whatever (this was perhaps not even the most ridiculous thing he proscribed). Reliance on principles 'can make large categories of differential reward hinge on decisions of little intrinsic importance.'[316] This feature underpins the fact that not only is it the case that lapses reduce one's ability to follow rules, but non-lapses reduce one's ability not to. Therefore not only can we feel that much is at stake in decisions of intrinsically little importance, but we can be led to do what is (radically) against our long-term preferences in decisions of great intrinsic importance. Sometimes lying is a very good idea. And while most of my audience will have many fewer occasions in which stealing is a very good idea, others do have them, and adherence to a moralistic principle that condemns it absolutely reduces one's ability to recognize this. Examples could be multiplied indefinitely.

The downsides of principled thinking arise from the same source as their motivational benefits, which is the deflection of attention away from our subjective preferences, and the rewards, or benefits, that we seek by our actions. The compulsion range preferences lie between our shorter-term interests and our largest, longest-term preferences. In having a principled rule like 'lying is (just) wrong', we effectively bundle together all instances of lying into a category which requires no further evaluation, or consultation of our preferences. This helps us to escape rationalizing self-deception but also prevents us from making rational judgments that take into account the specifics of

[316] Ainslie (2001, 148).

the situation. It prevents the information of our affective systems from informing our behavior.[317]

Even if F doesn't reach such an extreme state, the point is that the commitment to absolutist moral principles has the potential for serious pathology. Ainslie cites Paul Ricoeur's observation (alongside those of several philosophers, authors and psychotherapists recognizing essentially the same problem) that the freedom of our wills is undermined not solely by sinful temptations but by 'moral law, through the "juridization of action" by which "a scrupulous person encloses himself in an inextricable labyrinth of commandments"'.[318] Ainslie also notes that existentialism was in significant part born of Kierkegaard decrying the attack on the 'vitality of experience' entailed by Kant's and Hegel's insistence that all actions be taken in accordance with universal rational principles (144). Nietzsche (1888) (in)famously perceived that 'convictions are prisons'.

If Joyce and I are right, then moral convictions are prisons *by design*. Their *whole point* is to prevent freedom and flexibility with respect to actions falling under certain descriptions or categorizations. Joyce rightly sees the motivational benefits to be had by placing ourselves (or, in the normal case, being placed) in such prisons, but it is odd that he doesn't see much in the way of downsides to being in prison. Of course he *could* argue that the benefits of being in prisons *of our own creation* outweigh the costs, but what's strange is that he *doesn't*. He doesn't seem to perceive the need to do so, assuming that if we moralize what is 'already useful' that the long-term benefits of being

[317] This is of course too simple. Our affective systems are at play in motivating us not to violate the principle as well. So it would me more accurate to say that it drowns, stunts or blocks other forms of affective information due to the legalistic nature of the principles.
[318] Ainslie (2001, 144). Ricoeur (1971, 11).

in the right sort of prisons are likely to outweigh the costs.

But even if we put aside for the moment the (enormous) problem that we might be moralizing the *wrong* values, moralizing ones that seem good is deeply problematic. Even such prohibitions as seemingly 'useful' as that against killing innocent humans without their consent can lead to potentially very undesirable policies, such as opposition to abortion under any circumstances.[319] And when this moral prohibition runs up against the injunction against violating a woman's right to do whatever she wants to with her body (another very useful-sounding policy)[320], the stage is set for intense and saddening confusion on all sides, not to mention angry, moralistic divisiveness.

Even with such eminently sensible-sounding policies, there is great value to understanding—*and having access to this understanding*—that such policies are in the business of protecting things that we care about. The most obvious benefit is that such understanding can help get us traction on how to make our policies better and in what kinds of circumstance to make exceptions to them. But F's whole strategy is not to make any exceptions. The strategy is, I think it's fair to say, a kind of mindless moralism. True, the thoughts leading to deciding to adopt the strategy are far from mindless, but the goal is to attain a circumstance in which one does not think about whether or not to make exceptions to absolute moral rules, one simply declares them wrong and acts in accordance with them.

It sounds tendentious to describe the project as an attempt to attain a mindless moralism until we're reminded that thinking about whether to make exceptions opens the

[319] And moral injunctions against killing any animals, or things that can feel pain, or living things of any sort, just get increasingly insane.

[320] It also sounds good to say that parents have the right to raise their children however they see fit. But it seems we have to have exceptions to this policy too.

door to rationalizing self-deception and akrasia. So mindlessness of a sort can certainly be a bulwark against that.[321] And while I think Joyce is right to diagnose the anti-akratic function of morality as largely consisting in a certain kind of unquestioning adherence to (moral) principles, there are serious downsides, even if we tend to moralize roughly the right sort of values. Which we might not.

This worry brings us to Hinckfuss's (1987) criticism of morality and Joyce's response to it. Hinckfuss noted that there have been innumerable atrocities not only carried out by people who firmly believed themselves in the moral right, but often done in the very name of their moral values. Joyce responds that this just shows that it's important to moralize the 'already useful actions' (181), i.e., it's important to fictionalize the *right* moral values. However, what the right values are is not something to be gotten squared away in an afternoon or fortnight prior to setting about moralizing them or rendering one's existing moralizations fictional. When we recognize that what would make some values 'the right moral values' goes far beyond what is independently and instrumentally useful to us, this is even clearer.

What the 'right values' are is something we are always—if we are lucky—in the process of discovering and/or creating, both as individuals and societies. Those values are such that there will almost always be appropriate exceptions to make to any policy.[322]

[321] In fact, the mindlessness involved here is essentially a lack of self-awareness. Calling someone mindless or saying they lack self-awareness sounds insulting. But it is not for nothing that 'The Fall' depicted in Genesis (and in the book by Camus, another existentialist) is about the potentially very high costs of self-awareness.

[322] Perhaps there are some few policies that, at least for many people, can be expected to pay off in the long run without exceptions being made. Joyce has probably settled on the most promising one in his example, the prohibition against stealing. Especially when we restrict ourselves to a conception of stealing of the sort that involves taking things from shops or homes without paying for them, it is unlikely that Joyce's readers will have much more to gain than Hume's 'gewgaws' at the risk of potentially severe and/or long-lasting consequences. So perhaps this and some other very limited number of prohibitions or commands

Moral thinking tends to hide this from us by deflecting attention from what we care about for its own sake onto rules or principles that, at their best, tend to keep us from acting in ways that we are likely to regret (or would be if we understood what we were doing better). At their worst, they generate seriously pathological behavior and prevent us from gaining self-knowledge, knowledge of what it is that we *do* care about and thereby what other things we *should* care about.

4.3 Conclusion

I find it very unlikely to be coincidental that both willpower and morality have traditionally been thought of as unmixed blessings, aside from a handful of philosophers and psychotherapists beginning in the 19th century. Joyce has done a fine job in analyzing the benefits of the one in terms of the other, but in failing to examine the other side of the coin[323], his case is much weaker than it might seem. Employing Ainslie in this connection has allowed me to much better defend the claim that moral discourse is intimately connected with the will by means of its committing properties, and, unlike the other sources of support for this claim, including Joyce's work, to illustrate the serious potential for pathology that comes with such a 'technology'. The kinds of costs I've maintained that moral discourse has are just the sorts of *costs* one would expect if the discourse has the anti-akratic *benefits* Joyce thinks it does, to the extent one finds

would avoid serious versions of the downsides I've mentioned. But such restrictions would cause other practical problems, especially when we try to explain to others why we only consider such a restricted range of actions immoral.

[323] Not to mention the problems of groups and psychological feasibility and self-deception in the context of immersion, as well as posing a false choice between fictionalism and instrumental reasoning.

something like Ainslie's view compelling.

Joyce says that fictionalism becomes attractive if and only if it promises to recoup some of the benefits of the moral discourse that it finds systematically flawed (2007, p. 18). But the entailment is *only one-way*; it's just *only* if, not if. It does not become attractive simply by virtue of recouping some benefits anymore than the prospect of crawling everywhere becomes attractive by virtue of sparing us shoe leather. To be genuinely attractive it has to promise *net* benefit, not just benefit. When we compare it to some fairly readily available alternative modes of conducting ourselves, it is not at all clear that it can make any promises that can remotely make up for what a huge headache it otherwise seems to be. We really should try something else. Before talking more about what I do think would be a good idea, I have to discuss what might seem an obviously superior alternative to fictionalism, namely antirationalism about morality. If I am not committed to the claim that morality essentially presupposes categorical rational authority, but I am committed to the claim that belief in categorical rational authority is a central problem with moral discourse, then what is wrong with simply being an antirationalist aboutmorality? The answer to that question is the subject of chapter 5.

Chapter 5: Deflecting Attention From Moral Error Theory

5.0 Introduction

In this chapter I have the twin goals of arguing that antirationalism is not an attractive option and that arguments between error theorists and antirationalists that focus on the question whether moral concepts conceptually presuppose rationalism (or practical clout) are potentially interesting, but of quite secondary importance to the practical issues that come with a better understanding of the nature of our reasons and motivations. An antirationalist response to the error theorist that is focused narrowly on semantics and does not concern itself with the *threat* that the lack of a sense of practical clout represents does not address what is most significant about an error theory predicated on undermining the basis of our feeling that moral considerations have some special authority.

First, I will discuss antirationalism generally. I'll present a simple argument from David Brink that conceptual antirationalism (CAR) seems not only clearly possible but actual, at least for many people, since many people have conceptions of practical rationality and morality such that the demands of each can come apart. This simple argument is rejected by Joyce and others on the grounds that understanding morality in this way changes the subject. Then I'll examine Finlay's (2008) attack on Joyce's error theory and a defense of the latter from Olson (2010). I will disagree with Finlay's claim that moral discourse is not even *characterized by* practical clout (what Finlay calls 'absolutism'). However, Finlay claims that even if absolutism is pervasive in moral discourse, that discourse is conceptually end-relational; the absolutism does not infect the semantics, but only the pragmatics of the discourse. I do not dispute whether the discourse is conceptually end-relational or not. Instead, I argue that all these authors' overconcern with semantics obscures the more pressing problem of whether moral discourse can avoid the threat posed to it by the rejection of absolutism, which all three authors--and all antirationalists—do reject. I argue that a self-conscious antirationalism is poorly positioned to deal with this problem, and that the attempt to retain moral discourse in antirationalist fashion is, like fictionalism, the result of a failure to recognize, much less take advantage of, the opportunity in the crisis.

5.1 The Possibility of Antirationalism

I think David Brink (1997) has shown that it is clearly possible to conceive of morality in an antirationalist fashion. He argues that not only are there possible conceptions of morality and rationality on which moral and rational requirements could come apart, but that these conceptions are quite common. Here is his simple argument for this conclusion:

> 1. Moral requirements include impartial other-regarding obligations that do not apply to agents in virtue of their aims or interests.
> 2. Rational action is action that achieves the agent's aims or promotes her interests.
> 3. There are circumstances in which fulfilling other-regarding obligations would not advance the agent's aims or interests.
> 4. Hence, there can be (other-regarding) moral requirements such that failure to act on them is not irrational. (19)

As I said, the conceptions of morality and rationality presented above are hardly invented for the purpose of defending the mere possibility of antirationalist conceptions of morality. These are entirely familiar understandings of both morality and rationality. It is of course not necessary for either of these conceptions to be correct in order for CAR to be true. The point is that very many people do apparently conceive of morality and rationality in these or similar manners such that moral and rational requirements at least can possibly come apart. The extent to which one thinks they *will* come apart depends on many further questions about what sorts of things are really in people's interests and/or what sorts of aims they have and the specific nature and extent of morality's impartial, other-regarding requirements.

Surely this is too easy a victory for CAR. Many respected philosophers remain convinced of CMR despite this and other arguments denying morality's desire-transcendent rational authority.[324] What explains this? I think it is clear that the problem is that such a view threatens the authority of morality. As Brink goes on to explain, the argument above does threaten the authority of morality, on the assumption that morality and rationality are not independent perspectives from which one might judge what to do. If, as Brink, Joyce, I and many others believe, practical rationality is the ultimate currency in which to conduct practical deliberation, then where its demands and those of morality come apart, the latter should take a back seat.

I said that the perceived threat to the authority of morality is the primary explanation for resistance to CAR. To cite this as an explanation is not prejudicial to

[324] For some of these antirationalist arguments, see also Philippa Foot, "Morality as a System of Hypothetical Imperatives," *Philosophical Review,* vol. 81 (1972), reprinted, with postscript, in her *Virtues and Vices* (Berkeley: University of California Press, 1978); and David Brink, *Moral Realism and the Foundations* of *Ethics* (New York: Cambridge University Press, 1989), chap. 3.

either side. Not only does the (conceptual and substantive) antirationalist Peter Railton think this is the primary explanation (1986, p. 203), but it is at the center of Joyce's arguments in favor of conceptual moral rationalism (CMR). Joyce thinks that any attempted account of morality that leaves out practical clout isn't an account of morality at all, but has changed the subject (2001, p. 168). It can be agreed on both sides that the (perceived) authority of moral judgments is in all likelihood what explains why antirationalism seems implausible to many. The question then is what to make of this explanation.

For Joyce, the explanation is also a justification. The reason that people resist (both conceptual and substantive) antirationalism is that it fails to capture the sense of authority that moral judgments possess, and so cannot be a correct account of our moral concepts. Joyce argues that even antirationalists acknowledge that their accounts become more plausible the more they can account for the authority of morality.[325] Still, antirationalists can and do argue that so long as they can show that morality has rational authority for most of us most of the time, that is all the authority that is needed. Indeed, there are even advantages to divorcing them conceptually, for if we do not, then moral judgments are 'hostage' to judgments of practical rationality. If we have grounds for thinking that some action is (not) rationally required, then ipso facto we have grounds for thining that that action is (not) morally required. On the other hand, '[a]s long as we have not tied the content of morality to its rationality, we can reproach the immoralist

[325] Recall that Brink defends a substantive conception of practical reason that allows for the necessary rational authority of morality. Such arguments, if successful, could help to show that morality *in fact* has rational authority over all rational agents. But it would not show that that authority was a conceptual truth. Brink has also pursued metaphysical egoism to ground the prudential rational authority of other-regarding concern. Any successes in these endeavors would make conceptual antirationalism easier for many to accept.

with immorality. What is lost if we cannot also always reproach him with irrationality?' (Brink, 1997, p. 32).

Notice that it is important that the antirationalist can accurately charge someone with immorality no matter what their interests or desire set. That is to say that moral reasons can *apply* to people independently of their desires or interests, though the moral reasons might not have *authority* over them in all cases. In this sense, Foot (1972) argued that moral imperatives are categorical imperatives, i.e., they are reasons that apply to people independently of their interests or desires. As we saw in Chapter 2, rules of etiquette and institutional rules have this feature. The rule against talking with your mouth full *applies* to you whether or not you have any concern for etiquette or anybody who cares about etiquette. Likewise, if you are a member of a dues-paying organization, the rule that says you have to pay your dues applies to you whether or not you care about the organization or want to pay your dues or anything else you might want or not want.

To say that these rules apply to you is not to say that you should follow them. Not only might you have countervailing reasons not to follow them, you might have no reason at all. On an end-relational theory of practical reason, you will only have a reason to follow any rules (or do anything) if you have some end related to doing so. If you have no such reason, then the (moral or other) rules have no *authority* over you. Joyce argues against Foot's claim that only Kant and some Western intellectuals really think that morality has this 'magic force'. In chapter 2 I motivated but did not provide a sustained defense of the claim that moral judgments do purport to be both inescapable and authoritative. There I was careful to say that I thought moral judgments were characterized by this practical clout without claiming whether it was presupposed in the very concept of a moral judgment.

I have already said why I do not think that it is an important question whether the practical clout of moral judgments is to be understood as a conceptual commitment. The rest of this chapter will elaborate and defend that claim. For now what I want the reader to notice is the nature of the standoff. On the one hand, there is to all appearances the manifest *possibility* of an antirationalist conception of morality. Following Brink, I will call a person with positive conceptions of both rationality and morality such that the former has deliberative priority over the latter, and that the two perspectives can and sometimes do deliver different practical verdicts, a 'principled amoralist'.[326] I agree with Brink that such a person is possible, and it seems quite likely that many such people are also actual. The standard response to this claimed possibility is that people who make moral judgments with neither corresponding motivation nor the belief that they have any reason to act accordingly are only making 'inverted comma' moral judgments. This is understood as essentially claiming that any such putative moral judgments are best understood not as the claim, 'X is morally wrong' but as '(some group of) people think X is morally wrong.'

This presents us with a stalemate. The CMR who makes this move essentially agrees with Joyce that any person making a putative moral judgment without thinking they have any reason to act accordingly are not talking about morality in the same way

[326] I do not mean for this name to imply that this person never acts morally. It is meant to imply that when they do act contrary to the requirements of morality, it can be due to their (stable) perception of having reason on the side of amorality. It is contrasted to an 'unprincipled amoralist' who acts immorally due to weakness, depression, self-deception and the like. I include the 'stable' above so as to rule out a person whose judgment about what reason requires shifts in a motivated way that we would call akratic and self-deceptive.

that others are. Joyce finds such a conception of morality analogous to someone insisting that God exists because God is love and love exists; such a person has simply changed the subject (2001, p. 168). So what Brink and I think are clearly possible and actual conceptions of morality others rule out based on a conviction that such conceptions depart sufficiently from what people generally mean by morality to count as having changed the subject.

This can certainly seem question-begging. However, it raises the possibility that different people have different concepts of morality. This should not be surprising. Perhaps Joyce could admit the possibility that some people are CAR, but that the vast majority are CMR. He could argue that the principled CAR is the only kind. But arguably very few people have positive, general conceptions of morality and rationality at all. Joyce's primary evidence for CMR is based on how people feel about moral judgments, specifically that they do not feel that their authority evaporates on coming to find that someone doesn't have moral goals. But if someone were to internalize the principled amoralist's conceptions of morality and rationality, then that person might well feel that the authority did 'disappear' in the absence of relevant goals. It's just that such people are very rare.

Notice that on this view, while Joyce could still maintain that the (vast) majority of people are committing errors when they make their moral judgments, and in that sense an error theory was true of them, the error would not be essential to moral discourse as such. That is, if it is granted that CAR is possible, then the error most people are making is not essential to something's being a moral judgment. More to the point of this essay, the *practical* solution would certainly not be fictionalism in such a case. Rather, the solution would be to become antirationalists. That way we get rid of the problematic ontological commitment and can continue with the discourse.

But would this be a solution? Recall that Joyce advocated fictionalism on the grounds that moral discourse served an important committing function. That committing function worked in part by providing a 'calculation-silencing, desire-transcendent 'practical oomph that makes it often less vulnerable to succumbing to temptation than clear-headed … deliberation' (2006, p. 208). If Joyce and I are right that peculiarly moral discourse does serve a committing function that works in large part by deflecting attention from our motivations and silencing deliberation, then an antirationalist conception of morality can only take the place of a rationalist one if we can keep ourselves unaware of that very conception, and our relevant motivations, when it counts. In other words, the antirationalist solution on offer will look an awful lot like the fictionalist solution in the last chapter.

Fortunately, I have no horse in the race of whether moral concepts are (essentially) rationalist. Even better, the horse I am betting on says that this dispute is not important, and that the attention paid to this question by the authors discussed in this chapter suggests badly ranked priorities. As I said, Joyce attempts to answer the conceptual question by proxy. If we can discover the practical point or function of the discourse, then maybe we can judge whether the feature of clout is essential.

In the following sections I will address the current debate between Finlay and Joyce concerning Joyce's error theory. I start with Finlay's (2008) attack on Joyce's error theory. He argues that our moral concepts are end-relational and so many moral claims are true. It is useful to evaluate this debate between Joyce and Finlay since Joyce and I share a belief in the committing function of moral discourse, and Finlay and I share an end-relational conception of reasons.[327] While the central issue between Finlay and Joyce is whether the Joycean/Mackiean error theory is true, I will not attempt to adjudicate that issue. Rather, I will argue—contra Finlay—that peculiarly moral discourse is characterized by practical clout (or ‘absolutism’), and that Finlay’s antirationalism is only plausible on the assumption that people are unaware of the nature of their own concepts and, more importantly, their relevant motivations. Further, to remove this lack of awareness is to remove whatever (dis)value there is in peculiarly moral discourse. Therefore, self-conscious antirationalism and fictionalism have essentially the same strengths and weaknesses.[328] This result, in combination with arguments I have made in previous chapters, suggests that whatever (important) error there is in moral discourse is to be understood at the practical level, and mostly involves questions surrounding the practical benefits of being aware of the motivations that it is the peculiar business of moral discourse to obscure.

5.2 The (Important) Error in the "Error in the Error Theory"

Finlay (2008)[329] rejects the claim that morality presupposes what he calls ‘absolutism’. He thinks the practical reasons that people actually employ are meant end-relationally, that their given reasons are always relative to some end(s) or standard(s),

[327] Though mine is a substantive view of reasons and his is conceptual.

[328] Though antirationalism has less need for sophisticated philosophical apparatus; for example, the defense of a distinction between belief and acceptance.

[329] Finlay, Stephen (2008). The Error in the Error Theory, *Australasian Journal of Philosophy* 86/3: 347-369. All page-number references are to this work unless otherwise noted.

though they might not, and often do not, realize the meanings of their own reasons claims. Recall that for Finlay, for a fact, F, to be a reason to ϕ, relative to an end, E, is for F to explain why ϕ-ing would be conducive to E (2006, p. 8). These ends or standards are very often only contextually implicit. Since Finlay's is a semantic analysis of reason-claims (as opposed to a substantive account of reasons that could in principle diverge (wildly) from the actual semantics of reason-claims), it follows from his view that people do not make the kind of categorical reason-claims that Joyce's error theoretic argument relies upon. Even if people *think* they are making reason-claims of this sort, according to Finlay they are not.

Before we continue, it is important to make sure that Finlay and Joyce are not talking past each other, given that Finlay is denying that morality presupposes absolutism and Joyce asserts that it presupposes 'practical clout' or 'objectivity'. Joyce (2011a) delete regards Finlay's usage of 'absolutism' as orthogonal to 'objectivity', which is his and Mackie's concern. Joyce (2001, 2011a) approvingly interprets Mackie as saying that while there are categorical reasons, they are not 'objectively valid'. As we saw, on this view there can be reasons that apply to persons and are not relative to ends, but they are all 'institutional'. If one opts out of the 'institution', then though they might still *apply*, they do not have rational authority.

I think Joyce makes far too much of whatever difference there is in their usages. Here is Finlay (2008, p. 349) quoting Mackie:

> The ordinary user of moral language means to say something about whatever it is that he characterizes morally . . . [which] involves a call for action . . . that is absolute, not contingent upon any desire or preference or policy or choice [1977, p. 33].

I take what comes after the term ‘absolute’ to be clarifying what Mackie means by it, and also to be at least roughly how Finlay is using the term. Though Mackie allowed that categorical reasons exist, he claimed that they were not independent of ‘policy or choice’, in the sense that whatever ‘policy’ (or ‘institution’ as he says elsewhere) might entail categorical reasons, those reasons do not have rational authority independent of one’s willingness to participate in the institution and its policies. I think Finlay uses the term ‘absolutism’ in a sense that essentially denies that they do not have such authority. In other words, *pace* Joyce, I think ‘absolutism’ can be substituted for ‘practical clout’ without important loss, and I do so in this essay.[330]

5.2.1 Is moral discourse characterized by absolutism?

Finlay’s strategy is to 1) deny that the assumption of absolutism is ubiquitous in moral discourse and 2) argue that even if it is, it only results in the systematic falsity of moral discourse if it infects the semantics, which it does not. For Finlay, whatever abolutism there is in moral discourse, even if it is pervasive, is a pragmatic rather than semantic component of the discourse. I think the question whether absolutism is strongly associated with moral discourse is important and fairly clearly is to be answered in the affirmative. I think the question whether it ‘infects the semantics’ is not very important, provided that the answer to the first question is yes, and that the absolutism has roughly

[330] Update: Finlay (2011) has also rebutted Joyce’s claim that Finlay’s ‘absolutism’ and Mackie’s and Joyce’s ‘objectivity’ (or ‘practical clout’) are orthogonal. In fact, as we saw, Mackie also used ‘absolute’ in just this way. They are two different words doing the same work.

the function Joyce and I have argued that it has—and Finlay seems happy to concede that it has.

Appraising Appraisal Evidence

Finlay separates seven kinds of evidence in favor of the error theorist's claim that moral discourse is widely characterized by an assumption of absolutism. He argues that none of the seven push us toward the conclusion that moral talk assumes absolutism. I will discuss the two sources of evidence that I take to be the most important, both from my perspective and that of the error theorist as such.

The first is 'appraisal evidence'. When we appraise the (im)morality of others' actions, we do not look to discover their personal ends or standards as prerequisites of those judgments, nor do we withdraw our moral judgments even if we accept that, for example, Hitler had no ends or standards inconsistent with his actions. Generally speaking, not only do we not withdraw them, we find this either quite irrelevant or even perhaps an important component of their wrongness and/or blameworthiness, *viz.*, that they had no moral ends or standards.[331] Joyce argues that this shows that moral judgments are meant not hypothetically but categorically.

Finlay agrees that they are 'categorical' in the sense that we do not make moral appraisals with respect to the ends of those *judged*. 'But it is perfectly compatible with those appraisals being intended as relative to the ends, desires, or standards of the persons *judging*' (354). He notes that 'morally committed persons' care deeply about the ends

[331] From now on, unless otherwise noted, 'ends' will be short for 'ends or standards'.

embodied in their conceptions of morality. ‘Their interest in condemning immoral behaviour is not essentially connected with any concern that the perpetrators of such acts may be compromising their own ends or standards’ (354). He rightly notes that the difference between moral judgment and practical advice is just this difference; the latter but not the former is predicated on the addressee having ends to which the advice is conducive.[332]

Finlay explains the difference between the kind of ‘weak’ categorical reasons of etiquette, which Joyce accepts, and the ‘stronger’ ones of morality by appeal to the difference in the extent to which ordinary people care about one vs. the other. We don’t insist that people follow the rules of etiquette even at the clear expense of their own interests ‘because we typically are disposed to care more about the latter than the former. But we care much more about (e.g.) the welfare of children than we do about the happiness of those who may be abusing them …’ (355). Finlay concludes that appraisal evidence does not comport better with the ‘absolutist’ interpretation than the relational one.

Finlay doesn’t mention this, but in all three of the places he cites Joyce as presenting evidence that our moral appraisals are not sensitive to the addressee’s desires or interests (Joyce 2001, p. 42; 2006, pp. 60, 192), Joyce seems to confuse applicability with authority. An antirationalist has no need to deny that people do not withdraw or alter their judgments of immorality depending on anyone’s desires. Those judgments can apply without their having rational authority over the people in question. Nevertheless,

[332] Of course one might give prudential advice even in the absence of a belief that the addressee cares about prudence or her own self-interest, or that the advice presupposes such a concern. But this quibble is not important to the main point.

in other places Joyce is careful to distinguish between authority and applicability. For example, in attempting to show that morality is not meant to be an ‘institution’ whose rational authority is desire-dependent, Joyce imagines a ‘silly cult’ that requires that everyone dye their hair purple. They don’t mean it as advice, but rather as a rule that they conceive of as applying to everyone. Joyce is keen to make the following points. The first is that nobody who doesn’t care about this cult’s rules has any reason to dye their hair purple. The second is that that is not how we think about the authority of morality: ‘No human culture allows the authority of its moral rules to be so easily shrugged off’ (63).

Now what comes next in Joyce is fascinating. He recognizes that the fact that we don’t allow our moral rules to be so easily shrugged off could just be illustrative of the fact that ‘we’re all just too deeply embedded within the “morality cult” to recognize’ that there is no ‘real’ authority to our rules beyond our concern for them, including our willingness and ability to enforce them. He then says that ‘the price of accepting this is to acknowledge that the authority of morality is an illusion, that people who genuinely don’t care about it are as a matter of fact as legitimately free to ignore it as we all are free to ignore cult members telling us to dye our hair purple’ (63).

But of course we are not, as a matter of fact, free to ignore it. A moral community makes it its business to see that you are not free to ignore it. Joyce of course means not practically free, but rationally free. And though it is also true that a moral community makes it its business to see to it that they give you some genuine practical reasons not to ignore their moral rules, there is an important truth in this talk of being ‘really’ and ‘legitimately’ free to ignore moral rules. That is just to say that if you really don't care and you can really reliably escape punishment or costs of any kind as a result of ignoring them, then there is nothing necessarily irrational in your ignoring them. And there is also truth to the claim that accepting this view is accepting that the felt sense of the categorical authority of moral values or rules is based on 'an illusion'.

Return to Finlay's claim that moral judgments are 'intended as relative to the ends' of the judgers. I take it as obvious that the judgers in general would not accept that this is all there is to it. Not only do ordinary judgers not accept it, but it is quite a minority view among philosophers. When Finlay denies that moral reasons are intended to be authoritative for those addressed or judged, he is denying what the (vast?) majority of people who have familiarity with the question would say. So here, as elsewhere in his attack on the error theory, Finlay attributes pervasive errors of other sorts to ordinary speakers, as well as moral philosophers, in regard to the meanings and/or intentions associated with their moral judgments. I want to be clear. This is not meant as a critique of Finlay's argument, but as an illustration of the fact that even on Finlay's view, most people are mistaken about the nature of their moral claims. It is a very important part of my view that these mistakes are (highly) motivated.

We might ask, if Finlay's analysis is right, what explains why it is so rarely held? Why is it relatively easy to see that reasons of etiquette don't have authority over someone if they have no concern for etiquette or any of the consequences of not abiding by its strictures, but so many people, even after faced with persistent arguments to the contrary, continue to firmly believe that moral reasons *do* have this kind of authority? I take it that the answer is rooted in the difference Finlay acknowledges between the two kinds of judgments. We care a lot more about the actions associated with morality than with etiquette. But that *alone* does *nothing* to explain the difference. For if it were a simple matter of caring more, there shouldn't be any obstacle to acknowledging that we care more about morality than etiquette *and that's why* we're willing to insist on the one and not the other. But we don't do (only) this. 'We' think there are reasons to behave morally that do or should have authority over people no matter their ends. It is a challenge for any antirationalist to explain why it is that we don't engage in 'straight-talk' about what we want and what we're willing to tolerate and (eager to) punish. That is, an antirationalist account of morality needs to explain why we have any need for peculiarly moral discourse at all.[333]

If the gist of what I have argued in previous chapters is right, the explanation is at least in significant part because of the way our motivational systems work. Plausibly, the specific issue of punishment is especially difficult to engage with straight-talk due to the (moralized) notion that one may not force others into doing something simply because it is important to the one(s) doing the forcing.[334] For the moment, it suffices to note that whatever the explanation, we will, at least in the short-medium term, have a much harder time getting stably motivated to insist (where 'insisting' requires teeth) on others behaving in certain ways if we think of this insistence as based 'merely' on the ends that we care about.[335] Finlay acknowledges that we don't retract our judgments about the Nazis even if we suppose they don't have ends inconsistent with their actions. But we most certainly do not describe what was wrong with what they did, or why it was worth

[333] Joyce (2006, p. 208) makes essentially the same point.

[334] This is an example of a moral principle acting as a commitment to secure a general sort of behavior—noncoercion--but that then confuses us due to its overgeneralized and absolutist character.

[335] We will have made the most important step when we no longer feel moved to preface our (deepest) concerns with words like 'merely'.

the incalculable price to stop them, solely in terms of ends that were important to us, but didn't necessarily have any authority over the Nazis. We (at least very many of us) would say that what they did was wrong regardless of whether they *or we* cared about it or not, and that we care about it largely if not entirely *because* it was so hideously morally wrong.[336]

If something like what I've been arguing is right, the peculiarly moral character of our condemnation of the Nazis requires, if not the acceptance of something like practical clout, at least the lack of acceptance that the idea of practical clout has been discredited. Therefore up to this point, Finlay's arguments could be entirely correct and the reasons I've given for abandoning moral discourse would not have been much or at all affected. For what I have argued is important is that moral discourse is threatened by an awareness of the end-relational nature of our (moral) reasons and Finlay has not touched this view.

Does Moral Disputation Provide Evidence for Absolutism?

After dismissing appraisal evidence, Finlay takes up three more kinds of evidence that have been used to defend the claim that absolutism infects our moral discourse. I will focus on the last of these three, 'disputation evidence'. This evidence consists in the existence of apparent fundamental moral disagreement, where the disputants nevertheless take themselves to have a common subject matter and univocity of predicates. If people

[336] It is possible, and even plausible, to argue that in saying that we would be wrong to not have such ends, that we are judging some (lack of) ends we might possibly have relative to the ends we have now. I think it is plausible that we 'rigidify' our conception of wrongness in this way (cf. Prinz, 2007), but I also think that acknowledgement of this fact would have the same kind of *potential* undermining effects for the same kinds of reasons I have been and will continue to point to, and that these felt dangers are also part of the explanation of why this view is not more widespread.

with fundamentally different ends nevertheless argue with each other with the hope of convincing one another, as seems to (sometimes or often) happen, that appears to be evidence that they understand their moral claims not as relational but absolute. If they were to see their claims as being relativized to their respective ends, they would realize there was no point in their argument, since the claims of each could perfectly well be true, relativized to their respective ends. Under these circumstances, if reason-claims are meant relationally, we should expect these arguments to stop altogether once the disputants realized their different ends, or a least to take a form that does not seem to presuppose univocity.

Finlay first wants to insist that such disagreements are actually much less common than often supposed. ‘They all involve disourse between people who are morally alien to each other’ (356). He invites us to think of the last time we ‘engaged in moral discourse with someone like Charles Manson or a neo-Nazi’ (356). Second, he points out that in order for such disagreements to be evidence for absolutism, these disputants would have to *accept* that their interlocutors are morally alien. But since this is so uncommon, per Finlay, disputants plausibly assume some common ground between themselves until such an assumption is no longer tenable. And in the rare cases where diagreement continues in such a way that suggests absolutism, Finlay argues that it is ‘only’ the pragmatic, not semantic point of the discourse.

In responding to these arguments I’ll incorporate some material from Olson (2010). I think that the arguments in favor of moral discourse being characterized by absolutism are important and sound. If moral discourse were not even characterized by absolutism, my entire project would be undermined, and that would be wonderful since the project just is to rid us of absolutism. But it most certainly is characterized by it, at least peculiarly moral discourse. Olson is defending the conceptual claim that moral discourse essentially presupposes absolutism, i.e., that it does infect the semantics of the dscourse. I am much less confident of the soundness of these arguments and I do not concern myself with them in any detail. I am quite confident that they are not nearly as important as the arguments for the existence of the absolutism. Whether in the semantics or pragmatics (assuming there is an important distinction), absolutism is there and it is of practical importance.

5.2.2 Absolutely

First, Olson claims that Finlay dramatically underestimates the prevalence of fundamental moral disagreement (FMD). Olson cites the disagreements between ‘conservatives and feminists; socialists and neo-liberals; cosmopolitans and nationalists’; between those who think that human and animal suffering are morally on par and those who think human life is morally special; and between 'pro-choice' and 'pro-life' activists. Among philosophers, Olson sees FMD between ‘utilitarians and deontologists; Rawlsians and Nozickians; anarchists and communitarians, etc. .’ Olson rightly wonders why we should ‘assume that the person with whom you have a fundamental moral disagreement is such a depraved character’ as Charles Manson or a neo-Nazi. ‘She might rather be a utilitarian, a Nozickian, a liberal, a conservative, a socialist, a nationalist, an ethical vegetarian, or a 'pro-life' activist'.[337]

Olson also responds to Finlay's claims that 1) apparent FMD often arises from non-moral disagreement, such as empirical or theological disagreement (Finlay 2008, pp. 356-8), and 2) speakers have to accept that they are in FMD for this to count as evidence for absolutism. Olson says that it is both implausible and uncharitable to think that all or even most cases of apparent FMD are based on non-moral disagreement. Unfortunately, he offers no argument as to why it is implausible or uncharitable. I suppose the idea behind it being uncharitable is that people in these cases don't realize that their disagreement has a nonmoral basis and so continue to argue in moral terms. That is only as uncharitable as you think it is easy to recognize that their disagreements have non-moral bases. The vast majority of people have little to no training or practice in discerning whether the basis of their disagreements are moral or nonmoral and so plausibly should not be expected to recognize it.[338]

Philosophically trained people do have such training however, and often do recognize it. However, at least some of the academic philosophical debates that Olson cites seem to be among the best examples we have of cases where the apparent FMD is real and recognized and therefore is evidence for absolutism. They seem to be real cases at least insofar as they do not seem to be based on non-moral disagreement. The best evidence of this is that the professionals engaging in at least some of these disputes *take*

[337] All quotes from Olson here and below from pp. 19-21 of the prepublication draft of his (2010) article.

[338] Personally, I find it plausible that very many (at least apparent) FMD have non-moral bases, but certainly not all of them. So long as there are some in which the disputants take themselves to be in FMD (whether or not they really are), they will either tell against Finlay's analysis or require a separate, 'disunifying' analysis.

their disagreements to be fundamentally moral. These would seem to be the paradigmatic cases of FMD that Finlay's analysis cannot handle.

However the disputation evidence for absolutism is much wider than these cases, because the evidential requirement for *actual* FMD is too strong. It is only necessary that people *accept* that they are in FMD and continue to argue with one another in the relevant ways for their behavior to count in favor of an absolutist element in the discourse. Therefore it is not clear that it would help Finlay's cause even if it were to turn out that *all* ethical arguments *actually* turn on factual premises at some level. So long as people *take themselves* to be operating on fundamentally different motivations or standards, yet continue to dispute with one another with apparent univocity of predicates, they seem to be presupposing absolutism.[339]

I conclude that Finlay's claim that moral discourse is not even (widely) *characterized* by absolutism is untenable. First, Finlay's arguments against Joyce's appraisal evidence did effectively show that there is no need to suppose that what people are *actually* concerned about when making moral judgments is the addressee's reasons for action. But in showing that, it did nothing to undermine the evident fact that when people make peculiarly moral judgments, they do not *take themselves* to be 'only' concerned with (enforcing) their own (moral) rules or goals. Second, given that all that is required for evidence of absolutism in FMD is that people take themselves to be in FMD, I submit that Finlay grossly underestimates the amount of evidence there is for absolutism as a general characteristic of peculiarly moral discourse.

[339] Again, I'm not saying that they seem to presuppose it as a conceptual, much less essential conceptual element of the discourse, only that they seem to presuppose that there are the sorts of reasons that end-relationalism rules out.

Finlay's next move then is to argue that wherever and to whatever extent that the discourse is characterized by abolutism, the absolutism does not infect the truth-conditions of moral judgments. I am only going to address the dispute about whether the absolutism is a conceptual feature of moral discourse in order to illustrate both how difficult it is to even see how to adjudicate this question and that fortunately, the answer is of academic interest at best.[340] To the extent that the value of moral discourse is essentially practical, locating a problematic element of that discourse in its pragmatics does little or nothing to vindicate the value of that discourse.

5.2.3 You Say Semantic, I Say Pragmatic

Though he claims that it is not so widespread as Joyce, Olson and I think it is, Finlay grants to Joyce that "sometimes the point for which we engage in moral discourse is 'a desire to say something more – to imbue the moral imperative with a greater authoritative force' [Joyce, 2001, p. 41]" (357). But he challenges Joyce's claim that this greater authoritative force is part of the semantic point of the discourse. 'It may instead simply be its *pragmatic* point, or what we aim to accomplish by our speech acts' (357). It is a 'feature of the rhetorical use of relational moral language' (357).

Olson responds by claiming that while both the absolutist and relational interpretations capture the data, the former does so better. First, the absolutist interpretation fits better with the claim that moral language evolved in the first place at

[340] I see it as evidence that there is not a deep distinction between semantics and pragmatics, but I have no commitment to this general claim.

least in part to fulfill the ‘coordinating and regulative functions’ as has been argued for by Joyce (and myself). Olson submits that it is more plausible that they fulfill this function by entailing categorical reasons than if they are conceptually relativized to some particular end.

It is not clear to me how to evaluate a claim like this. If Olson means to say that moral discourse evolved biologically, I find it highly doubtful that biological evolution can have much to do with ‘entailing categorical reasons’. As in the case of Joyce’s biological evolutionary arguments, I just see nothing in the way of evidence that biological evolution selects for conceptual entailments or lack thereof, and good reasons to think that it does not. It is far less implausible in my view to argue that cultural evolution is able to establish something like conceptual entailments, but I wouldn’t know how to evaluate this claim either and Olson gives no help here either.

Somewhat helpfully, Olson tries to show that Finlay’s attempt to analyze the rhetorical force of moral claims as ‘simply’ pragmatic backfires, since this extra force is most straightforwardly explained if we assume the truth of the conceptual claim. He argues that the difference in ‘seriousness and intransigence’ between judgments of etiquette and morality are likely to be reflected in the different concepts we use to make them. Less plausibly, he argues that it is hard to see how moral claims could retain their greater rhetorical force if they were conceptualized as having no more intrinsic authority than those of etiquette. If neither entailed categorical reasons, Olson says, then we should be able to ‘waive’ moral demands as well as those of etiquette.

Olson’s latter claim is fairly easily rebutted as it stands, since by hypothesis moral claims are much more important to the moral community (which includes oneself) than those of etiquette. Therefore they will not *let you* waive them in the way they might *let you* waive those of etiquette. As Finlay says, 'it is moral criticism, not the authority of moral reasons, that is inescapable' (2007a, 142). However, this response is also inadequate as it stands, because it is just false that the moral community (again, including oneself) *presents or conceives of* its demands as being 'simply' *its* demands. Notice that this does not beg the question against Finlay, since he does, and must, acknowledge that people can conceive of the meanings of their own moral claims incorrectly, thinking that they are absolutist though they are not.

This lands Finlay back in the teeth of my earlier charge that an antirationalist has the burden of explaining why moral discourse is around in the first place. If our moral concepts are end-relational, why does it seem to so many people like they are not? Finlay interprets what we are doing when we use this rhetorical feature of relational language as 'demanding' that our interlocutor adopt the relevant ends. But what then is the explanation for why we do not do this more straightforwardly? Why do we not demand it outright?

Compare Joyce's (2001) response to Stevenson's (1937) argument that moral judgments are not assertions but expressions of emotions intended to influence people to approve or disapprove of actions. Stevenson thought that morality was 'largely a manipulative device' (Joyce's phrase, 14). But, crucially, the manipulative device *looks just like assertions look*. Instead of either attempting to persuade someone by reference to ends they are understood to share, or by overtly demanding that they do something, we, according to Stevenson, make what look like descriptive claims about what is simply right or wrong in order to achieve 'a quasi-imperative force which, operating through suggestion, and intensified by your tone of voice, readily permits you to begin to influence, to modify, his interests' (Stevenson, 1937, p. 22), quoted in Joyce (2001, 14)).

We can agree with Stevenson and Finlay that moral judgments have the function that they have, this greater rhetorical force, but when we ask *how it achieves* this extra force, the best explanation might be that it does so by entailing that certain actions are absolutely forbidden or required. The fact that we want to influence someone with our utterance does not make the utterance a command any more than claiming that Paris is in France is a command because I am saying it in an effort to get someone to believe it (Joyce 2001, p. 15).

I find this a nearly decisive objection to Stevenson's argument, and a somewhat powerful objection to Finlay. But more important, I don't think it matters much. If Stevenson were right, which I take it he is not, it would still be a major *practical* problem just insofar as the fact that moral language is primarily a manipulative device becomes commonly believed. As Joyce says, 'few of us are so scheming in our expressions of will to power' (2001, p. 15). That is, when we manipulate people, including ourselves, we are not aware that this is what we're doing (at least at the time). There are very powerful reasons to suppose that if we did know this, doing so would become much more complicated and difficult.[341]

Analogously, if Finlay is right, and the absolutist element of moral discourse is not part of its semantic but rather its pragmatic point, I don't see that it matters much, if

[341] Even if this view of morality were not *commonly* believed, it would be a problem for those who did believe it, just insofar as they are not 'so scheming in their expressions of will to power' as to consciously manipulate people in the way that they would have to if they came to accept Stevenson's analysis (and did not succeed in tucking that awareness out of sight in the spirit of Joyce's fictionalism). Also, unfortunate and frustrating as it may be, people are notoriously more difficult to manipulate if they know that is what you are trying to do.

at all. As Finlay says, the pragmatic point of a discourse is a function of ‘what we aim to accomplish by our speech acts’. What he doesn’t seem to see (or care about?) is that that very aim is threatened by a recognition of how it is achieved. It is, as far as I can tell and as far as any of Finlay’s or Joyce’s or Olson’s arguments are concerned, immaterial whether the hypothesized motivational effects or functions of absolutism work via its incorporation into the semantics of moral discourse or ‘only’ the pragmatics.

5.3 Not False, Only Important

Despite my contention that the outcome of this dispute about semantics vs. pragmatics is not highly relevant, it’s worth taking a fuller look at Finlay’s argument against the conceptual claim. My reasons for doing so are the same as they have been. What I want to continue to argue is that every attempt that Finlay makes to rescue morality from conceptual error just lands it directly into the kind of error that I have been and will continue to claim is the real essential error of moral discourse. I think it is very revealing that this continues to happen. That is, every plausible move that Finlay makes in response to Joyce’s error theory just serves as fodder for my dual claims that 1) any essential error in moral discourse is practical and 2) the heart of that error lies in the motivated mis- or lack of perception of our motives.

5.3.1 Truth Can Survive Systematic Error

To this point Finlay has been arguing against the idea that absolutism is a common characteristic of moral discourse, and that in the rare instances where it does or might occur (such as in FMD), it is a pragmatic rather than semantic component and so does not render even the claims made in FMD false. Now Finlay wants to show that this move is legitimate even if absolutism is *utterly pervasive* in moral discourse. Still, he insists, it is part of the pragmatic and not semantic, or referential, point of the discourse.

To see how it is possible that people might have universally false background beliefs associated with some discourse and yet an error theory for that discourse (apparently) not to be appropriate, we only have to consider water and motion discourse. Though for centuries water was almost universally believed to be an element rather than a compound, we do not think that everything said about water during that time was false (351). Most obviously, people could point to some water, say that it is water, and be right, since it was in fact water. Similarly, people might not realize that motion is relative to an inertial frame and still say truly that something moved. As Drier (2005) argues, and as these examples seem to show, there can be no simple assumption that even uniformly-held assumptions in some domain infect the content of our concepts about that domain (351).

Of course in other cases this is not so. As Joyce and Finlay agree, witch-discourse and phlogiston-discourse are appropriately handled with an error theory. Joyce thinks that moral discourse belongs with these and Finlay thinks it belongs with motion- and water-discourse. As mentioned, Joyce thinks that we can get a handle on which discourses should receive error-theoretic treatment by inquiring into the point of the discourse. He rightly thinks that the point of water discourse was primarily to refer to a certain kind of stuff all having some essential physical composition in common and the point of motion discourse to refer to changes in the position of objects over time. The unrealizable points of witch and phlogiston discourse, on the other hand, were to refer to women with supernatural powers and to the stuff stored in all flammable materials and released during combustion. (Finlay, 2008, p. 363; Joyce, 2001, p. 96).

Therefore Finlay regards Joyce's error theory as 'premised on the claim that the whole point of moral discourse is to refer to value with absolute authority' (364). He quotes Joyce as saying that motion discourse can easily survive our letting go of absolutism, while "'ordinary use of the concept of *moral rightness*, by contrast, is completely undermined without absolutism' [2001: 97]" (363). Now Finlay urges that this can't be right as stated, since Joyce wants to *preserve* moral discourse, precisely *because* it is useful. But then its *whole* point can't be to refer to that which does not exist or it would be pointless to try to preserve it. Finlay rightly notes that Joyce does not want to give moral discourse a new point but to preserve its current one. It therefore follows that the whole point of the discourse cannot be to refer to absolutely authoritative value (364).

What Joyce must mean, says Finlay, is that referring to absolute value is the '*referential* point' of the discourse (364). A discourse could have multiple points; for example, witch discourse could have the points of preserving villagers' health, detecting and punishing sinners, and suppressing female power and autonomy (364). Finlay claims that for cognitivists like he and Joyce, 'only intentions to refer are properly taken as determinative of content' (364), and so Joyce must be claiming—and Finlay is denying—that the referential point of moral discourse is to refer to absolute value. Once we see that a discourse can have several distinct points we can see the possibility that the absolutist feature of moral discourse can be explained in terms of some other, non-referential point of the discourse (364).

Rejecting the notions that the content of moral thought are given by either our introspective judgments on the matter or by what our intentions to communicate might be (since this does not distinguish semantic from pragmatic points), Finlay argues that what determines content are the ‘essential application conditions’, which are roughly to be understood as the ‘criteria on which a concept or term is applied’ by competent first-order users (364 - 5). And since Finlay thinks that the essential application conditions for moral wrongness, as well as for motion, are relational, then he thinks that something is judged to have moved just in case it has altered its position relative to a reference frame, and something is judged wrong just in case it violates certain standards or frustrates certain ends. And this goes even for those who explicitly defend absolutist conceptions of motion or morality (364).

After all, says Finlay, believers in relational vs. absolutist conceptions tend to (seem to) make the same first-order motion and moral judgments. But since by hypothesis there is no such thing as absolute motion or absolute value, these people’s judgments could not be responsive to these nonexistent properties. On the contrary, they are responsive to their respective relational properties, no matter what an absolutist might think they’re responsive to, which is why, per Finlay, we are justified in saying that any such absolutist ‘misunderstands his own language and thought’ (365).

5.3.2 One Objection Too Good

At this point in his essay Finlay makes an objection to his own proposal, his response to which is both confused and illuminating. He parodies his own arguments, pointing out the 'inconvenient fact' that people did manage to use the concepts of phlogiston and witch in ways that had significant agreement across competent users (366). Since the users of these terms could not be responsive to nonexistent properties any more than the users of moral discourse, they must instead be responsive to some complex disjunctive relational property, perhaps including having enemies who have suffered misfortune. It is then this property to which people are responding when they style someone a witch, and so their claims can be true after all. Similar remarks go for phlogiston (366).

Finlay responds by noting that our 'application of concepts is often responsive and sensitive to what we take to be *evidence* of—but not constitutive of—their instantiation' (366). He remarks that competent users of witch discourse would have no trouble imagining cases where there was good but misleading (lack of) evidence of someone's being (or not being) a witch. Can the moral error theory be 'rescued' in the same way? That is, can the moral error theorist claim that the makers of moral judgments, while not sensitive to (nonexistent) absolute value, are sensitive to relational properties that people take to be evidence for absolute ones? Finlay thinks not.

His first response to this move is to say that it is 'gratuitously uncharitable'. I ignore this response because I cannot see, and he does not explain, why it is more uncharitable than his claim that they 'misunderstand their own language and thought', and at any rate it trades on the same confusion I expose presently. Finlay's second objection is that it would be strange that people would take relational value[342] to be sufficient evidence for absolute value. 'Wouldn't they,' Finlay wonders, 'rather be sceptical that relational value … was any indication of the real thing?' (367). Presumably, Finlay claims, this inference would be based on an assumption that the 'moral standards or ends are the absolutely right ones to subscribe to … [but] we might wonder where this assumption comes from' (367). Further, many ethical theorists think we have direct insight into basic moral truths that neither require or even admit of evidence. Some theorists regard this as a platitude. We should therefore, according to Finlay, be skeptical about any theory that claims that even these seemingly basic truths are arrived at inferentially (367).

Finally, Finlay claims that this defense of the error theory 'undermines the very argument in its favor … the claim that ordinary judgment is absolutist by default because we are not ordinarily sensitive to *the significance of the relevant parameters*' (367, my emphasis). But, says Finlay, 'if we actually infer absolute moral value … from the evidence of the relational facts, then we are after all sensitive to *the relevant parameters* … ' (367, my emphasis).

5.3.3 The Significance of Significance

Of course the, or at least this, 'error theorist' is perfectly happy to accept that

[342] He talks about relational motion and value in this context, but I will stick with the latter for simplicity's sake.

ordinary people are sensitive to the ‘relevant parameters’, whatever those might be. But not to their significance! That is, one can perfectly well be sensitive to the fact that something frustrates one’s ends or violates one’s standards without being sensitive to the fact *that that is what one is sensitive to*! The whole idea--my whole idea at any rate--is that this is *exactly* what is going on (at least very much of the time). People are quite commonly sensitive to that which frustrates certain of their ends or standards, and respond in the ways they do at least in large part *because* their ends or standards are violated. That does *not* imply that they are sensitive to (aware of) *which ends* are the ones that are being frustrated, or to the fact that their moral judgments have anything to do with sensitivities to the satisfaction or frustration of their own ends or standards at all! The fact that one conceives of, interprets, and/or dresses up one’s ends or standards in the glad rags of moral judgments clearly does not imply that they are (if anything it suggests that they are not) aware that they are doing so.

And even if we *do* suppose that they acknowledge that their judgments are relative to their ends, it is quite common for people to think that they *are* the absolutely right ones. We might indeed wonder where such an assumption comes from, since that is a fascinating question. My wondering has led me to develop and articulate what I take to be a fairly well-supported hypothesis, drawing on empirical work in motivation, the will and social psychology, that such an assumption is widespread at least in large part due to its committing function.[343] We might also think there are important religious roots for it, which if true would only supplement, not undermine, the committing function argument.

[343] Of course this ‘assumption’ serves its function in a larger conceptual and cultural context or environment, but this does nothing to undermine the claim that it does have that function in those contexts or environments.

So now we have (once again) come to see that no matter the winner of this semanic dispute—to be clear, I am not saying that Finlay is wrong about the semantic claims—the best *explanation* for why people might 'misunderstand their own language and thought' in this domain is because the absolutist component of moral discourse serves the committing function that Joyce and I have argued that it does. It does not matter whether it is part of the semantics or 'only' the pragmatics.[344] And this is why the *spirit* of Joyce's claim remains intact, i.e., that without the absolutist component, talk of moral rightness loses its (main) point.

One potential objection from Finlay is nevertheless in order. He might claim that it is important that people in the witch case are *aware* that black hats, warts, suffering of adversaries, etc., are evidence of witchhood. But in the moral case, people are presumably not generally aware that the relational values to which they are sensitive and responsive are being treated as any such evidence. They are therefore not inferring from evidence in a way that they are aware of or that they would acknowledge, even on reflection. In fact, the way Finlay presents his witch examples suggests if not conscious inference, at least inference that would be acknowledged as such on a little reflection. But if the people will not ackowledge that any inference is going on in their perceptions of moral value, perhaps we should not count what they are doing as 'inference' at all, in which case the disanology between witch and moral discourse can be maintained.

Once again however, this is just grist for my mill. It is perhaps telling that Finlay chose to answer his own hypothetical objection, which included witch and phlogiston discourse, only with an example of how people could distinguish evidence for witchhood

[344] Or if there is any sharp distinction, which I doubt.

from what it is to be a witch, and used evidence such as black hats, warts and ill enemies as examples. It is easy to see that such evidence could be quite fallible. But suppose a typical 17th-century Salemite saw a woman apparently flying around on a broom in broad daylight, with a large black pointy hat, green hair, and shooting lightning out of her hands. Do you suppose that this fair citizen would think that they had good but fallible evidence that this woman was a witch, or would they simply perceive a witch, just as surely as phlogiston scientists perceived phlogiston escaping and scientists watching cloud chambers perceive electrons?[345]

The point here is the familiar one of the theory-ladenness of perception. What we (think we) see directly is very often conditioned by background beliefs in subtle and powerful ways of which we can be quite unaware. We can have extreme difficulty imagining not perceiving the world in some of the terms and categories to which we are accustomed, especially insofar as these categories are or seem fundamental to our identities and worldview. Millions of people claim to have seen miracles and magic of all kinds directly. A convincing performance, in the mind of someone predisposed to believe in such things, is not perceived as good but fallible evidence for a miracle or magic, but as a simple witnessing of the real thing.[346]

Finlay might want to insist that surely most of the witch-sighters, if pressed, will admit that what they really have is only very good evidence (at best), and that what they

[345] It goes without saying that on this view of perception that one can perceive something without that thing's actually being there. That is just a consequence of, or at least a natural way of thinking about the consequences of, a 'theory-laden' conception of perception. If the background theory is false, one will perceive what is not in fact there (Churchland, 1979)

[346] I once showed a friend some Derren Brown videos which I thought were very impressive; I could not begin to conceive of how he had done some of his 'mind-reading' tricks, absent the use of stooges which he swore he did not use. She was most impressed by the one trick that I could easily see how to do, a use of simple but clever misdirection. For everything else, she found it hard to see what all my fuss was about, since he was 'just reading their minds'.

have actually seen is not really constitutive of the phenomena at hand (it is possible that the witch-like display described above was an elaborate trick). Well, perhaps so, but this is *equally the case for morality!* It is a commonplace for people to think themselves directly certain of what they consider basic moral truths, like the immorality of abortion or racial miscegenation or homosexual sex or interfering with other cultures' values. Yet if pressed hard, many will acknowledge the bare possibility that they really only have good (intuitive) evidence (and if pressed harder, often become less confident and more confused and sometimes angry). I see no reason why such people would be generally different from those who perceive directly that they have seen or been on the receiving end of a miracle. If pressed, both might admit the bare possibility that they are wrong, and that really they 'only' have very good evidence, and that their background beliefs are playing a large role in their perceptions and judgments. But neither would experience themselves as making inferences from their evidence to conclusions. Rather, they would experience themselves as simply seeing the relevant truths.[347]

Now, we have been acculturated to perceive certain kinds of things as being *immoral*. We have been trained to experience certain kinds of emotional/affective responses as perceptions of *(im)morality*.[348] Others, including a large subset of 'us', have been raised just as certainly to perceive certain kinds of responses as perceptions of sin or sinfulness. For such people, watching cold-blooded murder for material gain is no more experienced as fallible evidence of sin than for others it is fallible evidence of moral

[347] Recall the brief description in chapter 2 of the Orissan child's coming to carve up his world in conceptual terms that will seem entirely natural and straightforwardly perceived at maturity.

[348] Complicating matters, the perceptions as of (im)morality also affect the responses themselves, interactively. At any rate, I don't mean to imply that we have purely noncognitive responses which are then interpreted cognitively. The perceptions are, in my view, bound up with cognitions of a sort.

wrongness.

And nevertheless, we (godless ones) might think that sin discourse is best handled by an error theory. Or, more apropriate to my interests here, we might think that sin discourse is in trouble for those who disbelieve in God. And we might think this *even if* we were convinced that the referential or semantic point of sin discourse is not, after all, to refer to acts that are 'prohibited-by-God'. We might think that whatever the result of this semantic dispute, we need(ed) some other discourse, given that we no longer believe in God, and that belief in (or acceptance of) God was doing important work in our sin discourse. And of course we *did* (further) develop our moral discourse in response to the unavailability of religious discourse to serve the functions it had been serving.[349] Now, what I'm saying is that we find ourselves in an analogous position with respect to moral discourse and its contemporary and successor discourse(s).

5.4 Antirationalism: What Is It Good For?

Now I hope it is becoming more clear why the 'error theory' as such is not very important, once we have taken on board (at least as decent approximations) the substantive claims that I (and to a large extent Joyce) have made with regard to the committing function of moral discourse and how it is accomplished. Joyce advocates fictionalism in an attempt to recoup the benefits of moral discourse without the costs of belief in it. Those benefits came from the absolutist element that Finlay says is 'only' its

[349] Even for those who believe in God, religious discourse cannot serve one of its crucial functions when discussants have deeply different conceptions of what God wants. As we'll see, this is one of the primary reasons for moving beyond moral discourse as well.

pragmatic point. And I have been arguing that it is (at best) a narrowly interesting question whether Finlay or Joyce is right on that score.

The argument of this section is that even if antirationalism is correct, it leaves us in a surprisingly similar situation to that in which we would find ourselves if Joyce's error theory were correct. Joyce had to decide what to do given the truth of his error theory. He chose fictionalism, a crucial component of which is to remain generally unaware of our underlying motivations by means of objectifying values in the moralistic style. What neither Joyce nor Finlay seem to recognize is that if antirationalism is true, we need to decide what to do about *that*. I think that the grounds for retaining or abandoning moral discourse are quite similar whether one is an error theorist or antirationalist.

Though Joyce, Finlay, and Olson are disagreeing intensely over whether moral claims are true or false based on the truth of the conceptual claim, I am unable to see how, for example, Finlay's analysis should make Joyce any less interested in fictionalism, only of a slightly different sort. Of course Joyce couches the problem with moral discourse in terms of the truth and falsity of the judgments themselves, and argues against continuing to believe them on grounds that doing so would infect our valuable truth-tracking machinery. I hope now to have shown that the true problem is not whether moral claims can be true, but whether they can do their distinctive work without the (stable) availability of the absolutist component. We saw in the last chapter that the kind of fictionalism that Joyce recommends would require a highly disciplined effort to manipulate one's thinking,[350] an effort which had at its core, or at least as a critical component, avoiding the awareness of what one was doing. But if he were to come to accept Finlay's analysis, a very similar 'remedy' would be required.

Joyce argues that it gives morality an 'extremely odd flavor' to say that one person murdering another is 'depraved and morally unacceptable and in the next breath [to assert] that he had no reason to refrain and that in fact committing the murder was what he ought, all things considered, to have done' (2006, p. 204). The point here is of course to argue that practical clout (or oomph) is an important part of moral judgments. Joyce doesn't deny that we could still have good reason to punish or despise such a person, but that next to the sincere judgment that they did what was the most rational thing for them to do, our hostile attitudes and/or judgments 'lose some of their teeth' (204).

I said above that Joyce acknowledges the possibility that the reason we do not allow moral rules to be 'shrugged off' as easily as strange cultish requirements is that we are 'too deeply embedded in the "morality cult"' to see that moral requirements have no more *intrinsic*[351] rational authority than wacky cult rules. Early in his book, he acknowledged that our motivation to resist the conclusion that someone could ever perfectly reasonably scoff 'Morality, Schmorality!' might lie in our 'being immersed in a particular normative framework' (2006, p. 63). Later, returning to the same theme, Joyce suggests that the reason we are uncomfortable granting that immorality could be practically rational might be

[350] For Joyce, the thought-control is for one's own good, but it could easily be extended to cover commitments in the service of other ends.

[351] They can have lots of derivative or relational authority, and that this is part of any plausible antirationalism, but by hypothesis this kind of authority is experienced differently.

> due simply to the fact that we are adherents of the moral institution, unwilling to step out from its linguistic rules. Yet this in itself seems troubling, for it seems to suggest a degree of self-delusion in the moral adherent's relationship to moral precepts. If it is ex hypothesi true that [a person] has no real reason to refrain from murder, but moral people won't admit this due to their adherence to the moral institution, then that institution is blinding them to something. (2006, p. 204)

In other words, Joyce thinks that being deeply immersed in the normative framework of morality can lead to self-delusion and blindness.

Joyce's solution to this problem was to find a way to be deeply immersed that avoided this self-delusion and blindness. He argued that one could be deeply immersed in the very same framework, but without technically believing any moral claims or judgments.[352] I argued that, for several kinds of reasons, this project was highly ill-advised. To the extent that the level of immersion which Joyce advocated was psychologically feasible, it looked for all practical purposes like the blindness and self-delusion that Joyce was trying to avoid. Indeed, such self-delusion and blindness are among, but do not exhaust, the entirely unacknowledged downsides of the motivational tools that Joyce was keen to preserve through his fiction.

Now it looks to me that the kinds of problems awaiting the fictionalist also await the self-conscious antirationalist. I cannot see that it matters whether the absolutism is at the semantic or 'only' pragmatic level, or in fact whether those can be sharply distinguished. What matters is the absolutist component is doing motivational work of a

[352] In the same way that, technically speaking, the pathological version of the Holmes-fictionalist believes that Holmes isn't real.

special kind such that awareness of its function tends to undermine it.[353] Now admittedly, to the extent that an antirationalist employs moral discourse in ways that do *not* employ the absolutist element, he will have no need to keep from himself the awareness that he is 'manipulating' himself or others, since he needn't be. However, to just that extent, *he will have no need of peculiarly moral discourse*. That is, in whatever contexts he does not want to make use of this rhetorical feature of moral language, it should in principle be just as profitable for him to speak in terms of his (group's) ends or standards.

This returns us to Finlay's claim that it is moral criticism, and not the authority of moral reasons, that is inescapable. Finlay said that reasons of etiquette appear to lack categorical force because we tend to care less about them than moral reasons. I think that is right. But I noted that the story must be more complicated, since we don't simply say—to others or ourselves—that we care more about them, and that we are far more prepared to discipline and punish for moral violations. On the hypothesis that it is *in fact* our own ends to which these reasons are relativized, there must be an interesting explanation as to why we don't say or think so. I think mine (and Joyce's) is, if imperfect and incomplete, on the right track. And if the arguments I made in Chapter 4 are roughly correct, regarding the difficulty of achieving these motivational benefits in the context of the knowledge of what one is doing (and especially if others do too), then either a self-conscious antirationalist faces many of the same practical difficulties as Joyce's fictionalism or it voids what is peculiarly (dis)valuable about moral discourse.

[353] See especially section 4.2.2 on why awareness of these functions is problematic.

This is why morality is different than water.[354] We don't want to say that people who thought that water was an element could not say anything true about water. One can get rid of that background belief and one's talk about water can continue to go on without having to embark on a course of mental gymnastics in order to avoid awareness of one's newfound knowledge about water. Same with motion-discourse. If my and/or Joyce's relevant theoretical claims about morality are in the right ballpark, moral discourse is not like these discourses in just this way. Now Joyce says (or can be read as saying) that since moral discourse is different in this way, the error theory is true. Finlay doesn't address the question directly, but seems to imply that even if moral discourse would lose an important function in the context of awareness of the truth of relationalism, even if people's commitments would be weakened,[355] that doesn't make the error theory true. But as I've said, I take it that the error theory has interest outside fairly narrow academic concerns only because of the (perceived) *threat* or *danger* it poses. That, in my view, is the most plausible explanation of why people object to it so strongly, and consider it a last resort, not because they are so concerned about uttering falsehoods per se.

The difference between Finlay's and Joyce's analysis is that for Joyce, what people mean by their moral claims commits them to error. For Finlay, this is not the case, but he grants that very many people might *believe* that they mean something absolutist by their moral judgments. Given the importance Joyce places on not believing falsehoods, Joyce would presumably find it just as important for people to know that *what they believe they mean by their moral claims is false*--and to know the truth about

[354] This is not the only reason, but an important one in this context. I also agree with Finlay (2008, 363) that morality, unlike water, 'lacks metaphysical depth'.

[355] Which he does acknowledge in his (2007a, p. 147).

what they *really* mean--as he finds it important that they know the truth about their moral claims themselves being false. What is common to either scenario is that we would need to learn to keep from view the fact that we are engaged in a 'manipulative' endeavor with others and/or ourselves. Similarly, even if Stevenson were right, Joyce would have the makings of a similar fictionalist project, on grounds that if moral judgments really are only commands then we should believe *that*, but since believing that threatens to muck up the motivational works, we should pretend that it isn't true.

Whether moral judgments are assertions or disguised commands, and whether their absolutist element is semantic or pragmatic (indeed whether there is any good answer to this question), the function(s) that they serve will tend to be compromised to the extent that we are aware that those *are* their functions and (roughly) how they work. Finlay says the fact that the absolutist component of moral discourse can be seen as pragmatic rather than semantic "belies Joyce's claim that nonabsolutist moral concepts 'could not so effectively play the roles to which we put morality, and thus we could not *use* it as we use morality' [2006: 208]" (358). Even if we bracket the possibility that the pragmatic point is accomplished *by means of* the semantic point, the spirit of Joyce's claim is still completely intact. That is, if we no longer have available to us the absolutist element of moral discourse, *even if* it is best analyzed as a rhetorical usage of semantically relationalist language, then we cannot use this discourse as we use moral discourse.

That is, we cannot put it to the *peculiar* use to which we put moral discourse. If

Finlay is right, much moral discourse doesn't employ the absolutist component. According to this view, on one end of the absolutism dimension of moral discourse, we should find people making moral claims neither semantically nor pragmatically employing any absolutist component. We would find people debating or discussing moral matters or making moral judgments with respect to ends that were perfectly well understood to be shared by the discussants. But I submit that such cases a) don't look like paradigmatically moral discussion at all and b) can in any case be readily be accomplished without moral discourse. There is just no need for the peculiarly moral terms of moral obligation, permissibility, duty and the like to have such discussions.[356]

What makes morality a 'peculiar institution', as Bernard Williams (1985) calls it, is that it does, at least sometimes, conceive of its demands as having authority over persons independently of anyone's particular ends or standards. There is an effective demand that if one doesn't have the relevant ends then one had better at least pretend to adopt them. If we are understood to share the relevant ends already (with the same priority rankings), then there is just no need for this peculiar element of moral discourse and therefore no special reason to employ it insofar as we have other means of discussing matters, and good reason to avoid it because doing so has the potential for confusion and counter-productiveness.[357]

As we move toward the other end of this dimension, we find people making heavy use of the absolutist component. Here we expect the participants to have a

[356] One might be tempted to object that these terms would be entirely appropriate if someone's acknowledged ends just are to fulfill their moral obligations, etc. But this would miss the point. The idea is that without absolutism, the notion of moral obligation has nothing special going for it in the first place over other alternatives. Further, it can easily be counterproductive as I will argue shortly.

[357] I'll say more about this below.

perceived lack of (or lack of perceived) relevant shared ends. However, if participants are quite *aware* of the pragmatic rhetorical nature of the absolutist component, then whatever pragmatic benefit it once had will be, at the very best, severely undermined. Under these circumstances, people might as well explicitly demand that others agree with them, or share their ends. With all due respect for the importance of empirical inquiry, I think it is clear that this strategy is a nonstarter.

So, even if Finlay is correct that semantically relationalist language can do what moral discourse can do, it cannot do it if we are aware of what Finlay grants that we are not generally aware of, namely the pragmatic nature of the absolutist component. To the extent that we are relevantly aware, we will not be able to use the discourse to whatever special purpose(s) it had been put to in the past.

Fortunately, neither will we be vulnerable to its serious downsides, many of which have to do with lack of awareness of our motivations. As I noted above, Finlay was wrong to require that people *correctly* perceive themselves as being in FMD in order to take their behavior as evidence for absolutism. It is sufficient that they think they are in FMD. And to the extent that I'm right about the motivational power of moral discourse relying on a deflection of our attention from our motives, we can expect that people will very often be mistaken about whether they are in FMD. As we saw in Haidt's research (though we certainly did not need his research to know it), people are very prone to interpret their own (group's) motivations favorably, and quite a bit less prone to that style of interpretation when evaluating others' motivations in the context of moral disagreement.

There is no simple means of dealing with others who have competing intrinsic ends (and/or rankings thereof) in the context of limited cooperation, but the problem cannot even be properly addressed in the absence of a better appreciation for what those ends are than we have at present.

5.4.1 The Antirationalist's Dilemma

I just argued that a self-conscious antirationalism faces problems very similar to those of fictionalism. However, there is an important difference between self-conscious antirationalism and Joycean fictionalism that is worth discussing in a bit more detail. As I mentioned above, I agree with Finlay that at least some moral discourse doesn't employ absolutism, or at least not to a large degree. Since the entire point of Joyce's fictionalism was to preserve the absolutist component of moral discourse and its motivational benefits, a critique of self-conscious antirationalism will have to take into account the possibility that the antirationalist is not attempting to preserve any or as much practical oomph as Joyce's fictionalist. I mentioned above that an antirationalist with no desire for oomph had no need for peculiarly moral discourse, and that using it can be counterproductive.

Here I want to summarize and extend my arguments against self-conscious antirationalism by presenting the self-conscious antirationalist with a practical dilemma. I want to show that 1) if she wants the robust oomph of the fictionalist, that invites the very serious problems of that proposal, and 2) seeking no special oomph voids the value of peculiarly moral discourse and threatens serious confusion in interpersonal communication as well as introduces fruitless and possibly intractable disagreement. As

I argued in Chapter 4, I am not interested in arguing that we should never employ moral discourse, including in an attempt to get some extra motivational oomph, whether in ourselves or others. There I said that an abolitionist as I see her is not someone who must fanatically avoid moral talk or thought at any cost, any more than an atheist[358] is someone who must avoid religious talk at all costs. I'm not out to prevent people from telling themselves or others under any circumstances that morality or God demands this or that. My point is that we often do and should reflect seriously about what to, and when we undertake these normative investigations, moral discourse is often if not always best put to the side in favor of other ways of conducting this reflection. This is especially true for those who already accept either an error theory or antirationalism about moral discourse.

I've been arguing that whether or not antirationalism is true of our moral concepts, awareness of this fact will tend to undermine whatever special point there is to moral discourse. But someone wanting to keep this special point might want to adopt a similar strategy to Joycean fictionalism. They could act as if moral judgments have a special sort of authority while doing their best to remain unaware of their own strategy. In chapter 4 I argued that this was psychologically implausible. But perhaps I am wrong. Perhaps (at least) some people can do it. Then I argued that if successful, people are in effect deluding themselves. But perhaps someone is not very concerned about this. Perhaps their concern for maintaining their moral commitments is greater than their concern for avoiding self-deception in this domain. Then I argued that there are serious

[358] I realize that there is a disanalogy between abolitionism and atheism in that the first is a practical proposal and the second reflects a belief about the world that is consistent with proposing a religious fictionalism (where is the champion of this proposal?). But since to my knowledge no atheists propose to conduct our normative discussions via god-discourse, the reader can safely read 'atheist' to mean someone who does not favor god-discourse as an important component of our normative discourse.

downsides to the motivational tools that peculiarly moral thinking employs. Those downsides include the undermining of the very commitments that one is attempting to strengthen, both in oneself and others. This happens due to the motivation to misperceive one's motives such that one is not in fact violating any commitment. The sense of practical oomph that morality employs makes one feel pain and/or weakness at the perception of a violation, which motivates one not to perceive violations. This is a very important and under-appreciated downside of both willpower and moralistic thought. There are other potential costs as well, such as loss of flexibility and a sense of being imprisoned by one's moral convictions. All these downsides militate strongly against undertaking the kind of significant effort required to achieve/maintain a peculiarly moralistic mindset in the face of an awareness of the nature and function of such a mindset and the nature of practical reason.

But what about an antirationalist who makes no attempt at retaining a sense of practical clout? Fortunately, Finlay seems to be just such an antirationalist, and some of the downsides of this strategy are evident in his writings. Joyce and I have both argued that a moralist with no interest in clout has no clear reason to use moral discourse at all. Why not simply speak directly in terms of what one wants or cares about, what one is willing to tolerate and punish, etc.? Perhaps the antirationalist will answer that doing so would come across as strange, since this is not how most people talk. Perhaps it is better to speak in moral terms when among others who are doing so.

I think that this is the opposite of correct, for this is likely to lead to serious confusion, or at best, excursions into metaethics that will rarely be practically advisable. For example, Finlay (2007a) believes that our concept of morality is purely and radically other-regarding. That is, the demands of morality are roughly on the order of what Singer would have us believe they are on the strong version of his argument that I presented in Chapter 2. According to Finlay, morality makes extremely severe demands on us, at a minimum to do whatever we can to alleviate harm to others. In fact, Finlay disagrees with Singer in that Finlay thinks moral demands take *no account* of our self-interest (2007a, p. 141). Of course the difference between Singer and Finlay is the way that they think about the rational authority of these demands. Singer, without saying so explicitly, seems to think that we rationally ought to be doing such things, while Finlay thinks that most people don't care much about morality and to that extent have no strong reasons to act as morality requires.

What Finlay thinks people *do* care much more about is the extent to which their moral communities demand that they conform to the demands of morality (2007a, p. 142). There is a lot to be said on this topic, but for the moment I just want to note how different a way of thinking this is from the way most people think about morality and its demands (which Finlay recognizes). If Finlay wants to have a conversation with someone about what would be the moral thing to do in some situation, his judgments are almost certainly going to be different than theirs, since by hypothesis most people make a strong association between the requirements of morality and what they ought to do, all things considered. So if Finlay wants to have a conversation with someone about what is morally required, he is going to have to explain that he thinks that what morality requires is just one source of concern among many, and one might perfectly rationally care about it a lot, a little, or even possibly not at all. He might even add that he thinks most people don't care very much, and that there is no rational pressure on them to do so, unless it would help them avoid punishment or reap other rewards. If he does not explain himself in some such fashion, he is going to run the risk of serious miscommunication, since in the vast majority of situations, people will infer from his judgment that he thinks X is morally required that he thinks that people really ought to be doing X. I agree with Finlay that it's plausible that people deceive themselves about the requirements of morality because they want to think of themselves as far more in keeping with moral requirements than they really are (2007a, p. 145). But they also resist this conclusion because they do not see moral requirements in terms of conceptual truths that could diverge wildly from what is practically rational for them to do.

To the extent that these suppositions about people's motives and beliefs about morality help to explain why people would have strong but erroneous intuitions, such observations might help Finlay's argument for what morality does in fact require, as a conceptual matter. However, it does not make the idea of actually employing the discourse with such people (or anyone else) attractive. On the contrary, it appears to hold out the promise of serious confusion, miscommunication and/or having to take a detour through metaethics and the nature of practical reason in order to get on with practical discussion. Much better to try to get at what the person actually cares about. If it turns out to be what is required of them in order to impartially help others as much as possible, so be it.

But perhaps Finlay is wrong about what is required from the moral point of view. Will this help? Not much, if at all. For an antirationalist, it is hard to see why it matters what is required from the moral point of view per se. If the moral point of view requires impartial other-regarding concern, then it is hard to see why that matters over and above one's impartial concern for others. If it requires never lying or cheating or stealing, it is hard to see why that matters over and above one's independent concern for these things and their practical upshots. For an antirationalist who has given up entirely on any special clout for morality, then the fact that something is required from the moral point of view has no particular relevance that I can see. Either the point of view that is considered the moral one is a point of view that attaches to some concern of the agent or it is not. Discovering that that point of view is, after all, the moral one should have no ability to create or strengthen any motivation to adopt or act in accordance with that point of view.

In his defense of conceptual antirationalism, Brink (1997) argues that even on an instrumental conception of reason, moral reasons can retain quite a bit of authority, just insofar as most people have impartial other-regarding attitudes. But as he notes, this makes those requirements hostage to those attitudes (32). He adds that the more traditional defense of moral requirements resides in an attempt to show the coincidence between moral requirements and those of self-interest (where the authority of self-interested reasons may be taken for granted). For example, we benefit when others cooperate with us, but those benefits are conditional on our own cooperation. Therefore we generally have reason to cooperate, where that is taken to mean following familiar other-regarding norms of restraint and aid.[359] This connection between morality and self-interest is admittedly contingent and only generally reliable, but it might seem that the fact that most of us do have other-regarding desires as well as rational concern for ourselves that depends on our acting 'morally' provides plenty of authority for morality, as well as providing grounds for reasoning directly about what is morally required of us.

[359] I would add that we have similar kinds of reasons to punish and condemn.

A defense that attempts to tie moral requirements to those of self-interest must assume that Finlay and Singer (and many others) are wrong about what morality requires. For on their views, morality makes demands on us that take us well beyond what we can hope to be repaid in terms of the kind of benefits we get from following other-regarding norms within our own communities. And even though many of us have other-regarding concerns, those concerns do not make it rational for us to adhere to anything like the full demands of morality if they are roughly what Singer and Finlay (and arguably the moral saints) think they are. My point isn't that Finlay and Singer are necessarily right about what morality requires (though I am sympathetic to their claims). My point is that there is famously quite a bit of disagreement over what morality does require, and that for an antirationalist to engage in such disputes strikes me as an entirely academic enterprise. I have nothing against academic enterprises as such, but the question I am posing here is whether an antirationalist who has given up entirely on moral clout should actually *employ* peculiarly moral discourse in her serious normative investigations and discussions. While it might be interesting to investigate conceptual truths about what morality requires of people, for such an antirationalist the answer to this question has no bearing that I can see on how she should live her life. Additionally, for an antirationalist seeking no moral clout, judging that morality requires impartial concern adds nothing to whatever level or kind of impartial concern one already has.

I think normative discussion is very important, both intrapersonally and interpersonally. I also think it can be very difficult. There are often very many things to take into account, and it is hard to know how to evaluate and weigh competing considerations, and in many cases it is hard to be quite honest with oneself and others, for so much of our reasoning is motivated in ways that on reflection we find incompatible with what we care about most. I find myself unable to see how a cloutless antirationalist has anything to gain by introducing a layer of potential disagreement that even if resolved would have no clear relevance to the question of what is to be done. In a group of cloutless antirationalists, any such discussion and disagreement must be purely of intellectual interest. But plausibly, the vast majority of *actual* disagreement over what morality requires, even among avowed antirationalists, is due to competing conceptions of how to live. Since the vast majority of people regard morality as having some kind of authority over us that is not contigent on how much we care about moral requirements, disagreements with such people will not be of purely academic interest. But a cloutless antirationalist who in such company argues about what morality requires is incurring not only the likely cost of miscommunication, but also introduces or reinforces a layer and kind of disagreement that is, by hypothesis, not well-suited to investigating what we care about and how to promote it.

Now it might be thought that I have foisted a false dichotomy on the antirationalist. I have given her the option of attempting full clout or none at all. But why couldn't a certain kind of antirationalist take on board everything I've said and just tweak her practices a bit? It's plausible, as I've just said, that the moral point of view is the point of view of our impartial concern for other people. I have said nothing to rule out the possibility that (some) people do have very strong impartial concern for others. In fact, I take myself to have very strong other-regarding concerns.[360] While I can

[360] Though I think it can be misleading to call them impartial. I like, and have concern for, certain kinds of people more than other kinds. I think this is true of almost if not all people. To the extent that the moral point of view is that of true and full impartiality, I agree with Finlay that people don't care much about this

acknowledge that I don't care whether this turns out to 'really' be the moral point of view after all, I can also acknowledge that I care a lot about this point of view and that this conception of the moral point of view is a common one. Therefore, why not use the language of morality to give this point of view a boost? I needn't become so deeply immersed that I'm self-deluded, and I also avoid the downsides of will to the extent that I am not deeply immersed. But since moral terms do have emotional resonance, why not use some of that resonance to motivate myself and others to adopt and act from (what we can at least tentatively call) the moral point of view?

It is not my intention to say that such a strategy isn't workable, but to point out its limitations. The first limitation is that with increased psychological plausibility comes decreased motivational efficacy. In the back--and often in the front--of one's mind will be the thought that one is manipulating one's own motivations, which will have an undermining or destabilizing effect. Second, the clout mechanism still has the downsides. True, with an 'intermediate strategy', the downsides will be lessened, but so is the upside. Third, and more important, given the relatively shallow level of immersion, one will not be strongly motivated to engage in much protracted discussion about what is actually morally required. Knowing that one is using the language of moral requirement to give one of one's motivated perspectives a boost, arguments about whether morality 'really' requires what Kant or Mill or Singer or Finlay or Brink or anyone else thinks it does are not going to be perceived as practically useful. That is not to deny that they might be intellectually stimulating, but I am considering the value of employing moral

point of view. But I and many people care about others not of their own nationality, race, gender, sexual orientation, political affiliation, family, tribe, and many other characteristics that are the paradigmatic sources of partiality.

discourse as a principal means of conducting our normative investigations. For such an antirationalist, I see no reason or motivation to engage in serious discussions about what our actual moral obligations are, where the point of the discussion is to figure out what to do. Without any significant attempt at immersion, one would expect her to begin to say something more along the lines of 'Look, these people are [starving, dying, being oppressed, destroying the environment] and I care about that, and I bet and hope that there is a part of you that does too, and we can do something about it without too much sacrifice.'

Of course I see this as a good thing. So my point is not that the final sort of antirationalism I described above isn't workable; on the contrary, to the extent that one does not engage in fruitless discussions about what is morally required, it is just the kind of antirationalism I encourage. It is an antirationalism that does not attempt to use peculiarly moral discourse to do any heavy normative lifting, and so does not allow moral discourse to (significantly) blind us to what we care about or the nature of our commitments and motivations.

5.5 Conclusion

Joyce (2011a, p. 520) recognizes a form of error theory that claims no false presupposition at the heart of moral discourse. This error theory is arrived at by judging 'all alternative metaethical theories irreparably defective'. An error theorist of this stripe has high confidence that no successful vindication of morality will be forthcoming in light of the over 2,000-year failure, by some of history's greatest minds, to achieve it.

Joyce calls this "the 'Enough Already!' Argument". Presumably this can be cashed out as saying we've had 'Enough Already!' of rationalizations in support of the 'practical oomph' of morality.[361]

Joyce and Finlay both eschew any such rationalizations, though both do so at the price of having to acknowledge widespread if not systematic ignorance about the nature of (moral) motivation. I've been arguing that any *essential* error there is to moral discourse is practical, not theoretical. It is the error involved in the motivated avoidance of certain kinds of knowledge or awareness, mostly involving one's own motivations.[362] I am of course keenly aware of the danger of this sort of awareness (and of mistaken beliefs or confusions about one's motivations), and that there are arguments to be made that we should remain unaware of our motivations in much the way that moral discourse helps us to do.

I argued in Chapter 4 that this is what Joyce's fictionalism amounts to. I tried to show that Joyce's arguments for fictionalism, despite his intentions, amounted to arguments that we should deceive ourselves about our underlying motivations in order to retain the motivational benefits of moral discourse. It is to his credit that his arguments for fictionalism identified the primary considerations in favor of *not* developing a better awareness of our motivations, and it is also to his credit that he recognized that a successful fictionalist would have to be deeply immersed in the fiction, with his attention

[361] Of course neither Finlay nor any other cloutless CAR needs any such rationalization; but they can only get away with not having one by ackowledging that people can be ignorant of their own concepts and motivations in just the way that motivates the attempted rationalizations at issue.

[362] Again, in saying that this sort of error typifies moral discourse, I mean to recognize that there are plenty of its users who are quite self-aware in many important ways. In fact, in Chapter 4 I argued that moral emotions are likely to increase commitment in those who tend to be aware of their own relevant motivations. Nevertheless, I believe there are other respects in which self-awareness is much less common.

very rarely loitering on his most fundamental goal(s).

I have a lot of sympathy for the ‘Enough Already!’ Argument. I agree with Joyce that we have had enough rationalizations for why morality has its practical oomph. I disagree with Joyce’s arguments to the effect that what we need instead is fancier and more complicated methods of evasion and jamming[363] in order to retain it. I would like to double down on the ‘Enough Already!’ Argument by extending this sentiment to not only rationalization, but to the other mechanisms of self-deception that are essential to his fictionalist program, as well as any antirationalist attempt to retain the practical clout of moral judgment.

Joyce rightly perceives that morality cannot perform its peculiar function without the absolutist component. While Finlay does not seem to recognize (or care about) the threat posed by a general awareness of the function of the absolutist component, Joyce doesn’t see that whatever essential and important error there is in moral discourse is reproduced by the successful carrying out of his fictionalist program. The important error in both Joyce’s error theory and Finlay’s antirationalist response is two-fold. They fail to recognize the essential nature of the error as consisting in a motivated lack of awareness of our own motivations, as well as the exciting opportunity we have to begin to develop greater awareness of ourselves by no longer adverting to (or leaning heavily on) moral principles, obligations, rights, duties, responsibility, desert, and the like in our normative investigations.

Joyce and I have both argued that a design feature of moral concepts is that they help to avoid the kind of rationalization that undermines commitments. The loss of this

[363] Bach’s methods of self-deception from Chapter 4.

feature constitutes the threat that Joycean fictionalism is motivated to avert. Nothing I have said above about the kind of rationalization involved in moral discourse should be taken to imply that I disagree with Joyce that it is *also* very successful in *preventing* rationalization.[364]

Responding only to the perceived danger, Joyce presents his fictionalism as a means of allaying the fear and loathing with which people often respond to the error theory. Ignoring the danger entirely, Finlay misses that his case for the truth of some first-order moral claims doesn't address the *spirit* of Joyce's error theory, which is that peculiarly moral discourse cannot survive the discrediting of absolutism. Unfortunately, the extent to which Joycean fictionalism is successful in averting this perceived danger is the extent to which the *essential* error, which is not theoretical/conceptual but *practical*, is reintroduced.

In the next chapter I'll try to convince the reader to welcome the important and admittedly dangerous challenge of attempting to know ourselves better, most specifically what it is that we care about.

[364] See section 4.2.4 for discussion of how moral discourse can be expected to both prevent and promote specifically akratic forms of rationalization, depending one's disposition to self-awareness.

Chapter 6: High Hopes

6.0 Introduction

My primary goal in this chapter is to make a case for engaging in 'straight talk' as (a large part of) a replacement for peculiarly moral discourse. The way in which I use this term is different than Joyce's sense of 'straight talk'. First, and most importantly, I want to broaden it by including the active inquiry into our values and motivations. Second, for Joyce, 'straight talk' is limited to self-interested or instrumental reasoning. What I mean by it includes this, but also includes talk about our cares, concerns, values, emotional and/or aesthetic responses, motivations, commitments, and the like, including which seem to us most central, important, deep, to have greatest priority, etc. Straight talk is not limited to descriptive claims however, but can include normative ones about what we should do or value or be motivated by. But insofar as they remain forms of straight-talk, there will be no attempt to disguise the fact that these claims proceed from end-relational, or motivated perspectives. Claims to the effect that one should give to the needy, or keep one's promises, or not be (so) envious are straight talk to the extent that one is aware (or does not resist awareness) that the normativity of these goals is bound up with one's motivational set—that they are made 'from the point of view' of one or more motivated perspectives. It would be a distraction and probably foolish to attempt to characterize what I have in mind by straight talk more rigorously at this point, beyond the crucial idea that the whole point is lost if we don't try to be *honest* with each other—and most of all with ourselves—in this endeavor.

A helpful way to think of the value of straight talk is as a means of internalizing descriptive and normative subjectivism. Descriptive subjectivism is the claim that there are no practical reasons or values metaphysically independent of our contingent (dispositional) motivational states. That view leads to a kind of practical nihilism unless one also rejects normative objectivism, which is the claim that the *authority* of our values—or perhaps only our 'highest' values—*depends* on their having objective standing. To the extent that we accept descriptive subjectivism, unless we thoroughly reject normative objectivism in favor of normative subjectivism, we have a nihilistic cocktail. That is, normative objectivism and descriptive subjectivism entail the rejection of the normative authority of not only our current values, but any values we might come to possess.

Chapters 1 and 2 had as their central goal to explain away the strong attractions of both normative and descriptive objectivism, and Chapter 3 argued directly against normative objectivism (though in different terms). I argued there that practical justifications do not require reasons that are unconditional, or rationally necessary. I argued for the 'Schopenhaeurian' picture of practical reason, on which our inclinations are not limitations on our practical reasoning, but rather conditions of its possibility. I tried to show that our motivated perspectives 'provide the terms in which we think and reason'. If we give them up, we are not left with an 'undistorted, unadulterated representation of the 'good as such,'' but with 'no evaluative judgment at all'.[365]

[365] Reginster (2008, p. 85). I am indebted to him for his discussion of the problem of nihilism as resulting from the combination of normative objectivism and descriptive objectivism. He is giving an account of how Nietzsche both conceived of and attempted to resolve the problem of practical nihilism. My project bears many important similarities to Nietzsche's, as both Reginster and I view the latter. Sadly, I cannot discuss those similarities here.

Therefore I think internalizing normative subjectivism is important to avoid a kind of practical nihilism, where we accept descriptive subjectivism but still hold on to (shadows of) normative objectivism.[366] Moral discourse is built for descriptive and normative objectivism; as we have seen and will see, it is very poorly suited to subjectivism. Therefore to internalize normative subjectivism, it's important to stop speaking in moral discourse, and to start getting comfortable locating the normativity of the moral domain in our actual concerns.

In chapter 4, I described my imagined abolitionist 'A' as someone who

> has not only interests but values. He not only has values but he understands the great importance of commitments in securing those values against specious rewards. He also understands that commitments, important as they are, sometimes warrant exceptions. He knows that there is no decision procedure for how to decide when to violate a commitment; he knows that employing good judgment and developing good habits and (self-) perceptual abilities and being vigilant against self-deception are all important but not sufficient for living a good life.

A knows these things and makes no sustained attempt to keep her awareness from the fact that her reasons are relative to her values. But there is one important aspect of a straight-talker that I did not emphasize there, and that is the importance of attempting to *discover* what one's (deepest) values are.

At the end of that chapter I said that Joyce thought that he could avoid Hinckfuss's worry that morality commits us not only to good but bad things, by simply moralizing those of our values that are 'already useful'. Again, it is hard to overstate the importance of rejecting the idea that the only way to 'step outside morality' is to regard

[366] That many people currently hold something like this combination I believe partly explains the apparent prevalence (among undergraduates at least) of what Williams (1985) called 'vulgar relativism'.

our moral(ized) values as valuable *only* to the extent that they are instrumentally useful for some nonmoral(ized) goals. Rather, we can and should step outside morality by regarding the value of *all* our values in terms of what we actually care about, or what matters to us. But more important for this chapter, I think the primary hurdle in the task of figuring out what we *ought* to value, in what ways and how much, is that we do not know nearly as well as we could what it is that *does* matter to us—what it is that we *do* care about, and with what priority rankings. I said in response to Joyce that figuring out the answers to these questions is not the work of an afternoon or a fortnight subsequent to which we can set about moralizing what we come up with (even if that were a good idea). Figuring out the answers to these questions is nothing less than the work of getting to know (and perhaps creating) oneself, which is the work of a lifetime.

This chapter can be seen as responding to three possible reactions on the part of a normative subjectivist to my call for straight talk. The first reaction has two parts. It questions the importance of an investigation into what we care about. This could be because it is unclear what the normative significance is of what we do care about to the question of what we should care about and/or because an 'investigation' is unnecessary since it is fairly trivial to discover what we care about. Section 6.1 will argue for two related claims. The first is that for a normative subjectivist, practical justifications must bottom out (roughly speaking) in terms of what we care about. This is perhaps better construed as a reminder than an argument, since this sort of conclusion seems to fall right out of normative subjectivism. And yet the reminder seems required, since even among those philosophers who believe that our reasons are relative to what we actually care about,[367] there is surprisingly little discussion of the importance of finding out what those things are. That this is a difficult task is my second claim. Indeed, my idea of a post-moral normative discourse centrally involves much more effort and attention devoted to investigating the nature of, and relationships between, our actual commitments and values.

In section 6.2 I'll address a view that takes the project of internalizing normative subjectivism seriously, and agrees that finding out what our values are is an important task, but doesn't see any significant conflict between these projects and 'morality'. I take Jesse Prinz as the representative of this viewpoint. He explicitly rejects any attempt to move beyond 'morality', understood as the moral sentiments. I have no problem with the sentiments as such, but think that by his own lights as a subjectivist, moral discourse is very ill-suited to promoting what we both roughly agree would constitute 'moral progress'. Indeed, I'll argue that his own metaethical view is plausible only if moral discourse (including the moral sentiments) systematically deflects attention from our motivations and values. I'll show that by his own lights, that is a very bad thing.

The final reaction I'll address in section 6.3 acknowledges that moral discourse distracts us from what we care about in important ways, but thinks that the dangers of opening up that pandora's box are too great, or at any rate not worth the trouble and risk. My response to this reaction will be to admit the danger, admit that I don't know how well it will work, but argue that it is nevertheless—at least for many people—the best option. An important part of my response here will be to try to bring home the danger of

[367] Joyce (2001) does not. There he defends an idealizing account of practical rationality.

not internalizing normative subjectivism, given a disbelief in, or lack of commitment to, descriptive objectivism. But I will also allow that in the end, the answer will probably depend on what the person cares (most) about.

Before I start, I want to forestall a large potential confusion. I can see how the reader could get the impression from my discussion so far that I am fanatic about always having one's attention directed toward toward one's goals or motivations. This is very far from the case. I am far from promoting a view that advocates having our attention on our fundamental or underlying motivations all or even most of the time. The view I am promoting is that peculiarly moral discourse at its very best addresses these motivations indirectly, and at its worst departs wildly from them in severely pathological ways. My point then is not that in conducting our daily affairs—not even when these are squarely in the socio-moral realm—should we be attending always to our own motivations. It is not always, or even usually, the time for reflection on what we should do or why we should do it. My point is that when it *is* time for such reflection, that reflection could use much more—and more honest—investigation into what really motivates us and what we actually care about than moral discourse characteristically affords.

6.1 The Importance of (Finding Out) What We Care About

Moral discourse (including associated sentiments) contains powerful motivations not only to misperceive our motives, but also our values. Norms about which motivations are blameworthy or praiseworthy tend to motivate perceptions of one's motivations as being among the praiseworthy, or at least not among the blameworthy. I hope this has achieved the status of truism. At any rate, that we can and frequently do misperceive our motives surely is a truism. Yet the same is true if we substitute 'values' for 'motivations'. This *should* be a truism, and yet the fact that we can mis- or not perceive our own values seems very much less commonly appreciated.

Presumably this fact helps explain why there is so little—to my knowledge virtually no—discussion of the (normative) importance of *finding out* what our values are. It is not so surprising that normative objectivists don't find this project terribly important. After all, what we happen to value is not in itself of any clear normative significance for them. But it is fairly surprising to find such an apparent lack of interest in this subject among normative subjectivists.

In his landmark (1988) paper 'The Importance of What We Care About', Harry Frankfurt argued that what we actually care about has normative significance for us by simple virtue of the fact that we care about it. Also, given the importance of what we care about, for anything that we care about we can and should ask if we care about caring about it. Indeed, he thought that 'if *anything* is worth caring about, it is worth caring about what to care about (92). For we can surely make mistakes in caring about certain things, by caring about them at all or in the wrong ways or too much.

Frankfurt imagines a person who cares about avoiding stepping on the cracks in the sidewalk. It seems that this person is making a mistake by 'caring about, and thereby imbuing with genuine importance, something which is not worth caring about' (93). He immediately forestalls any objectivist interpretation of what this mistake would amount to: 'The reason it is not worth caring about seems clear: it is not important to the person to make avoiding the cracks in the sidewalk important to himself' (93). As we saw in

Chapter 3, an objectivist might be tempted to answer that for all we know, it *is* important for some person to make it important to himself, and this shows that the foolishness of caring about crack-stepping cannot be cashed out in terms of contingent motivational sets. Rather, it is unconditionally, or intrinsically foolish. But we are past that now.

Nevertheless, there is something unsatisfying about Frankfurt's answer to the question of what makes this person's concern about crack-stepping an error. It connects with Brink's worry that appeals such as this presuppose rather than explain (ir)rationality or what makes values important (even to us).[368] We might put the worry in a slightly different way than Brink did in Chapter 3. We might ask, 'How do you know that the person has this higher-order concern or lack thereof?' Of course Frankfurt or anyone else can stipulate the concerns involved in any given case. But that won't do here because Frankfurt's answer is clearly not meant to rely on his own stipulation. Rather he is supposing that if we were to find an actual person who cared about not stepping on the cracks, we would also find that they would not care about caring about it.

I expect Frankfurt is right in this case. But it's important to see that his claim that the person *is* making a mistake is legitimate if and only if he actually does not care about caring about stepping on the cracks.[369] Now, although I suspect that there aren't many people who care about caring about avoiding stepping on cracks in the sidewalk, I also suspect that there *are* patterns of concern, in ourselves and in others, that would surprise

[368] See section 3.21. Importantly, neither Frankfurt nor I (nor Finlay, as we'll see) appeal to *idealized* desires or concerns to explain the irrationality (or mistake) of this person. Our conception of the error in caring about stepping on cracks is in terms of one's *actual* concerns.

[369] Please take the complicating features discussed in Chapter 3 as given. For example, he might care about caring about it for reasons derivative on some other intrinsic concern in conjunction with (presumably) false beliefs. I will avoid such complications in this chapter unless they are directly relevant to my point.

us.[370] And yet Frankfurt doesn't indicate that any part of 'The Importance of What We Care About' is *discovering what those things are*.

It's worth emphasizing how strange this is before moving on. The title of his paper is to be taken in two (main) senses. The first is that what we care about is by that fact alone important to us. The second is that since the first thing is true, it also must be important to us to care about things that are worth caring about. That is, it's important (to us) what we care about. And finally, when Frankfurt addresses the question what makes something (not) worth caring about, the answer is given in terms of what else we (don't) care about—specifically what we (don't) care about caring about. So the answer to the vitally important question of what we *ought* to care about depends (in large part at least) on the answer to the question of what we *do* care about.[371]

Normally, if you were to read a highly-regarded philosopher conclude that the answer to the vitally important question of what to care about depends on the answer to the question of X, I think you would expect them in short order to give some indication that we had better turn our attention to X. If someone argued that the answer to the question of what to care about depends on which things have intrinsic value, for example, you would expect them to think that we had better go boldly forth and find out what those things are. But not here.

My point is not to criticize Frankfurt, whose paper I otherwise admire greatly. My point is to illustrate that even where one might most expect to find it, there seems to be no interest in the question of finding out what we actually care about. There is

[370] In fact, as I indicated at the beginning of this section, we should be very surprised if we were *not* surprised.

[371] It also depends on what things we are able to care about, as he says at the very end of his paper.

certainly widespread interest (as we saw in Chapter 3) in the question of what the normative significance is of what we care about, and even in the question of the normative significance of different orders of desire relative to one another. For example, does the sheer fact that we don't care about caring about something remove or lessen the normative significance of the thing we care about? Does the sheer fact that we do care about caring about something that we don't actually care about give us good reason to care about it? I think these are interesting and important questions, but the answers to them will be of quite limited importance if we don't even know what those various things are that we care about, or care about caring about.

Perhaps the failure to recognize the importance of finding out what we care about stems from the belief, either implicit or explicit, that we already either know the answers or that they are in any case easily discovered. I think this is the opposite of the truth in many important contexts. It is for several kinds of reasons quite difficult to know what is important to us, even when we try to find out.

Finlay (2007a) avoids the first mistake, but makes the second one. That is, Finlay seems to realize (in fact, his views require that it be so) that people are commonly mistaken, even self-deceived about what they care about. As we saw in section 5.4.1, he holds that most people care more about what their moral community thinks of them than they do about their actual obligations. I'm fairly sure he would agree that most people don't realize that. But he does seem to think that one can read off people's cares and their relative levels of strength fairly straightforwardly from their behavior. In pointing out where Finlay goes wrong here, one of my aims is to illustrate to the reader that it is in fact quite difficult to tell—in many cases at least—what someone cares about and how much.

Recall that Finlay thinks reasons are end-relational. In his (2007a), he is explicit that the question of where one's strongest reasons lie depends on the answer to the question of what matters most to that person (153). This view of the strength of reasons plays an important role in his conclusion that—contrary to the opinion of most philosophers in the history of (Western) philosophy, neither reasons of self-interest nor those of morality have any special rational authority or significance. That is, not only does Finlay think that neither of these sources of reasons has any *intrinsic* rational authority, he thinks that they also don't generally constitute our strongest reasons because most people simply don't care much about them. What I'm interested in now is how Finlay comes to this last conclusion.

First, let's get clear about what Finlay regards as the content of the concepts of morality and self-interest. I'm not going to recapitulate his arguments, but we need to get a general idea of what he means—and what he thinks most other people mean—by these terms. First, he thinks that moral concern is (conceptually) purely and radically other-regarding. Second, self-interested concern is that which is meant to promote 'a life containing intrinsically rewarding pursuit and/or accomplishment of goals that I strongly and intrinsically care about at the time of my pursuit/accomplishment of them' (149). So for him the question whether we ought to comply with morality or self-interest depends on the question whether we care more about our own interests or those of anonymous others (154).

One might expect that if this is correct, morality is in poor shape indeed. However, Finlay thinks that self-interest may fare no better:

> First, it's vital to note that for virtually all of us *both* moral and self-interested ends are relatively low in our order of priorities, ranking well below our personal projects … We care more about having a nice house, car, or lawn for example, than we do about distant human misery; more for our favorite food and sports teams than for some level of health or future happiness … therefore … "Morality *or* self-interest" is a false dichotomy. (154)

I think that 'morality or self-interest' is a grievously false dichotomy. But here I'm concerned with how Finlay is addressing what he explicitly takes to be a complicated psychological question, what it is that we care (most) about (154).

A final complication Finlay addresses is that there seem to be large differences in the relative priorities of morality and self-interest for different people. His obervations lead him to conclude that while some people seem to care greatly about their self-interest, others seem to care hardly at all. And while some people seem to care a lot about anonymous other people, others seem to care very little. Now for all I know, this is correct, but I want to suggest that the psychological question is much more complicated and difficult than Finlay suggests here.

The claim that some people care hardly at all about their self-interest is supported solely by the parenthetical observation that there seems to be a lot of 'carefree' drug use and other willfully self-destructive behavior. The claim that many people don't care much about anonymous others is supported solely by the parenthetical invitation to 'witness widespread attitudes in the United States toward wars on foreign soils' (154). I certainly don't read Finlay as thinking that he has made convincing arguments based on only these comments. Again, my goals in what follows are not so much to critique Finlay as 1) to illustrate both the difficulty and importance of finding out what people actually do care about, and 2) to show that even for someone who is explicit that the question of what to do comes down to the question of what one cares about, there seems to be strikingly little interest in taking seriously the question of what people actually do care about.

While Finlay avoids the mistake of thinking that people can reliably introspect their own concerns, he makes the equally seductive error of thinking that one can read off what people care about fairly straightforwardly from their behavior. Can one infer from the fact that a person uses drugs that they don't care about their self-interest? Of course many people with bad drug habits try to stop, one plausible explanation of which is that they care about their self-interest. But in describing their substance abuse as 'carefree', Finlay was presumably referring to people who aren't trying to stop. Of course for their 'abuse' to be evidence for low self-concern, the person had better have good reason to believe that it is genuinely and significantly harmful in the long run. It is far from clear that many forms of casual drug use are significantly if at all (net) harmful in the long run, or that those engaging in it have good reason to think it is.

But of course there are very many cases where people do have good reason to think it is harmful. So must these people not care much about their self-interest? Many of these people have good reason to think that it is harmful to them, but deny it nevertheless. Presumably this is often motivated and self-deceptive. If it is not self-deceptive, it is not evidence of low self-concern but simply poor information and/or reasoning. But if it is self-deceptive, why do they deceive themselves? The most natural explanation in many cases is that they *would care*—perhaps a lot—to know that they are harming themselves in this way. Indeed, wherever one finds people engaging in behavior that they are in an epistemic position to believe with high confidence will be harmful to them in the long run, but self-deceptively believe that it will not be harmful, the most straightforward explanation of why they do so is that they *do* care about their long-term well-being. So, if such people's mistaken beliefs are not the products of self-deception, it is no evidence for lack of concern; and if they are, it is (prima facie) evidence *for* self-concern.

But surely in the (much narrower set of) cases where people *admit* that they care more about, say, smoking, than their long-term well-being, we can reliably infer that they actually do care more about smoking, which suggests low self-regard (or very high smoking-regard). Certainly we can at least rule out self-deception in these cases. No, we cannot! Imagine a person who decides to quit smoking for reasons explicitly to do with their long-term health. This happens all the time. Suppose they fail. They re-commit, try again and fail. Perhaps a third time. Now they are faced with two broad possibilities. They can see themselves as weak or they can see themselves as not caring so much after all about quitting smoking. Perhaps they don't care as much as they thought they did about their long-term health, or perhaps they care more about smoking than they thought they did. Or perhaps not. There is always pressure to re-evaluate one's commitments and values when one finds oneself violating those commitments and values. Perceiving ourselves as violating our values makes us feel weak, guilty, ashamed and so on. There is always pressure in such cases to either 1) not violate the value, 2) not perceive oneself as violating the value, 3) re-evaluate/modify the value or 4) *perceive oneself as having re-evaluated or modified the value*.

These are of course not exclusive options. A smoker might well deceive herself into thinking that she doesn't care very much about her long-term health (or that she cares very much about smoking) to avoid feelings of weakness or shame, and in doing so, and in continuing to act accordingly, she may shape her actual concerns to match or resemble her view of them. The point here is just that these things are awfully complicated. Given what we already know about motivational psychology, there should be nothing shocking about a person sincerely asserting that they don't care much about even their own self-interest, but being mistaken via self-deception.

The same thing goes for concern about 'distant misery'. I'll use Finlay's example because I think it is instructive. Again, many Americans do believe that U.S. wars on foreign soil are to the ultimate benefit of (most of) the foreign people—and if not to them then to their descendants. No matter how mistaken, it doesn't follow that they don't care. Now Finlay (or anyone) might respond that if people really cared, they would do a much better job than they actually do finding out whether views like this are defensible. Let's grant that many people don't put much effort into finding out whether such wars are really likely to benefit foreign others. Does this mean that they don't really care? Of course not. Again, a plausible reason why they might not look very hard is that if they found out otherwise they would be very upset. They need not of course be conscious of this reason for their not looking. This just brings home the point again that self-deception and willful ignorance are most naturally explained in terms of a person trying to keep from themselves something that they *do* care about.

I think a plausible response on Finlay's behalf is that in many such cases, people don't actually care, but they do care about caring. That is, they don't really care (much) about foreigners' welfare, but they self-identify as someone who does, so they care about caring for the foreigners' welfare, even though they don't care (much) about it. To foreshadow the discussion of Prinzean sentimentalism below, it could be that they don't have much in the way of sentiments of disapprobation resulting from the perception of harm done to foreigners by the U.S., but they do have a sentiment of disapprobation toward this fact about themselves, were they to perceive it. That is, they might not really feel bad about the U.S. harming foreign others in the (perceived) national interest, but they would feel bad to know that about themselves. And that very fact, as well as the predictable sentimental responses they could expect to receive from others in their 'moral community' were they to recognize and voice such a lack of concern for foreigners, could go a long way to explaining why they might think they care much more than they really do.

I think this kind of response is entirely plausible, and is perfect grist for my mill, requiring as it does that people are often strongly and successfully motivated to be to quite mistaken about what their concerns actually are. Recall from chapter 5 that Finlay's account of moral concepts as end-relational is only plausible if people are by and large radically ignorant of the nature of their own concepts.[372] We've seen excellent reasons to think people would in fact be motivated to be relevantly ignorant. Many of these same reasons, in addition to others, make it very difficult for us to know our own or others'

[372] He also defends his account of moral obligation as radically other-regarding in part based on the claim that people are motivated to deceive themselves into thinking that they largely fulfill their moral obligations, and so must regard them as *not* radically other-regarding.

values, and even more to know their relative priority rankings (even assuming a fact of the matter). Therefore, anyone who thinks that the sources of normativity are to be found in our actual values should want to conduct serious normative inquiry quite differently than we are prone to do now.

Specifically, we should want to attend to our actual concerns. If confronted with a question about what we *ought* to value, the very first thing we should look to do is figure out what we *do* value. Do we care—and how much and in what ways—about terrorism, racism, justice, military aggression, democracy, sexism, impartial concern for others, human rights, fairness, equality, what God wants, world peace, budget deficits, the ascendency of the proletariat, individual rights, moral principles of all kinds, torture, the suffering of strangers, the suffering of various animals, avoiding cruelty, and/or knowing the truth about these and other things? Or for many of these things do we care a lot about about caring, but not actually care very much? How much do we care about these things in comparison with, say, avoiding embarrassment, saving face, respecting authority, status of various kinds, ingroup loyalty(ies), strengthening the group as such, traditions as such, revenge, what other people think of us and/or having power of various kinds?

I'm not suggesting any answers to these questions, though there are some important studies suggesting that people's motivations, especially in the moral realm, are quite a bit different than they think they are.[373] I'm sure the answers differ for different people. Is Finlay right that people tend to care (a lot) more about their lawns or sports teams' success than they do about strangers dying of malnutrition? Do people who spend

[373] See Kurzban and Aktipis, 2007; Batson, 2011; Sabini et al., 2001.

money on phone sex but not on famine relief really care more about the former than the latter? It would be good to know.

Now I turn to address a normative subjectivist who seems very interested to internalize normative subjectivism, but doesn't see any significant conflict between that project on the one hand, and 'morality' on the other.

In order not to see any deep conflict between internalizing normative subjectivism and moral discourse, I think one would have to reject an important component of what I argued in Chapters 1 and 2. There I argued that moral discourse has a motivational (committing) function that is threatened by an awareness of its workings. This claim should be broken into two related theses. The first is that moral judgments, including the moral emotions involved in them, are best understood as commitment devices. I've argued that these commitment devices work at the level of culturally and biologically evolved mechanisms, and also at the level of individual psychology. The account is surely incomplete and likely incorrect in some of its details, but I hope the reader finds the central argument compelling. At any rate, let the general thesis that committing its users to ways of thinking, feeling and/or acting is central to moral discourse be called the Commitment-Device Thesis (hereafter 'COMMITMENT').

I've also argued that these commitment devices are generally (but not *necessarily*) threatened by an awareness of their workings. This is largely because an awareness of their workings directs attention to one's own (and often others') motivations and commitments, and this kind of attention can undermine the motivational efficacy of one's commitments. I've argued that features of the discourse that direct attention away from our own motives and commitments are profitably understood as serving the function of deflecting attention from these very motives and commitments. Again, I certainly have not provided a complete and flawless account of the role of the deflection of attention in moral discourse (or commitment generally). But let the general thesis that a core functional component of the commitment devices at work in moral discourse is the deflection of attention from our own motivations the Deflection of Attention Thesis (DEFLECTION).

I think to view employing moral discourse and internalizing normative subjectivism as compatible, you would prima facie have to reject DEFLECTION. I think I've said a lot to make DEFLECTION plausible, but it's important for my abolitionist argument to try to minimize skepticism about it. I would not attempt to make a case against moral discourse on the grounds that it centrally employs commitment devices as such. On the contrary, I doubt very much whether we could get on either as individuals or societies without employing them liberally. However, if our reasons are always relative to what we care about, but we are often quite ignorant of what we care about and how much we care about it (especially in the moral realm) due in large part to a discourse that systematically deflects attention from our own values, then that does present strong prima facie reasons to avoid such a discourse, especially for a normative subjectivist.

Jesse Prinz is an excellent representative of the viewpoint that internalizing normative subjectivism and moral discourse are compatible, though he is not directly responding to my concerns. His view is important for me to address, because unlike Finlay, he defends morality not only in terms of the truth of its claims, but also in terms of its *value*, which I have argued at length is the much more important question. Further, he at least roughly agrees with some of my central theoretical and practical commitments. On the theoretical side, he (implicitly) has a sort of commitment/motivational device view of morality. On the practical side, he thinks that we should 'constantly remind ourselves' of the subjectivist nature of our morals, that our sentiment-based values are independent sources of reasons, but also that we should train and modify them in light of our other values. That is very close to my view! It looks for all the world that he is just as interested in internalizing normative subjectivism as I am, but explicitly argues in favor of (keeping) 'morality'.

I want to show that that is a bad move by his own, and any similar subjectivist's lights. I think moral discourse is very poorly suited to the task of moral progress as Prinz conceives it, even if we take our current views of our values for granted. But it is even worse-suited if we realize that we don't typically know what our values are very well. Prinz, though he claims that his moral and nonmoral values are 'subject to revision', also displays no real interest in a sustained effort to know what our values are. I'll first describe his view, and in the course of doing so argue that an important aspect of it is plausible if and only if DEFLECTION is broadly correct. To the extent that his view is plausible then, that constitutes more evidence for DEFLECTION. Then I'll argue that by all the things he holds dear, he should be in the vanguard of the abolitionist movement.

6.2 Prinz's Sentimentalist Defense of Morality

Jesse Prinz (2007) explicitly disagrees with skeptics like John Mackie, Michael

Ruse and Joyce who claim that all moral claims are false. In fact, he thinks that moral truth is very easy to come by. According to Prinz's version of sentimentalism, moral properties are response-dependent properties; specifically, they depend on sentimental responses. On Prinz's view, if I say that something is morally wrong/right, I have spoken truly iff I have a sentiment (a disposition to feel an emotion) of dis/approbation toward it. That is the case because the standard concept (morally) WRONG, according to Prinz, is a 'detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the disapprobation range' (94).

This is as extreme a version of relativist subjectivism as one could hope to find. It holds that you speak truly in claiming that something is morally wrong just in case you have a disposition to feel emotions in the disapprobation range toward that thing. I cannot begin to recapitulate Prinz's arguments for it. As resourceful as he is in rebutting objections from moral objectivists, I am here more interested in how he fends off objections that are closer to home, namely those of metacognitive sentimentalists. For a metacognitive sentimentalist, employing the concept WRONG requires that one deem one's sentiment of disapprobation appropriate or merited. Call this the 'merit objection' to straightforward sentimentalism.

I'm going to briefly motivate the merit objection, then give Prinz's response, then discuss the relevance of those responses for my project. The relevance is of two main sorts. The first is that Prinz's response allows him—unlike the metacognitive sentimentalists—to distinguish first-order and second-order moral judgments. This is important not only because it contributes to making Prinz's view a more plausible version of sentimentalism than theirs, but because I want to show that this distinction is best captured not in terms of whether one has ‘really’ made a moral judgment one way or the other, but rather in terms of whether one cares about something vs. cares about caring about it. Focusing on this question rather than the question of the conditions for moral judgment is more tractable as well as more normatively significant.

The other reason his response is relevant to my project is that in order for Prinz to fully answer the merit objection, he rightly invokes (speculates as to) the central committing role that sentiments play in maintaining and stabilizing norms, and that something like DEFLECTION seems required to explain why it is that the committing role of the sentiments is so hard to see. Therefore Prinz’s account is only plausible if DEFLECTION is broadly correct, and if DEFLECTION is broadly correct, then by Prinz’s own lights we should abandon moral discourse, as I will try to show below.

Here is how the merit objection runs. Suppose a mob hit man is undergoing rehabilitation and now sincerely asserts that killing is wrong. However, it seems possible that he does not feel any sentiment of disapprobation toward killing, implying that sentiments aren’t necessary for moral judgment. Now imagine a person who sincerely insists that there is nothing morally wrong with homosexuality, though continues to feel sentiments of disapprobation when she contemplates it in herself or others. She is trying to rid herself of these sentiments, but has not yet. This example seems to imply that sentiments are also not sufficient for moral judgment (Prinz 2007, 112). Rather, for there to be a ful-blooded moral judgment, one must judge one’s sentiment of dis/approbation appropriate, merited, warranted, or some such thing.

Metacognitivism seems more attractive than Prinzean sentimentslism at first blush, but a daunting problem is that there are various senses of appropriateness. A joke could merit amusement because it is clever and witty, but for other, perhaps moral, reasons, it might not be appropriate to laugh or even to find it amusing. Standing one's ground when one knows one is in the wrong could be tactically appropriate but morally inappropriate (D'arms and Jacobson 2000; Prinz 2007, p. 113). Therefore a metacognitivist needs to find the right sense of appropriateness or fittingness. But this presents a very difficult dilemma. One cannot build in the idea that the sense of appropriateness is the moral sense without introducing a vicious circle, and one cannot propose a different sense on pain of failing to capture the judgment that something is *morally* wrong, as opposed to, say, tactically ineffective.

Of course showing that metacognitivism has problems doesn't show that Prinz has escaped the force of the apparent counterexamples involving the recovering hit man and homophobe. Prinz's strategy here is to say that if the hit man doesn't have any sentiments against killing, then he doesn't really think it's morally wrong (113), and that if the woman still has sentiments of disapprobation toward homosexuality then she does morally condemn it (114).

Prinz avoids fist-pounding here by providing us with an alternative description of the case. He says that the former hit man doesn't judge that killing is morally wrong, but he does judge that it is morally wrong not to have that attitude toward killing. The homophobe doesn't really judge that homosexuality is morally permissible but does judge that it is wrong to have those attitudes toward homosexuality. In each case, a sentiment is at work, but the sentiment is not directed toward certain kinds of actions but toward their own (lack of) certain kinds of sentiments toward those actions. So the homophobe does morally condemn homosexuality, but ‘condemns the fact that she condemns homosexuality’ (114).

Now for the second, related part of Prinz’s response to the merit objection. His response to the hit man and homophobe still leave us with a puzzle that the metacognitive sentimentalist seems much better positioned to answer than Prinz. And that is the undeniable fact that we do think that wrongness warrants or merits sentiments of disapprobation. But according to Prinz, moral properties are response-dependent. He analogizes them to colors in this respect. But we do not consider color responses *merited*, but only caused. They are then unlike emotional responses to that extent. So even if judging that emotional responses are fitting or appropriate is not built into moral concepts, it is still clearly something that we frequently do, and it seems an important part of moral thought and discourse, unlike thought and talk about color.[374] How can Prinz’s view explain this?

> To deem guilt morally appropriate is to have a moral sentiment toward guilt. It is to have a meta-sentiment …. Moral emotions are often implemented by a second layer of moral emotions … If our first-order norms are backed up by meta-norms, they will be harder to lose …
>
> [unlike color-responses,] Moral emotions are not merely caused; they are merited by their causes and they are regulated by meta-emotions. There is a sense in which we deem our moral emotions appropriate, but that is not essential to those emotions, or to our concept of morality. It is, instead, a powerful mechanism for sustaining our sentiments. (115)

This concludes my account of Prinz’s answer to the merit objection. I think it is on the right track, which is part of why I chose to discuss Prinz in the first place. I also

[374] This is not exactly the way Prinz characterizes what remains to be explained (115), but nothing of importance for my purposes here is lost or altered in my rendering.

think that what's right about it tells in favor of COMMITMENT, DEFLECTION, and the importance of attending to what we care about rather than what is morally WRONG.

Return to Prinz's contention that the former hit man and homophobe don't really think killing is wrong and that homosexuality is morally ok, respectively. It is easy to imagine a lot of fruitless disagreement ensuing at this point concerning the truth-conditions of when one 'really' does or does not morally condemn something. The (recovering) homophobe sincerely asserts that she doesn't think it is wrong, and yet has negative sentiments toward it. So, does she *really* judge it morally wrong or not? For Prinz's project, it is important that she does in fact morally condemn it iff she has sentiments of disapprobation towards it. But that should not be our concern. We should focus on what is normatively significant for each, namely, what each cares about.

The important thing that Prinz highlights here is that our sentiments can take different objects, including our (lack) of sentimental responses. What seems fairly straightforward in the case as described is that the hit man doesn't care (much) about not killing people. But he does care about caring about it. The relationship between caring about something and having sentiments (or motivational dispositions) toward it is much less controversial than that between making moral judgments and having sentiments or motivations. It is hard to see what arguing about whether a moral judgment is made in one case or another gains us, while it is easy to see how it might lead to pointless and distracting disagreements.

Once we put that distraction to the side, we can see more clearly what we have before us. The hitman cares about not being the kind of person who doesn't care (much) about killing people. This concern might comprise a disposition to feel guilty when perceiving his lack of guilt at the thought of killing for money. His metasentiment of guilt might over time generate first-order sentiments of guilt. What it is also likely to do is to make him think that he *does* care about killing innocent people (more than he really does). That kind of mistake is potentially very common, for reasons that we have discussed at some length. We are likely often *mistaken* about what we care about (and/or how much) due to our having metasentiments toward (not) caring about certain things, in particular ways, to lesser or greater extents.

As I pointed out in Chapter 4, if the perception of some action of ours is aversive, we are motivated to avoid the perception. That can be done by acting differently or perceiving our actions differently. Since our control over our sentiments is typically less reliable than our control over our actions, we should be even more likely to mis- or not perceive our sentiments than our actions. For a normative subjectivist, this kind of mis- or lack of perception is a large problem of great normative significance, as I argued (or pointed out) in the previous section. Placing our attention squarely on questions of what we care about rather than questions of what we judge morally wrong promises to mitigate this problem, if only by removing a (large!) distraction.

Now let's turn to Prinz's response to the claim that wrongness merits disapprobation. First let me point out that it is no part of Prinz's project to argue that moral sentiments or moral discourse generally has a committing function. And yet what he's done is explain what seems to be a central feature of moral discourse in terms of commitment. Of course I think he is on the right track, and I think the arguments I've given in favor of COMMITMENT and DEFLECTION give valuable support to Prinz's argumentative strategy here, which is based on almost pure speculation and intuitive plausibility. But I'm also keen to point out something Prinz doesn't. Prinz's account explains[375] why people regard sentiments as merited but not color responses. But it leaves something else crying out for explanation—namely why it is so hard to see or accept this fact! That is, on the assumption that Prinz is on the right track here, why have so many other very smart people who have spent many years working on this problem failed to regard this sense of emotions being merited as fundamentally in the business of stabilizing ways of feeling and acting? In slightly different terms, if COMMITMENT is correct, why is it so hard to see, and why do people actively resist it?

DEFLECTION predicts and explains this phenomenon. It also explains why many people would find speaker-relativism (and relativism generally) so 'implausible'.[376] Recall that for Prinz, an action is wrong just in case it has the property of disposing a person to respond to it with emotions in the disapprobation range. If DEFLECTION predicts anything, it predicts that people would generally be loathe to accept this, and the more so the more 'moralistic' they are. DEFLECTION predicts that 'objectivity' would be an important part of moral discourse. But a community of people who have accepted Prinz's thesis, for all Prinz tells us, could just as well tell one another that they have these or those sentiments toward these or those actions, without losing or changing anything important to moral discourse. DEFLECTION, on the other hand, predicts and explains why this would be a momentous departure from what we recognize as moral discourse.

It would be easy to misconstrue what I just said as a criticism of Prinz's view. On the contrary, for all I know or care, he is right about what makes moral claims true. My

[375] Much more convincingly once we've seen some arguments for COMMITMENT and DEFLECTION!
[376] Admittedly, many versions of relativism are implausible for other reasons.

point, like the point I made in response to Finlay's view about moral truth-conditions, as well as the point about Joyce's fictionalism (in which moral claims would be 'true-in-the-fiction'), is that every way of getting moral claims to come out true requires just the sort of radical, motivated lack of awareness that DEFLECTION predicts.

Prinz says, '[o]nce we recognize that morality exists to serve our wants and needs, we can try to adjust current morals so that they serve us better' (307).[377] But the vast majority of people don't think that morality is around to serve us, or is a social construction at all. Prinz tries to alleviate the worry people have that if morality is a social construction then it loses its distinctive value. He counsels these people to look to art and medicine and governments. These things 'come from us' but are none the less valuable for it. 'The fact that art is a social construction does not deprive it of value. We don't expect institutions of art to collapse upon discovering that art is a product of human invention' (8).

Precisely so. And yet many people expect and/or fear that *morality* might collapse upon some such (widely believed) discovery. Prinz suggests that such people must just be confused and their fears unfounded. DEFLECTION suggests rather that they perceive (if dimly and in other terms) something important about the peculiar motivational strategies employed by moral judgments and discourse.[378] The fact that people don't get upset about art being a social construction, but balk at the idea when applied to morality—and even more at the idea that it is a tool to serve us—is not

[377] I don't quite agree that morality *currently* exists (only or mostly) to serve our 'wants and needs'. I do agree that we should bend it in that direction, but to the extent that we are self-conscious about that, we will destroy what is peculiar to moral ways of thinking.

[378] Similar remarks apply to thinking of religion as a social construction. Religion as social construction is easy to accept if religion doesn't play a central role in committing you to ways of acting and being, not so easy otherwise.

evidence that people are confused or inconsistent, but rather evidence that morality is different than these other social constructions (if that's what they are), and different in ways that amount to further evidence for COMMITMENT and DEFLECTION. And if both that and Prinz's account of moral truth are correct, then moral truth is as easy to come by as it is irrelevant to deciding what to do and how to live.[379]

I just concluded that on Prinz's own view, moral truth is irrelevant. Now I'll contend that on his view of what is necessary to make moral progress, he should be energetically waving the abolitionist flag.

According to Prinz, for us to make moral progress requires 'norm-pitting', i.e., pitting norms against one another. It would be more accurate to describe his view as value-pitting, since some values are not associated with norms. Since there is 'no transcendental stance from which we can assess competing moral theories … we must retool our values from within' (289). I think this is precisely correct. I think that any normative subjectivist should agree. In order to do this, Prinz thinks we must bring nonmoral values to bear on our moral values. One moral system can be better than another if it fares better according to our nonmoral values. The nonmoral values Prinz considers are: the consistency of our moral values (with each other); their demandingness; their conduciveness to social stability; their contribution to 'our welfare' and (separately) to 'our subjective sense of well-being'; their generality, universality, and 'consistency with premoral biological norms' (291-2).[380]

[379] That is not to say that one's sentimental responses are irrelevant in deciding what to do, but rather to say that 'discovering' that such responses are what constitute 'moral truth' adds nothing useful to our deliberations. Anything that one might have once wanted from the venerable idea of moral truth has been quite thoroughly removed on this account.

[380] This list is not meant to be exhaustive, and subject to empirical revision.

Prinz notes that these nonmoral values can but need not compete with one another, and so it will often be possible for a system of moral rules to fare better along multiple dimensions. He also rightly points out that these standards are our own values; if they were not, we would not see them as advantages when comparing moral systems. Prinz insists that none of these standards are moral standards; in deploying them to assess moral systems we are 'stepping outside of morality' (292).

I see no reason to accept this last claim even on Prinz's own account, since it seems fairly clear that many people have sentiments connected with the values of promoting 'our' welfare and sense of well-being. But let that pass. The point I want to make is that Prinz is keen to insist that the moral rules/values are no more foundational or important than the nonmoral values. Every conviction—moral or nonmoral--is potentially relevant to evaluating any of our values (304). There is no special authority to moral values for Prinz; they are 'merely' the ones we have internalized via the relevant sentiments.

If that is the case, I think it is fairly clear that moral discourse is well-suited to hiding it from us. Indeed, Prinz (unlike Finlay) is quite clear that he thinks that moral progress faces a major obstacle due to our tendency to think that morality is 'immutable', or a 'window into absolute truth' (301). He warns that if we regard moral truths as objective, we lose sight of the possibility of moral progress.[381] One 'prophalytic' he recommends against this threat is to regard morality as instrumental to the nonmoral goals listed above. Yet he recognizes that we have a hard time viewing morality

[381] What Prinz must mean here is not moral progress in the sense of people or institutions being morally better, but moral progress in the sense of improving on our sense of what is (im)moral in the first place.

instrumentally, since ‘our sentiments present our basic moral values to us as if they were intrinsically good. [However,] If we remember that morality is a tool … we can be open to the possibility of moral growth’ (301).

Due to the tendency of our moral values to present themselves as intrinsically good, Prinz also recognizes a ‘deep vulnerability to self-deception about moral progress’ (302). Since our current moral standards by their nature present themselves as superior to the competing standards of the past, ‘when we reason about whether things have improved … we are subject to the confirmation bias: a tendency to cherry-pick evidence that will support our current convictions’ (302):

> Against this very serious concern, I can only recommend hard labor. We must *constantly remind ourselves* that our values are not reflections of an absolute truth … Morals are inculcated, and often shaped, as Nietzsche would say, by power struggles and happenstance. *Reminding ourselves of this deep contingency of morality* is just a first step. *We must also subject our values, including those we treasure, to rigorous reconsideration* in light of extramoral concerns … (302, my emphasis)

I humbly submit that a decicedly poor way to achieve a project like this is to keep referring to the sentiments one happens to have as moral truths and as one’s moral obligations. To do so radically misses the point of peculiarly moral discourse, which is in large part to *keep from ourselves* that our moral values are not objective or absolute, to *avoid awareness* of any deep contingency of our moral values, and to *prevent* a rigorous reconsideration of our moral values.[382]

[382] At any rate, that is essentially a restatement of DEFLECTION, which is itself made more plausible if Prinz’s view is correct, as I argued above.

The pointlessness—or radical counterproductiveness—of retaining the language of moral obligation is perhaps most striking when we see that on Prinz's view it is also the case that the sheer fact that someone *else* has a sentiment of disapprobation toward failing to do something means that *they* have a moral obligation to do it. Therefore if we find a white father who has a sentiment of disapprobation toward failing to kill his daughter's black lover, it follows straightforwardly that he has a moral obligation to kill him. Most people's reaction to such a claim is that this account of moral obligation must therefore be false. I think Prinz does a good job of rebutting such objections. But it is at the price of ridding the language of moral obligation of its primary function.

I've been arguing that Prinz has cheapened the language of moral obligation to the point that it loses any normative or motivational force that simply reporting the relevant sentiments doesn't have. But the problem is worse than that, from Prinz's own perspective. The language is not just pointless, but counterprodutive. For it is surely the case that *actually* having a moral obligation is *thought* to have more normative force than 'merely' possessing a moral sentiment. And I have explained that belief in normative force largely in terms of the motivational force accomplished by COMMITMENT and DEFLECTION.

So if Prinz *really wants us to recognize* that when we say that one value is morally better than another one, or when we say that we have a moral obligation, 'we simply reveal that we have internalized that value' (297), then continuing on with the language of moral truth and obligation seems a very poor idea indeed. Again, it would be analogous to claiming that it is easy to know God's commands, since some action being 'commanded by God' simply means that we have a sentiment of approbation toward that action. I am not claiming that one could argue as effectively for this claim as Prinz does for his claims. What I am claiming is that *even if* one could, and if one *also* thought it very important for people to recognize that their religio-moral values are social constructions ultimately in their own service, it would hardly be a good idea to go on with god-discourse. In other words, *if we want to promote normative subjectivism, we shouldn't use a discourse built for normative objectivism.*

At the end of his book, Prinz explicitly addresses the question whether we should 'forgo morality' in favor of exclusively nonmoral values. His answer is that we probably could not if we tried, but that it would be a bad idea in any case. However, what he means is that we should not get rid of moral sentiments and replace them with 'cool principles'. He does not address the question whether we should conduct ourselves in peculiarly moral discourse. Indeed, it is very hard to see why Prinz would object to directly discussing our sentiments and values, rather than thinking and talking in the language of moral obligations and so on.

The first reason Prinz gives for not forgoing morality is its motivational importance. But again, what he has in mind is ridding ourselves of certain sentiments. He argues that rules that are not sentimentally grounded are not effective guides to conduct. For example, 'the threat of guilt and shame help us resist temptations. Anger, contempt, and disgust help regulate the behavior of others' (306). But we don't need *moral discourse* to feel these emotions. Perhaps we need it to (stably) feel the peculiarly moralized versions of these emotions, but there are clearly nonmoral versions of these emotions. Here Prinz fails to recognize the extent to which the ways one conceives of the nature and role of one's emotions can change their character.

For example, Nietzsche was opposed to morality, but not contempt. He felt that contempt served to maintain a sense of elevation or rank. When he says, ‘One must be superior to mankind in force, in loftiness of soul – in contempt’,[383] it would be misleading at best to regard his contempt as a *moral* sentiment. There are two reasons for this. The first is that the motivation to maintain a sense of superiority, conceived in terms of loftiness of soul, is not what we consider a moral motivation. The second is that conceiving of one’s emotions as in the service of one’s ‘higher goal’ (as Nietzsche put it), or as in the service of other nonmoral values (as Prinz advises us to do) changes their character. When Prinz says that regarding moral values as instrumental is difficult because our sentiments present our values to us as if they were intrinsically good, he ignores the contribution of moral concepts to (the stability of) that very presentation.

Since Prinz is focusing on the value of the punitive moral emotions, we should notice something important about our concept of moral responsibility. When one looks at the conditions under which we hold people morally responsible, they seem strikingly appropriate as conditions under which it would make sense to condemn and/or punish people on broadly consequentialist grounds. For example, we hold that one cannot *really* be morally responsible unless they were in control of their actions. Likewise, it would seem *pointless* to condemn or punish people who were not in control of their actions. But our judgments of moral responsibility—and the sentiments associated with them—are very resistant to such ‘forward-looking’ considerations. Choosing to ‘hold responsible’ for broadly pragmatic reasons is famously quite a different affair from the standard conception of judging someone ‘actually’ morally responsible; the (moralized)

[383] *The Antichrist*, Preface.

sentiments commonly associated with the latter do not flourish in the captivity of the former.[384] If we think of our (moralized) anger, contempt and disgust as in the business of regulating the behavior of others, then we are already well on the road to their being nonmoralized and perhaps (much) less effective.

At any rate, the sentiments that Prinz rightly finds motivational don't require peculiarly moral discourse or concepts, and neither does their regulation. For what it's worth, I offer as evidence the fact that I haven't made a serious claim as to moral obligation or responsibility for many years now. Yet I can assure readers that my anger, disgust and contempt are in good order. If anything, they flourish too well. They are just not moralized (or less stably than in the past).

Moving beyond moral discourse does not consist then in the application of 'cool principles', or in the project to 'eliminate anger, contempt, disgust, guilt, and shame' (307). Rather, it consists primarily in a commitment to know ourselves (our motivations and values) better than we do, and in part to facilitate just the kinds of progress that Prinz wants to make—progress that on precisely his own view, he should think that moral discourse greatly inhibits. By ignoring the extent to which that discourse contributes to (the stability of) our experience that our sentiments reflect 'absolute truth', Prinz misses an opportunity to advance moral progress as he conceives it. For a good way to 'constantly remind ourselves' of the sentimental nature of our 'moral' values is to stop dressing those values in the glad rags of moral duties and obligations.

[384] My point here is not to make any large claims about the nature of moral responsibility, only to point out that conceiving of the sentiments associated with holding people morally responsibile as worthwhile due to their ability to regulate others' behavior changes their character.

6.3 Taking Charge of the Spectacle

> "It is from the will to truth's becoming conscious of itself that from now on—there is no doubt about it—morality will gradually *perish*: that great spectacle in a hundred acts that is reserved for Europe's next two centuries, the most terrible, most questionable, and perhaps also most hopeful of all spectacles…"[385]

Now I turn to address the final objection to straight talk. The objection proceeds roughly as follows. 'There has been a lot of moral progress it seems, and so whatever we've been doing is working pretty well. We are gradually living lives better and better—by our own lights—than those we'd have lived in the past. At any rate, the case against conservativism isn't strong enough. We probably don't know ourselves as well as we might, but doing so in the ways you propose is very dangerous—as you've done much to convince us.'

As preface to my response, I want to repeat something I said in Chapter 4 but have not emphasized. Philosophers (and many others) ubiquitously argue for what 'we' should do—even for the most sweeping and radical proposals—where 'we' is implicitly or explicitly meant to cover everyone. I would no more set myself the task of arguing that everyone ought to try to move beyond moral discourse than I would do so for religious discourse. To my mind, it would be absurd not to recognize that very many, perhaps the vast majority, simply could not conduct their lives effectively in anything like the ways I recommend. It would be mad to suppose that every religious person is best advised to drop their belief in God and accompanying religious discourse and instead to start attending to their own concerns as the sources of normativity. Neither would I make

[385] Nietzsche; *On the Genealogy of Morality*, Book III, section 28.

any such claim with respect to moral discourse. When I say ‘we’, I mean to be addressing *primarily* (but not only) those people who accept—at least in broad outline—normative subjectivism, COMMITMENT and DEFLECTION.

But even among these people, there might be those who feel that I have made my arguments for DEFLECTION too well, and that we had better not be like ‘those Egyptian youths who endanger temples by night … and want by all means to unveil, uncover, and put into bright light whatever is kept concealed for good reasons’.[386] Joyce seems to be in this camp. My response is to first repeat that I think that for my audience, it is probably too late for peculiarly moral discourse—too late not to become more aware of our motivations in the socio-moral realm, for the kinds of reasons I gave in Chapters 4 and 5. Second, I acknowledge that it is dangerous, but have tried to illustrate the many ways that *not* doing so is also dangerous. I view the crisis facing a society with increasing numbers of normative subjectivists as analogous to that of a society with increasing numbers of atheists. The unavailability of religious discourse as a means of guiding and justifying ways of acting and being did and does represent a crisis for the relevant individuals and society as a whole. The unavailability of moral discourse also poses a great crisis, but also an analog to the tremendous opportunity we had and took in moving beyond religious discourse. How terrible or how hopeful the coming ‘spectacle’ will be depends very much on how we approach it. I welcome it as an exciting if (and because) dangerous adventure. I’ll end with a last attempt to inculcate a similar sense of optimism in the skeptic.

[386] Nietzsche, *Gay Science*; Preface, Section 4.

Let’s briefly review the motivational *costs* of moralizing, by which I mean the tendency for moralized thinking to undermine one’s own consciously-held values. It does this primarily by motivating misperception of one’s actions and motives. First, recall that the motivational benefits of moral emotions[387] are a double-edged sword. While they motivate one to avoid behavior one perceives as wrong, they often do this by means of motivating one to perceive what one does as not wrong, and often positively righteous. So moral emotions motivate misperception of motivations and actions. This is ‘akratic’ in the sense of violating one’s values, though the misperception avoids the awareness of the violation. Second, principled thinking is also double-edged, and in a similar way. By bundling actions into classes, more significance is accorded to individual instances. This strategy supports greater motivation than using case by case evaluation, but also leads to a greater sense of weakness—and therefore lessened willpower in the future—if one perceives oneself as having lapsed. For just this reason, there are powerful motivations to misperceive one’s actions as not having been violations after all. Again, one’s values or commitments might be violated without one ever having the sense of having violated them or having been ‘weak’.

As I noted in chapter 4, the way to avoid the akratic costs of moralizing is by achieving an increase in the relevant kinds of self-awareness, but greater self-awareness threatens moralizing generally, especially once one is aware of COMMITMENT and DEFLECTION. I also tried to show that protecting oneself from the akratic costs of moralizing without improving self-awareness increases one’s vulnerability to the

[387] Moral emotions are not unique in this regard, but they are exemplary.

downsides of willpower by supporting the dominance of ‘compulsion-range’ interests, threatens pathological levels of inflexibility, and makes for saddening confusion and unnecessarily hostile divisiveness.

The above costs of moralizing are ‘akratic’ in the sense of generating powerful motivations to avoid perceptions of having done wrong *by one’s own lights*. But that is not the only, or the most serious problem with the motivated deflection of attention from one’s motivations. A discourse that systematically directs attention from our introspected concerns runs the risk of committing us in ways that are severely pathological from the perspective of what we most deeply value. We saw Huck’s abiding commitment to return Jim to his owners. Fortunately, the value of his friendship with Jim overcame this commitment. Unfortunately, Huck saw this as moral weakness. By his own lights, he did wrong, though he was very lucky to have done so. Examples in which people were not so lucky can be multiplied indefinitely, with no need to rely on fiction. As Hinckfuss pointed out, very many atrocities have been carried out in the name of a people’s moral values.

In some of these cases, such atrocities might not have violated the people’s deeper values. But in many cases, I am sure they did. We saw that Joyce’s advice on how to avoid this enormous liability is to moralize the ‘already useful’ values. But the value of our values doesn’t only depend on their instrumental ‘usefulness’ to us, but also something like the depth and/or extent to which we care about them. And a discourse that systematically deflects attention from the things we care about has the predictable result that we don’t know very well what those things are. Ameliorating this problem is one of the primary motivations for abandoning moral discourse, and attending instead to the things we care about. Recall Prinz's point that the nature of moral sentiments is such that our perceptions of moral progress are suspect. Our sentimental values (perhaps with a large boost from our moral concepts) represent themselves as superior to real or imagined alternatives. Claims of moral progress should not be made or accepted glibly.

Still, the objection might continue, it appears that we have a standoff. There are many potential dangers and benefits on both sides. But on the side of moral discourse—and religious discourse for that matter—is that they are ubiquitous and have *demonstrated* payoffs (if also demonstrated costs). What I am recommending no large stable society has ever done. There is, according to my own DEFLECTION thesis, a good explanation for that.

Just this sort of argument has been and continues to be made for the long-term necessity of religious belief (and discourse). I suspect I am more sympathetic to this argument than most atheists are. Of course for many of us, it it too late for religious belief. But even if it were not too late, there are good reasons to think that neither moral nor religious discourse are as important as they once were.

When moral and religious discourse were developing, we knew little or nothing about the nature and indispensable value of commitments. Until quite recently, it seemed paradoxical to many people that it could be rational to restrict one's options. To some, it still is. As we saw in Chapter 1, the ubiquity of people setting out to restrict their own behavior has been a puzzle to utility theorists. Even to the people engaging in such behavior, it might strike them as a mark of their own weakness or even insanity to place restrictions on their own freedom. With the increasing appreciation of the value of commitment, we can rationally hope and expect our society to develop various methods for committing themselves in ways that don't require deep and pervasive ignorance about the nature of those very commitments.

The tactics that people employ to secure their values against undermining and erosion are overwhelmingly discovered by trial and error and often—perhaps overwhelmingly—without conscious awareness. But the same conscious awareness that in some contexts can threaten commitments can also be directed at developing better commitment strategies. One of the best resources we have are other people. Telling people who care about us that we want their help upholding our commitments is a great way to take advantage of the fact that they are not prone to many of the akratic rationalizations we are in our own case. Far too often, when people perceive what they regard as moral transgressions, they react moralistically, motivating defensive reactions in the targets of their condemnations. These reactions include the motivated misperception of one's own actions or motives.

On the other hand, if we take a cooperative rather than condemnatory stance, we might help a person to see that their actions are not in accord with their own values. Despite the fact that awareness of the committing function of morality can undermine its effectiveness, awareness of the importance of commitment and the ubiquity of self-deception, mis- and lack of perception of one's own motives hold the promise of making up for what is lost.

However, even if we cannot make up for all of the motivational power we lose by moving beyond moral and religious discourse, when we recall that these discourses also have great power to commit us to *pathological* ways of acting and being, they look much less attractive. Any argument that holds that the committing power of moral discourse is worth keeping must confront the fact that that very power requires a deflection of attention from our motives and values, and that this in turn precludes an effective investigation into which of our values are *worth* having.

I presented the considerations above not only as reasons to exchange moral discourse for straight talk, but as reasons for doing so in the right way and with the right attitude. For as I have said, I think it is too late for a normative subjectivist to reap much in the way of stable benefits of moral discourse. My hope is that recognizing the costs of moralizing and the benefits of straight talk will make us glad that it's too late. Thinking that it's too late removes it as an apparently available option and helps commit us to a better alternative. But the benefits of straight talk are very much bound up with our attitude about engaging in it.

An analogy from martial arts might be helpul. In the grappling art of Brazilian jiu-jitsu, it sometimes happens that an opponent is succeeding in his effort to put you in a position that you don't want to be in. A beginner continues to fight against being taken there beyond the point at which it is clear that the effort won't succeed. Even as he is being swept onto his back, for example, his body and mind are still fighting it.

A more advanced grappler can recognize that he is being taken somewhere he did not want to go, but rather than fight what he cannot stop, he goes there on his own terms. Sometimes one can only hope that this results in the position not being as bad as it otherwise would have been. At other times, going *in the manner of one's choosing* somewhere one did not originally want to go can greatly improve one's position. One of the great strengths of jiu-jitsu is its repertoire of attacks that can be initiated lying on one’s back. The repertoire is so powerful that many experts feel most comfortable fighting on their backs with an opponent on top of them—a position that without the relevant expertise and mindset would be a decidedly unattractive place to be. But with the right experience and mindset, one can turn a crisis into an opportunity.

That is roughly how I propose to view the move from moral discourse to straight talk. First, I think it is too late for moral discourse to fulfill its peculiar role; by that I mean that peculiarly moral discourse will prove increasingly unable to meet the demands of an increasingly knowledgeable, reflective and inquisitive culture. That presents a crisis that may be turned into an opportunity, much as the inability of religious discourse to do normative duty for millions of people gave rise to a crisis, as well as an opportunity to move beyond it. There was—and still is—no guarantee that post-religious ethics will work out. And yet I think that post-religious moral discourse has proven (radically) superior to religious discourse as a means of conducting our normative investigations. Likewise, there is no guarantee that post-moral discourse will work out. But there is a great *opportunity* to discover, evaluate and promote our values much better than we do now.

Second, having the right experience and mindset are important. Much of what is lost in moral discourse must be made up for and surpassed by means of practice in another way of thinking, but not only that. Recognizing the central importance of being able to secure our values against the varieties of akrasia, we will want to gain practice in strengthening ourselves in the relevant ways. Willpower and attention are importantly connected. Practices such as yoga and meditation increase discipline and our attentional powers. These are of course just two examples to stand for indefinitely many. The point is that if the motivational power of moral discourse relies on a deflection of attention from our introspected motives, then we will want to be able to harness that power without the pathology. The ability to direct and maintain one's attention on one's commitments and away from one's impinging desires is something that requires practice. The crisis of awareness generates the requirement and opportunity to manage that awareness in the service of one's higher goals.

Nevertheless, though I stand behind all I've said, I am an advocate who does not resent that name.[388] There are dangers of the kind of self-awareness I am promoting, and there are dangers of avoiding it. Both can be described independently of the intrinsic concern one has for the relevant kind of self-awareness, both in oneself and in others. I think it is clear that for some people, knowing what motivates them and what they really care about has high intrinsic interest; for others, it has much less. For all I know, the question whether to go to the trouble of trying to find out the answers to questions like these by changing the way one thinks and talks in the moral realm will depend largely on how much we care intrinsically about knowing ourselves in this way.

[388] "What provokes one to look at all philosophers half suspiciously, half mockingly, is … that they are not honest enough in their work, although they all make a lot of virtuous noise when the problem of truthfulness is touched even remotely ... They are all advocates who resent that name…" (Nietzsche; *Beyond Good and Evil*, section 5)

Concluding Remarks

At the most general level of description, practical normative discourse should consist in 1) finding out what our values are and 2) finding out how to best promote those values. When values conflict, the first question to ask is which one we care about more. Some things we care about intrinsically, like our own welfare, that of our loved ones and many other things, including perhaps the welfare of strangers and that people be punished for acting contrary to (some of) our values. There will be many things we do or should care about derivatively, like monetary policy. Most people don't care intrinsically about monetary policy, but they should care about it because it impacts things they do care about intrinsically. Other things will be mixed, like foreign policy. We might care intrinsically about how we treat other peoples, and also derivitavely about how that treatment affects us in return.

In order to answer very many questions of what we ought to value, we will want a lot of information and the ability to think clearly. What makes idealizing accounts of practical rationality so attractive is that they highlight the value of knowledge and clear thinking. It's a mistake, however, to suppose that one must launder one's preferences through omniscient and vivid appreciation of all nonnormative facts for those preferences to have normativity. But it is not a mistake to think that good reasoning, knowledge of relevant facts and/or vivid appreciation of those facts is often crucial to promoting many of our most important values.

These things *might* even be enough, in principle, to settle the question of what to do when values conflict. But I doubt it, especially in what are some of the most important and difficult cases. My skepticism about this is the same kind of skepticism I have about idealizing accounts of rationality. Vividly appreciating the attractions of one kind of life likely make the attractions of competing ones inaccessible. In such cases one cannot answer the question which kind of life appeals more to a person by having them appreciate fully what each kind of life will be like.[389]

Sidgwick thought that there was a duality of practical reason; that between the demands of morality and those of prudence, neither had priority over the other. I broadly agreed with Finlay that this is a false dichotomy and that we have many other values that might be as or more important to us than these, and that what we have most reason to do depends on what we most (strongly, deeply, and/or stably) care about. There are probably some and perhaps very many cases in which we cannot answer the question which of two competing values we do or can hold more strongly, deeply and/or stably.

What to do about competing values of this sort is not something I addressed here in any detail.[390] However, I want to warn against two mistakes, one substantive and the other both substantive and interpretive. The substantive mistake I have already warned against, but it is so apparently seductive that I repeat it here. The mistake is that higher orders of desire (value) have intrinsically greater normative significance. My usage of Frankfurt might have given that impression. There it seemed that the reason the person is making a mistake by caring about not stepping on the cracks in the sidewalk is that it is not important to him to make this important to himself. The examples in Prinz were also

[389] See David Sobel, "Full Information Accounts of Well-Being" Ethics 104 (1994), pp. 784-810 and Connie Rosati, "Persons, Perspectives, and Full Information Accounts of the Good" Ethics 105 (1995), pp. 296-325.

[390] I strongly suspect it is not something that can be addressed by traditional philosophical methods.

of this sort. In the cases of the recovering hit man and homophobe, we saw the higher order desire as the more normatively significant. Likewise, the example of multiple orders of value from Chapter 3 gave the impression that the higher-order desire has the greater normative significance. There the person wished that she didn't wish that she didn't wish to eat the cake. That is because she wished she were less vain, which seems a more normatively significant value than valuing one's appearance (too much).

It is not coincidental that these authors represent the higher orders of desire as more significant. I explained this in section 3.4.3 as intimately connected with the workings of the conscious will. Higher orders of desire are associated with our *conscious, statable* evaluative beliefs and our conscious experience of (longer-term) willing and reasoning. We identify ourselves with these perspectives, for they seem to us the only genuinely evaluative perspective; however, there is very much evaluative cognition going on that we don't recognize as such.

This is what is so compelling about the Huck Finn case (and others like it); it blows up this identification. Though Huck might identify himself with the motivated perspective associated with his conscious evaluative beliefs and experience of willing, we needn't do so. And once we see that we needn't do so in his case, we can learn we needn't do so in our own case either. So even if higher-order desires often have greater normative significance, *it is not by sheer virtue of being a higher-order desire or commitment*. Our higher-order commitments are typically if not always those we consciously endorse. But what we consciously endorse is hugely influenced by deeply contingent norms and often outright propaganda.

The second potential mistake is both substantive and interpretive. On my view, there simply is *no transcendent account* of what makes something a (stronger) normative reason. That is, there is no criterion of the strength of practical reasons that transcends all motivated perspectives. So it is an interpretive mistake to read me as offering one[391] and I think it is a substantive mistake to think there is one (and therefore a practical mistake to look for one). Most if not all neo-Humean theories founder on this attempt to give a transcendent account of the strength of reasons. We saw that Hume foundered badly on it.

But it might look like I have been saying that the fact that we care more about something is what makes it the case that we have stronger reason to promote it. Offering as a criterion of one's strongest reasons that which one cares about most can be interpreted as an account that transcendends all motivated perspectives. Such an account is even naturally interpreted as transcendent, just as Hume's criterion is. Which desire one cares the most about, just as which desire is strongest, is—at least on the most natural reading—a fact that transcends any particular motivated perspective. Perhaps Finlay and/or Frankfurt think of the strength of reasons in this way, but I don't. I am not offering any such transcendent fact as a criterion of our strongest reasons.

My judgment that our strongest reasons are relative to what we care most (strongly, deeply and/or stably) about is itself made from a motivated perspective. It is, roughly speaking, the motivated perspective that asks what is worth caring about. I think the answer to the question what our strongest reasons are is tantamount to the question what we should care most about. And as I have argued, I think that the answer to what

[391] Though I realize that one could reasonably get that impression; hence this explicit denial.

we should care most about is (roughly) a function of what we do care about most strongly, deeply and/or stably, including to what extent we will be *able* to care about it on reflection.

Now I think what makes that true is to be found not in motivation-transcendent facts relating strength of reasons to strength of caring, but in the fact that in asking ourselves the question what to care about, we activate the motivated perspective (or cluster thereof) that we call caring about what to care about, and that perspective seems to take as its inputs *something like* information about how strongly, deeply and stably we are committed to various sorts of concerns.[392] I think the answer to the question of what constitutes our strongest reasons is the empirical question of what sorts of information we seek (again, not necessarily consciously or under that description) when we ask what is worth caring about. In my view, the answer to this empirical question is that we seek information about how deeply, strongly and/or stably we do care and are able to care about something. My answer is admittedly vague and rough; I stand ready to be corrected, and certainly improved upon. But I nevertheless think my answer is important and on the right track—and that answer clearly does not transcend our actual motivations.

For example, if I ask if it is worth caring about our children very much, it seems that the fact that we do care—and care about caring—very strongly, deeply and stably comprises the vast bulk of the information needed to answer the question (the fact that we care about caring can be a big part of the stability). On the other hand, if we ask ourselves if we should care much about maximizing the total amount of pleasure in the universe, the fact that we don't actually care about this (or anything to which such

[392] Not that it does so consciously, or under this description!

concern conduces), or if we do, that this kind of concern is shallow, unstable, and weak (and almost certainly based on a confusion) also provides us with most if not all of what we need to know to answer that question as well.

In the final pages of his *Reasons and Persons*, Parfit expresses sentiments for 'Non-Religious Ethics' very similar to those I want to convey in favor of Non-Moral Ethics. Parfit explains that the widespread belief in a god or gods 'prevented the free development of moral reasoning' (454). Since open disbelief in God by 'a majority' of people is so recent an event (indeed, not yet the case!), Non-Religious Ethics is very young. Since it is so young, it is too soon to tell how it will develop. However, precisely because we cannot know how it will develop, Parfit ends his book with the claim, 'it is not irrational to have high hopes' (454).

I feel much the same way about Non-Moral Ethics, specifically what I'm calling straight talk. Much as religious ethics prevented the free development of moral reasoning, peculiarly moral reasoning prevents the free inquiry into what we actually care about, at least directly and under that description. I think the primary reason that non-religious ethics is better than religious ethics is that very much ethics does investigate what we care about, including and especially Parfit's work. As I argued in Chapter 3, we should see McMahon and Parfit's investigations into what is objectively required for rational prudence as unwitting investigations into the nature of the motivated perspective of our 'egoistic concern'. I remarked there that the intuitions they employ to make their arguments are unmysterious if conceived of as tapping into an actual source of concern, while they are quite mysterious if conceived of as (somehow) tapping into motivation-independent facts about rational requirements.

However, by doing this indirectly, and especially under a misconception of what one is about, one runs the risk of very serious confusion and pathology, especially in the moral realm. In supposing that one is on the hunt for attitude-independent moral principles, interpreting one's attitudes ('intuitions') as (somehow) data for or against them, Parfit runs into just this sort of confusion.[393] As a result, he spends a considerable amount of time in his book attempting to avoid 'The Repugnant Conclusion'. The Repugnant Conclusion is that if there are enough people in a world B whose lives are just barely worth living, then that world would be better than some other world A in which the people lived lives vastly superior to the lives in B. So long as there are enough people in B, then the total value of their lives could exceed the total value of the lives lived in A, and then that world would be better, and so we would be morally obliged to promote it.

Parfit is led to try so hard to find a theory that would avoid this conclusion because another principle that implies it seems to him hard to reject, namely a version of the Principle of Beneficence. The details are not important. What is important is the idea that we might be, if we cannot find some suitable theory that can avoid this and other problems, 'forced to accept' the Repugnant Conclusion. Parfit is right to find the conclusion repugnant, but wrong to think we could ever rationally be forced to accept it. By fundamentally misconstruing the nature of moral principles, he thinks we might be

[393] In what is nevertheless an excellent work of philosophy. I choose to discuss Parfit not because his example is particularly egregious, but because it is paradigmatic of what I think is a deep confusion in much moral philosophy.

forced to accept this conclusion much in the way that a person who fundamentally misconstrues the nature of personal rules might be forced to accept the conclusion that he should write in his journal rather than rescue his child from drowning in the pool outside because it's very nearly 11 p.m., he has yet to write today, and he 'must write an hour a day'.

Both Parfit and this person would do well to be more aware of the nature and function of principles and rules. They would do well to not worry about ever having to accept this or any Repugnant Conclusion due to their being entailed by moral principles. We find the Repugnant Conclusion 'hard to accept' in Parfit's words for the simple reason that we don't give a damn about promoting a world like that or anything that doing so would conduce to, and we care quite a bit about not doing so. And that is all we need to know to know that we shouldn't do it. If we are ever 'forced' to accept a repugnant conclusion as the outcome of straight talk, it will be a conclusion that we find repugnant from one motivated perspective, but attractive (or less repugnant) from another—and hopefully one that we care more (strongly, deeply and/or stably) about. Sophie's Conclusion (Choice) was a repugnant one to have to make, but in the circumstances the choice itself was not. Accepting Parfit's Repugnant Conclusion could never be the result of sound practical reasoning, but only a deep—but all too common—misconception of the nature and value of moral principles.

We want the discipline of moral discourse without the pathology. We want to know what we care about and to be able to act accordingly. That is a difficult task. But there is reason for optimism. Parfit finds reason for optimism in Non-Religious Ethics because so few people have made it their life's work. He points out that 'Non-Religious Ethics has been systematically studied, by many people, only since about 1960' (452). In that very year, Thomas Schelling pioneered the modern study of commitment. Several people have studied commitment systematically since then, but *very* few people have made Non-Moral Ethics (straight talk), grounded in an understanding of the relationships between commitments and values, their life's work.

While there is reason for optimism, I agree both that it is dangerous and that we cannot know how it will develop. Like Parfit, for just this reason I think it is not irrational to have high hopes. However, Nietzsche better captures the attitude about moving beyond moral discourse that he and I both want to promote:

> At last the horizon appears free to us again, even granted that it is not bright; at last our ships may venture out again, venture out to face any danger; all the daring of the lover of knowledge is permitted again; the sea, *our* sea, lies open again; perhaps there has never yet been such an 'open sea'[394]

[394] *Gay Science,* section 343. Parfit prefaces his book with just this quotation, I assume, because it captures the sense of freedom, adventure and danger that he associates with moving from religious to non-religious ethics. I empathize.

References

Ainslie, G. (1992). *Picoeconomics*. The Strategic Interaction of Successive Motivational States Within the Person. Cambridge University Press.

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press.

Ainslie, G., & Haslam, N. (2003). Altruism is a primary impulse, not a discipline. *Behavioral and Brain Sciences*, *25*(02), 251. Cambridge University Press.

Ainslie, G. (2005). Precis of Breakdown of Will, *Behavioral and Brain Sciences 28,* 635-673.

Allman, J. M. (2000). *Evolving Brains*. W. H. Freeman.

Anderson, S. W., A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex, *Nature Neuroscience*, 2 (11), 1032-7.

Axelrod, R. (1986). An Evolutionary Approach to Norms. *The American Political Science Review*, 1095–1111.

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

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325-44.

Bargh, J . A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition*, Vol. 1, 2nd ed., 1 - 40. Hillsdale, NJ: Erlbaum.

Bateson, P. P. 1966. “The characteristics and context of imprinting.” *Biological Reviews* 41: 177–220.

Barr, A. (2001). Social Dilemmas and Shame-based Sanctions: Experimental results from rural Zimbabwe. *The Centre for the Study of African Economies Working Paper Series*, Working Paper 149.

Batson, C. D. (2011). What's Wrong with Morality? *Emotion Review*, *3*(3), 230–236.

Baumeister, R. F., & Heatherton, T. F. (1996). Self-Regulation Failure: An Overview. *Psychological Inquiry*, *7*(1), 1–15. Taylor & Francis, Ltd.

Becker, G., & Murphy, K. (2008). A Theory of Rational Addiction*, Journal of Political Economy,* 96 (4), 675-700.

Berger, P. L., & Luckmann, T. (1967). *The Social Construction of Reality*. New York: Doubleday.

Berns, G. S., McClure, S. M., Pagnoni, G. & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience* 21:2793–98.

Blair, R. J. R. (1995). “A cognitive developmental approach to morality: Investigating the psychopath.” *Cognition*: 57: 1–29.

Blair, R. J. R., Jones, L., Clark, F., and Smith, M. (1997). “The psychopathic individual: A lack of responsiveness to distress cues?” *Psychophysiology* 34: 192–198.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*(1), 1–45. American Psychological Association.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*(1), 17–28.

Boyd, R., & Richerson, P. (2002). Group Beneficial Norms Can Spread Rapidly in a Structured Population. *Journal of Theoretical Biology*, *215*(3), 287–296.

Brandt, R. B. (1979). *A theory of the good and the right*. Oxford University Press.

Bratman, M. (1987). Intention and Evaluation. *Midwest Studies in Philosophy*, *10*(1), 185–189.

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

Brink, D. (1997). Moral Motivation. *Ethics*, *108*(1), 4–32.

Brink, D. (2007). The significance of desire. In *Oxford Studies in Metaethics* (Vol. 3, pp. 5–45). Oxford: Oxford University Press.

Brink, D. O. (1989). *Moral Realism and the Foundations of Ethics*. New York: Cambridge University Press.

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, *425*(6955), 297–299.

Brosnan, S. F., & Waal, F. B. M. (2002). A proximate perspective on reciprocal altruism. *Human Nature*, *13*(1), 129–152.

Carpenter, J. P., Matthews, P. H., & Ong'ong'a, O. (2004). Why Punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, *14*(4), 407–429.

Caspi, A., Moffitt, T. E., Silva, P. A., Stouthamer-loeber, M., Krueger, R. F., & Schmutte, P. S. (1994). Are Some People Crime-prone? Replications Of The Personality-crime Relationship Across Countries, Genders, Races, And Methods *Criminology*, *32*(2), 163–196.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge University Press

Cialdini, R. B. (1991). Altruism or Egoism? That Is (Still) the Question. *Psychological Inquiry*, *2*(2), 124–126.

Clore, G., Schwarz, N., & Conway, M. (1994). Affective Causes and Consequences of Social Information Processing. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of Social Cognition: Basic processes* (pp. 323–404). Psychology Press.

Cordell, J., & McKean, M. (1992). *Making the Commons Work: Theory, Practice, and Policy*. ICS Press.

D'Arms, J., & Jacobson, D. (1994). Expressivism, Morality, and the Emotions. *Ethics*, *104*(4), 739–763. The University of Chicago Press.

D'Arms, J., & Jacobson, D. (2000). Sentiment and Value. *Ethics*, *110*(4), 722–748. The University of Chicago Press.

Damasio, A., Tranel, D., & Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioural Brain Research*, *41*(2), 81–94.

Damásio, A. R. (1994). *Descartes' error*. emotion, reason, and the human brain. Avon Books.

D'Andrade, R. (1984). Cultural meaning systems. In R. A. Shweder & R. A. Levine (Eds.), *Culture theory: Essays on mind, self, and emotions* (pp. 88–119). Cambridge, England: Cambridge University Press.

Dancy, J. (2000). *Practical reality*. Oxford University Press, USA.

Darwall, S. (1988). Perspectives on Self-Deception. In B. P. McLaughlin & A. O. Rotrty

(Eds.), *Perspectives on self-deception*. University of California Press.

Darwall, S., & Gibbard, A. (1992). Toward Fin de Siecle Ethics: Some Trends. *The Philosophical Review*, *1*(1), 115–189.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.

Davidson, D. (2001). *Essays on Actions and Events*. Oxford University Press.

Davis, J. L., & Rusbult, C. E. (2001). Attitude alignment in close relationships. *Journal of Personality and Social Psychology*, *81*(1), 65–84. American Psychological Association.

Del Raine, R. (1993). USP Marion's Version of Orwell's 1984 and Beyond. *Journal of Prisoners on Prisons 4*(2), 1-11.

Deluty, M. Z., Whitehouse, W. G., Mellitz, M. & Hineline, P. N. (1983) Self-control and commitment involving aversive events. *Behavior Analysis Letters* 3:213–19.

Dennett, D. C. (1995). *Darwin's Dangerous Idea*. evolution and the meanings of life. Simon and Schuster.

Douglass, F. (1963). *Narrative of the Life of Frederick Douglass, an American Slave*. Forgotten Books.

Dreier, Jamie (2005). Moral Relativism and Moral Nihilism, in *Handbook of Ethical Theory*, ed. David Copp, 240-64, New York: Oxford University Press.

Dworkin, R. (1996). Objectivity and Truth: You'd Better Believe it. *Philosophy & Public Affairs, 25*(2), 87–139.

Elster, J. (1984). *Ulysses and the Sirens*. studies in rationality and irrationality. Cambridge University Press

Elster, J. (1987). *The Multiple Self.* Cambridge University Press

Elster, J. (1989). *The Cement of Society*. A study of Social Order.

Elster, J. (2007). *Explaining Social Behavior*. More nuts and bolts for the social sciences. Cambridge University Press.

Emerson, Ralph Waldo. 1841. *Self-Reliance.* Reprinted in The Complete Essays and Other Writings of Ralph Waldo Emerson. Ed. B. Atkinson, 1940, 145–69. New York, NY: American Library

Enzle, M., Hansen, R., & Lowe, C. (1975). Humanizing the Mixed-Motive Paradigm. *Simulation & Gaming*, *6*, 151-65.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature*.

Fessler, D., & Haley, K. (2003). The Srategy of Affect: Emotions in Human Cooperation. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*. Dahlem Workshop Reports.

Finlay, S. (2007a). Too much morality. In P. Bloomfield (Ed.), *Morality & self-interest*. New York: Oxford University Press.

Finlay, S. (2008). Against all reason? Skepticism about the instrumental norm. In *Hume on Motivation and Virtue* (pp. 155–178). Basingstoke: Palgrave Macmillan.

Finlay, S. (2011). Errors Upon Errors: A Reply to Joyce. *Australasian Journal of Philosophy*, *89*(3), 535–547.

Foot, P. (1972). Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, *81*(3), 305–316.

Frank, R. H. (1988). *Passions within reason*. the strategic role of the emotions. W W Norton & Co Inc.

Frankfurt, H. (1988). The Importance of What We Care About. In *The Importance of What We Care About* (pp. 80–94). New York: Cambridge University Press.

Garrard, E., & McNaughton, D. (1998). Mapping Moral Motivation. *Ethical Theory and Moral Practice*, *1*(1), 45–59.

Gazzaniga, M. S. (1985). *The social brain*. discovering the networks of the mind. New York: Basic Books.

Gazzaniga, M. S., Bogen, J. E., & Sperry, R. W. (1962). Some Functional Effects of Sectioning the Cerebral Commissures in Man. *Proceedings of the National Academy of Sciences of the United States of America*, *48*(10), 1765. National Academy of Sciences.

Gibbard, A. (1990). *Wise choices, apt feelings*. a theory of normative judgment. Harvard University Press

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *J. theor. Biol. 206*(2), 169-179.

Gintis, H. (2005). *Moral Sentiments and Material Interests*. The Foundations of Cooperation in Economic Life. The MIT Press.

Glover, J. (1985). *Matters of Life and Death*. New York Review.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge UniversityPress

Greene, J. D. 2002. The Terrible, Horrible, No Good, Very Bad Truth about Morality and What to Do About It. Doctoral dissertation, Princeton University.

Greene, J. D., and Haidt, J. 2002. "How (and where) does moral judgment work?" *Trends in Cognitive Sciences* 6: 517–523.

Greene J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment" *Science*, 293 (5537), 2105-8.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Haidt, J. (2001). The Emotional Dog and its Rational Tail. *Psychological Review*, A Social Intuitionist Approach to Moral Judgment, *108*(4), 814–834.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. Unpublished Manuscript, University of Virginia

Haidt, J., Rozin, P., McCauley, C., & Imada, S. (1997). Body, Psyche, and Culture: The Relationship between Disgust and Morality. *Psychology & Developing Societies*, *9*(1), 107–131.

Hardin, C. L. (1988). *Color for philosophers*. unweaving the rainbow. Hackett Publishing Company.

Henrich, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767–1770.

Henrich, Joe, & Boyd, R. (1998). The Evolution of Conformist Transmission and the Emergence of Between-Group Differences. *Evolution and Human Behavior*, *19*(4), 215–241.

Henrich, Joseph, Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(06), 795–815. Cambridge University Press.

Herrnstein, R. J. and Prelec, D. (1992) Melioration. In G. Loewenstein and J. Elster (eds.), *Choice Over Time.* New York: Sage, pp. 235–264.

Heuer, U. (2004). Reasons for Actions and Desires. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *121*(1), 43–63. Springer.

Heyman, G. M. (1996) Resolving the contradictions of addiction. *Behavioral and Brain Sciences* 19:561–610.

Hinckfuss, I. (1987). "The Moral Society: Its Structure and Effects," *Discussion Papers in Environmental Philosophy* Canberra: Australian National University

Hirschi, T., & Hindelang, M. (1977). Intelligence and Deliquency: A Revisionist Review. *American Sociological Review*, *42*(4), 571–587.

Hirshleifer, J. (1999). There are many evolutionary pathways to cooperation. *Journal of Bioeconomics*, *1*(1), 73–93.

Ho, M. Y., Al-Zahrani, S. S. A., Al-Ruwaitea, A. S. A., Bradshaw, C. M., & Szabadi, E. (1998). 5-Hydroxytryptamine and impulse control: prospects for a behavioural analysis. *Journal of Psychopharmacology*, *12*(1), 68–78.

Hollerman, J. R., Tremblay, L. & Schultz, W. (1998) Influence of reward expectation on behavior-related neuronal activity in primate striatum. *Journal of Neurophysiology* 80:947–63.

Hornstein, H. A., LaKind, E., Frankel, G., & Manne, S. (1975). Effects of knowledge about remote social events on prosocial behavior, social conception, and mood. *Journal of Personality and Social Psychology*, *32*(6), 1038–1046. American Psychological Association.

Hume, D. 1740. *A Treatise of Human Nature*. Clarendon, 1978.

Hussain, N. J. Z. (2004). The Return of Moral Fictionalism. *Philosophical Perspectives*, *18*(1), 149–188.

Huxley, A. (1998). *Brave New World.* New York: Harper Collins.

Jowett, B. (1937). *Dialogues of Plato*. Translated Into English, With Analyses and Introduction. Cambridge UniversityPress.

Joyce, R. (2001). *The Myth of Morality*. Cambridge Studies in Philosophy

Joyce, R. (2006). *The Evolution of Morality*. The MIT Press.

Joyce, R. (2007). Morality, schmorality. In P. Bloomfield (Ed.), *Morality and self-interest*. Oxford University Press.

Joyce, R. (2011). The Error In "The Error In The Error Theory." *Australasian Journal of Philosophy*, *89*(3), 519–534.

Kalderon, M. E. (2005). *Moral Fictionalism*, Oxford University Press.

Kant, I. (1793/1960) *Religion within the limits of reason alone*, trans. T. Green & H. Hucken, pp. 15–49. Harper and Row.

Karoly, P. (1993). Mechanisms of self-regulation: A systems view. *Annual review of psychology*.

Kern, F. (1948). *Kingship and Law in the Middle Ages* (S.B. Chrimes, trans.). Westport, CT: Greenwood Press.

Knutson, B. (2004). Sweet Revenge? *Science, New Series*, *305*(5688), 1246–1247. American Association for the Advancement of Science.

Kohlberg, L. (1963). The development of children's orientations toward a moral order: I. Sequence in the development of moral thought. *Vita Humana*.

Kohlberg, L. (1969). Stage and Sequence: The Cognitive-Developmental Approach to Socialization. In D. A. Goslin (Ed.), *Handbook of Socialization Theory and Research* (pp. 347–480). Chicago: Rand McNally.

Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151–235). New York: Academic Press.

Korsgaard, C. (1997). Normativity and Instrumental Reason. In G. Cullity & B. Gaut (Eds.), *Ethics and Practical Reason*. Oxford: Oxford University Press.

Korsgaard, C. M. (1986). Skepticism about Practical Reason. *The Journal of Philosophy*, *83*(1), 5–25.

Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge UniversityPress

Kuhn, D. (1991). *The Skills of Argument*. Cambridge, England: Cambridge University

Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. American Psychological Association.

Kunst-Wilson, W., & Zajonc, R. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558.

Kurzban, R., & Athena Aktipis, C. (2007). Modularity and the Social Mind: Are Psychologists Too Self-ish? *Personality and Social Psychology Review*, *11*(2), 131–149.

Kurzban, R., & DeScioli, P. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, (28), 75–84.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, *112*(2), 443–477. The MIT Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. the embodied mind and its challenge to Western thought. Basic Books (AZ).

Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society Supplementary Volume*, 63: 113–37.

Lieberman, D., Tooby, J., and Cosmides, L. 2003. “Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest.” *Proceedings of the Royal Society: Biological Sciences* 270: 819–826.

Locke, J. (1965). *Treatise on Civil Government and a Letter Concerning Toleration*, Charles L. Sherman, ed. New York: Appleton-Century-Crofts (Originally published 1689).

Loewenstein, G. (1992). *Choice over time* . Russell Sage Foundation Publications.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational behavior and human decision processes*. *65*(3), 272-292.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267–286.

Lorenz, K. (1937). The Companion in the Bird's World. *The Auk*, *54*(3), 245–273.

Mackie, G. (1996). Ending Footbinding and Infibulation: A Convention Account. *American Sociological Review*, *61*(6), 999–1017.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Mahapatra, M. (1981). *Traditional structure and change in an Orissan temple*. Calcutta, India: Punthi Pustak.

Mazur, J. E. (1997). Choice, delay, probability, and conditioned reinforcement. *Animal Learning & Behavior*, *25*(2), 131–147.

McDowell, J. (1978). Are Moral Requirements Hypothetical Imperatives? In *Proceedings of the Aristotelian Society* (Vol. 52, pp. 13–29 & 31–42).

McDowell, J. (1987). Projection and Truth in Ethics, Lindsay Lecture, University of Kansas. Reprinted in S. Darwall, A. Gibbard, and P. Railton (eds.), *Moral Discourse and Practice: Some Philosophical Approaches*, pp. 215–27. New York: Oxford University Press.

McDowell, J. H. (1996). *Mind and world*. Harvard University Press

McMahan, J. (2002). *The ethics of killing*. problems at the margins of life. Oxford University Press.

McNaughton, D. (1988). *Moral Vision: An Introduction to Ethics*. Oxford: Blackwell.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. American Psychological Association.

Mischel, W., Cantor, N., & Feldman, S. (1996). *Principles of self-regulation: The nature of willpower and self-control.* Guilford Press.

Moll, J., de Oliveira-Souza, R., and Eslinger, P. J. (2003). “Morals and the human brain: A working model.” *NeuroReport* 14: 299–305.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., and Pessoa, L. (2002). “The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic moral emotions.” *Journal of Neuroscience* 22: 2730–2736.

Nagel, T. (1970). *The Possibility of Altruism*. Clarendon Press, Connecticut.

Nelson, E. E., & Panksepp, J. (1998). Brain Substrates of Infant–Mother Attachment: Contributions of Opioids, Oxytocin, and Norepinephrine. *Neuroscience & Biobehavioral Reviews*, *22*(3), 437–452.

Nesse, R. M. (2002). The Evolution of Subjective Commitment. In R. M. Nesse (Ed.), *Evolution and the Capacity for Commitment*. Russell Sage Foundation.

Newcomb, T. M. (1943). *Personality and social change; attitude formation in a student community.* New York: Dryden Press.

Newman, J. (1981). The fictionalist analysis of some moral concepts. *Metaphilosophy*. *12*(1), 47-56.

Nichols, S. (2004). *Sentimental rules*. on the natural foundations of moral judgment. Oxford University Press, USA.

Nichols, S., & Folds-Bennett, T. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition*, (90), b23–b32.

Nietzsche, F. (1998 [1887]), *On the Genealogy of Morality*, trans. Maudemarie Clark and Alan Swenson. Indianapolis, Indiana: Hackett Publishing Company.

Nietzsche, F. (1888/1920). *The Antichrist*, trans. H. L.Mencken. New York, NY: Knopf.

Nietzsche, F. (1887/1974). *The Gay Science*, trans. Walter Kaufmann. New York: Random House

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

Nolan, D., Restall, G., & West, C. (2005). Moral fictionalism versus the rest. *Australasian Journal of Philosophy*, *83*(3), 307–330.

Nucci, L. P. 1986. "Children's conceptions of morality, societal convention, and religious prescription." In *Moral Dilemmas*, ed. C. Harding. Precedent.

Nucci, L. P. (2001). *Education in the Moral Domain*. Cambridge University Press.

Olson, J. (2010). In defence of moral error theory. In M. Brady (Ed.), *New Waves in Metaethics*. Bakingstoke: Palgrave Macmillan.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review*, *86*(2), 404–417.

Palmer, R. C. (1993) *English Law in the Age of the Black Death, 1348–1381.* Chapel Hill: University of North Carolina Press.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Piaget, J. (1965). *The moral judgment of the child*. (M. Gabain, Trans.). New York: The Free Press.

Polivy, J. (1998). The Effects of Behavioral Inhibition: Integrating Internal Cues, Cognition, Behavior, and Affect. *Psychological Inquiry*, *9*(3), 181–204. Taylor & Francis, Ltd.

Price, M. (2005). Punitive sentiment among the Shuar and in industrialized societies: cross-cultural similarities. *Evolution and Human Behavior*. *26*, 279-87.

Prinz, J. J. (2007). *The Emotional Construction of Morals*. Oxford University Press.

Putnam, H. (1971). *Philosophy of Logic*. New York: Harper and Row.

Pyszczynski, T. & Greenberg, J. (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in Experimental Social Psychology* 20:297– 340

Quinn, W. (1993). *Morality and Action*. (P. Foot, Ed.). Cambridge University Press.

Rachlin, H. (1995) Self-control: Beyond commitment. *Behavioral and Brain Sciences* 18:109–59.

Railton, P. (1986). Moral Realism. *The Philosophical Review*, *95*(2), 163–207.

Railton, P. (2006). Humean Theory of Practical Rationality. In *The Oxford Handbook of Ethical Theory* (pp. 265–282). New York: Oxford University Press.

Reginster, B. (2008). *The Affirmation of Life*. Nietzsche on Overcoming Nihilism. Harvard University Press

Richerson, P. J., & Boyd, R. (2005). *Not By Genes Alone*. How Culture Transformed Human Evolution. University Of Chicago Press.

Ricoeur, P. (1971) Guilt, ethics, and religion. In J. Meta (ed.), *Moral Evil Under Challenge,* New York: Herder and Herder.

Rosati, C. (1995). Persons, Perspectives, and Full Information Accounts of the Good. *Ethics*, *105*(2), 296–325.

Rozin, P., Haidt, J., & McCauley, C. (1993). Disgust. In M. Lewis & J. Haviland (Eds.), *Handbook of Emotions* (pp. 575–594). New York: Guilford Press.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586.

Russell, B., & Pigden, C. R. (1999). *Russell on ethics*. selections from the writings of Bertrand Russell. Psychology Press.

Sabini, J., Siepmann, M., & Stein, J. (2001). The Really Fundamental Attribution Error in Social Psychological Research. *Psychological Inquiry*, *12*(1), 1–15. Lawrence Erlbaum Associates, Inc.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. 2003. “The neural basis of economic decision making in the Ultimatum Game.” *Science* 300: 1755–1757.

Scanlon, T. (1998). *What we owe to each other*. Belknap Press.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology (Columbia University)*.

Shepher, J. 1971. “Mate selection among second generation kibbutz adolescents and adults: Incest avoidance and negative imprinting.” *Archives of Sexual Behavior* 1: 293–307.

Shizgal, P. & Conover, K. (1996) On the neural computation of utility. *Current Directions in Psychological Science* 5:37–43.

Shore, B. (1996). *Culture in Mind*. Cognition, Culture, and the Problem of Meaning. Oxford University Press, USA.

Shweder, R., Much, N., & Mahapatra, M. (1997). The "Big Three" of Morality (Autonomy, Community and Divinity) and the “Big Three” Explanations of Suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.

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

Singer, P. (2009). *The Life You Can Save*. Acting Now to End World Poverty. Random House Digital, Inc.

Smilansky, S. (2000). *Free will and illusion*. Oxford University Press, USA.

Smith, M. A. (1995). *The moral problem*. Wiley-Blackwell.

Sobel, D. (1994). Full Information Accounts of Well-Being. *Ethics*, *104*(4), 784–810.

Strathearn, L., Fonagy, P., Amico, J., & Montague, P. R. (2009). Adult Attachment Predicts Maternal Brain and Oxytocin Response to Infant Cues. *Neuropsychopharmacology*, *34*(13), 2655–2666. Nature Publishing Group.

Stevenson, C. L., (1937). "The Emotive Meaning of Ethical Terms," *Mind* 46, 14–31.

Sully, J. (1884) *Outlines of psychology*. Appleton.

Thomson, J. (1971). *In Defense of Abortion*. *Philosophy & Public Affairs*, *1*(1), 47-66.

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *Quarterly review of biology*. *46*(1), 35-57.

Twain, M. (2001). *The Adventures of Huckleberry Finn*. W.W. Norton and Co.

Van Inwagen, P. (1983). *An essay on free will*. Oxford University Press, USA.

Wagenaar, W. A. (1988). *Paradoxes of gambling behaviour*. Psychology Press.

Wallace, R. (1990). How to argue about practical reason. *Mind*, 99, 355-385.

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, *16*(10), 780–784.

Wiggins, D. (1987). *Needs, values, truth*. essays in the philosophy of value. Blackwell.

Williams, B. (1981). *Moral Luck*. Philosophical Papers, 1973-1980. Cambridge University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy*. Fontana Press.

Wundt, W. (1969). *Outlines of Psychology*. (C.H. Judd, Trans.). St. Clair Shores, MI: Scholarly Press. (Original work published in 1897)

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*(20), 7398–7401.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of*

Personality and Social Psychology, *51*(1), 110–116.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, *39*(2), 117–123. American Psychological Association.

Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. *Nebraska Symposium on Motivation*. University of Nebraska Press.

Zimbardo, P. G., LaBerge, S., & Butler, L. D. (1993). Psychophysiological consequences of unexplained arousal: A posthypnotic suggestion paradigm. *Journal of Abnormal Psychology*, *102*(3), 466–473.