Moral Mental States

Four Methods in Metaethics

Christopher Alen Sula

A dissertation submitted to the Graduate Faculty in Philosophy

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy,

The City University of New York

2011

This manuscript has been read and accepted for the
Graduate Faculty in Philosophy in satisfaction of the
Dissertation requirement for the degree of Doctor of Philosophy.

David M. Rosenthal

_____ _____

Date Chair of Examining Committee

Iakovos Vasiliou

_____ _____

Date Executive Officer

Catherine Wilson (advisor)

Stefan Bernard Baumrin

Samir Chopra

Douglas Lackey

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Moral Mental States: Four Methods in Metaethics

Christopher Alen Sula

Advisor: Catherine Wilson

Metaethics has traditionally focused on the meaning of moral statements, the referents of moral terms, and the justification of moral claims. This dissertation focuses on the *thoughts* in virtue of which our moral statements have their meaning and our moral claims have their content. These "moral mental states," as I call them, are best understood as having both intentional and volitional aspects and, corresponding to each aspect, as having satisfaction conditions involving both correctness and success in action. Put simply, our moral mental states are not only responsible for changing the world; they are responsible for changing the world in better ways, rather than worse ones. Using this framework, I examine three current metaethical frameworks (moral realism, constructivism, and expressivism/quasi-realism) and assess their adequacy in accounting for both kinds of these satisfaction conditions. I argue that none of the approaches succeeds in meeting basic constraints imposed by a robust view of moral mental states and propose my own view (interactionism). On my account, we often change our moral stances once the outcomes of the plans they recommend becomes known. Taking this as its basis, interactionism holds that a moral mental state is correct or incorrect

based on the content (roughly, approval or disapproval) of the moral mental state one would hold in light of this improved information. Accordingly, the procedures for confirming moral content are similar to those for confirming scientific content: hypotheses and observations play a central role, and we must occasionally defer to other inquirers on issues that exceed our capacity for investigation. I conclude by presenting three empirical hypotheses that claim that we do, as a matter of course, have other-regarding interests and that, in a significant number of cases, our self-regarding and other-regarding interests overlap.

Acknowledgements

Like all participants in a discussion, I owe a great deal to those who have talked with me and especially those who taught me how to talk.

Chief thanks are owed to my advisor, Catherine Wilson. Her comprehensive approach, gentle commentary, and fine example have all made this dissertation possible. Her abilities to engage a new problem, structure a careful response, and articulate that response clearly and elegantly—always with an eye to the historical and scientific—are unmatched.

Stefan Baumrin first directed my attention toward issues in value theory. Before his tutelage, I was an uncomplicated Nietzschean; ever since, I have been appropriately puzzled about the moral life. We have talked frequently, agreed rarely, but have ever been friends and allies.

David Rosenthal has been a great source of advice, rigor, and friendship for many years. His very carefully structured courses and writings instilled in me a deep conviction in naturalism and the central place of intentionality in philosophy. He has been one of my greatest teachers, and it is no overstatement to say that David changed my philosophical life.

I also wish to thank my examiners, Samir Chopra and Doug Lackey, for their time and generosity in conducting the defense. Samir, in particular, has proved a vibrant and refreshing thinking partner over the past few months. I regret that we did not meet sooner, but I look forward to many fruitful discussions to come.

I am especially thankful for colleagues who helped me work through the ideas in this dissertation. First, there were members of my dissertation workshop—James Dow, Ellen Fridland, and David Pereplyotchik—who remained patient as I first developed the ideas presented here. Jen Corns and Myrto Myopolous were helpful interlocutors over several years, and Amanda Favia and Kyle Ferguson provided invaluable comments on drafts at various stages.

There are also colleagues from other fields without whose personal support I could not have finished this project. The list is long, but memorable: Anick Boyd, Gregory Donovan, Allyson Foster, Ian Foster, Eero Laine, Kim Libman, Leigh McCallen, Kate McPherson, Claudia Pisano, Jared Simard, Suzanne Tamang, and especially Jill Belli and Shawn Rice.

I am also grateful to my students at Lehman College, with whom I tested many of the ideas and arguments in this dissertation.

Most importantly, I owe an enormous professional and personal debt to my friend and colleague David Morrow. Through him, I was first introduced to metaethics and began to develop my own approach to the issues examined here. His keen insights and careful commentary have proved simultaneously challenging and rewarding, and I cannot imagine a finer partner throughout my years in graduate school.

Finally, I owe incalculable thanks to my dear friends Sean Noyce, Katya Usvitsky, Kristin Miller, Weishun Cheung, Caitlin Healy, Shawn, Kate, Amanda, and, most of all, Abbi Leman.

Contents

Chapter 1

Situating Morals in a World of Minds

> [L]anguage is meaningful because it is the expression of thoughts—of thoughts which are about something.
>
> —Roderick Chisholm, "Chisholm-Sellars Correspondence on Intentionality"

When I say, "Murder is wrong," what kind of remark am I making? Am I making a statement of fact about the world, as I am with other observations such as "It's raining," or "The Earth is round?" Am I, instead, voicing an attitude and thereby, perhaps, attempting to change the world in certain ways by goading or guiding others to abstain from murder and prevent it whenever possible? Or am I doing both—that is, am I making some statement of fact and also performing some action, though it might be unclear where the statement of fact ends and my own preferences, desires, and vision for the world begin?

These questions about the nature of moral language belong to the field of metaethics, which is not concerned with the rightness or wrongness of actions *per se*, but rather with the content of moral claims, their truth conditions, and other phenomena of moral judgment, including moral psychology. Most prominently,

metaethics has focused on the meaning of ethical terms.[1] C. L. Stevenson perhaps put this best in saying: "These [ethical] questions are difficult partly because we don't quite know what we are seeking. We are asking, "is there a needle in the haystack?" without even knowing just what a needle is. So the first thing to do is to examine the questions themselves. We must try to make them clearer, either by defining the terms in which they are expressed or by any other method that is available" (1963, 10). Since Stevenson's time, the pursuits of metaethics have extended far beyond philosophy of language, into philosophy of mind, epistemology, logic, and even the empirical sciences of psychology, sociology, economics, and biology. There is even a suggestion, at present, that metaethics extends naturally into discussions of metanormativity, which has wider application throughout other areas of philosophy.[2]

In considering the nature of moral statements, we consider, in the broadest measure, the nature of the connection between the mind and the world—those features of the world that give occasion for moral thought and reflection and the mental–linguistic vehicles by which we plan and accomplish changes within the world. This connection runs in both directions, as it were. A speaker may be liable for making his/her moral assertions match the world just in case there is some fact of the matter at stake (e.g., whether torture really is wrong). This matching, in turn,

---

[1] Welchman (1989) credits G.E. Moore with this turn, not by directly focusing on the meaning of terms but by setting the agenda for later emotivists who took those considerations to be an interesting and important task.

[2] Gibbard (2006) suggests that the concept of mental content is normative, as well as meaning. "What's crucially puzzling in metaethics," he says, "extends to the puzzling topic of normative concepts in general. That makes for a broader subject matter, the metatheory of normativity, covering concepts of justified belief and justified feeling" (321). For other examples, see Field (2000) and Chrisman (2008).

involves bringing his/her thoughts, beliefs, attitudes, etc. in line with reality—that is, making his/her mind like the world. Call this direction of fit "mind to world." On the other hand, when a speaker says "Torture is wrong" in a world where torture is being practiced, he/she may be responsible for reducing the amount of torture that occurs, arguing against it, persuading others of its immorality, and behaving in other ways that attempt to bring the world in line with his/her thoughts—that is, making the world like his/her mind. Call this direction of fit "world to mind."

These two directions of fit and the interplay between them lie at the heart of metaethical issues. We want to know what, if anything, is true about the content of our moral thought and talk, and we want to know what, if anything, we are responsible for bringing about in the world (in virtue of our capacity to think and talk in certain ways). While ordinary language or formal analysis of moral statements may make some inroads to addressing metaethical concerns,[3] we are ultimately drawn back to the *thoughts* in virtue of which those statements have their meaning. As Roderick Chisholm put it, "language is meaningful because it is the expression of thoughts—of thoughts which are about something" (1957, 529). Metaethics, then, considers chiefly what we might call "moral mental states" akin to other kinds of mental states such as everyday perceptions, observations, thoughts, wishes, intentions, and so on.[4] In this respect, what is called for in metaethics is a

---

[3] Moore was perhaps the biggest proponent of this view, believing that armchair reflections on the term 'good' would reveal metaphysical truths about the property of goodness, but it was no less common to his emotivist rival A. J. Ayer, who discussed the nature of ethical propositions and challenged the notion that they were truth-conditional. Even today, figures like Frank Jackson and Allan Gibbard make liberal use of conceptual–linguistic analysis in their accounts.

[4] By 'mental state', I mean to capture the broad class of mental phenomena that includes sensations, qualitative states, intentional states, and volitional states. I leave open the question whether the term can

thorough analysis of moral mental states and their role in generating linguistic behavior and non-linguistic behavior. This very general approach is helpful in at least two respects.

First, it is of intrinsic interest for metaethicists to understand the nature of moral mental states and how those states interact with language and extra-mental reality. In fact, what seems at issue between moral realists and constructivists is the proper way to analyze the content of moral mental states. Whereas realists hold that these states are about something in the external world apart from human experience, constructivists believe they are about a distinctive kind of experience, that of occupying the moral point of view. Expressivists, by contrast, believe that these states are not primarily aimed at getting something right about the world, but are instead aimed at expressing attitudes and influencing others. The failure of a century's worth of work to settle basic metaethical questions suggests that the problem may lie at a deeper theoretical level, and it is worth examining the ways in which competing theories approach the issue of moral mental states.

The second reason why these states should interest metaethicists is that one's overall view of mental content must hang together, in a certain sense. One's view of mental content generally may determine one's views of moral mental content and everything that follows from it. For example, a committed direct reference theorist who believes that each thought directly represents and is

---

apply more broadly to diachronic mental processes, which might be of interest to virtue theorists interested in overall character traits or habits rather than particular moral thoughts, judgments, beliefs, and so on. For further discussion of moral mental states with reference to mental states in general, see §§2.1.3, 2.2, and 2.3.

connected to an object in the real world will almost certainly, for reasons of consistency, be a realist who believes that moral judgments refer directly to some stance-independent facts and properties.[5] For these reasons of theory coherence, it is important to address the issues of mental states that underlie metaethical theories. It may even turn out that metaethical considerations could help us adjudicate overall questions of mental content; that is, moral mental states might prove to be a sort of test case for mental content, and conclusions reached in metaethics might help tip the scales in favor of some theories of mental content more generally. I think this is rather unlikely because moral mental states appear to be a relatively small part of our overall mental economy, but that point itself remains an open question.[6]

In this dissertation, I will argue that the existing metaethical approaches have failed to address the full range of problems posed by moral mental states. They have, at some times, ignored important aspects of these states and drawn improper metaphysical conclusions from them; at others, they have ignored the importance of these states altogether, failing to give a robust explanation of their genesis and consequences. These difficulties may not be insurmountable. It is entirely possible that these approaches could be revised to address or avoid them, or that developments in other areas of philosophy, such as reference or

---

[5] In fact, I think many philosophers of mind and language have wound up being error theorists because they believe moral terms *must* refer in this way, and, doubting that any such referents exist, they conclude that moral statements must be systematically false or, worse yet, meaningless. Though error theory and moral realism disagree sharply on their views of mental reality, they both share the methodological view that questions about the correctness of moral beliefs, judgments, etc. must be determined by querying the existence of moral facts, properties, etc. in the world.

[6] Heidegger, Habermas, and Levinas, to my mind, all offer accounts on which normative or ethical reflection underlies much more of our mental/conceptual economy.

epistemology, will obviate the problems I am about to sketch. Still, these accounts must come to grips with the fact that their implied views of the mind–world relation lack the explanatory power they need, frustrate the aims of their own theories, or are flatly inconsistent with larger philosophical views that garner much allegiance.

To begin, it will help to formulate a working understanding of moral mental states in terms of this broader mind–world link. Following that, in Chapter 2, I consider, in detail, the nature of moral mental states and establish several criteria by which metaethical theories may be assessed and critiqued. In the next three chapters, I examine ways in which existing approaches fail to satisfy these criteria: in Chapter 3, I argue against moral realism; in Chapter 4, I examine the prospects for constructivism; and in Chapter 5, I consider a family of theories I call "performativist" views, which includes emotivism, expressivism, and quasi-realism. Following these critical remarks, I sketch my own metaethical approach in Chapter 6 and highlight the ways in which it incorporates the best elements of these other accounts but also distances itself from their associated problems.

In the remainder of this chapter, I want to develop an intuitive notion of moral mental states for use in metaethical theory. I begin by surveying the various kinds of moral mental states one can hold and by eliminating certain states that *seem* to be moral mental states (i.e., moral doubts and wonderings) but are, on further analysis, quite disanalogous from the standard examples. I also offer some brief remarks on the nature of mental attitudes and moral content, with further discussion reserved for the following chapter.

1.1. Moral Mental States

Consider Gilbert Harman's (1977) infamous example of seeing children douse a cat with gasoline and light it aflame. From this simple observation, a number of moral mental states may spring: I may have a *thought* that the children shouldn't be acting in such a way. I may *judge* that their action is wrong, or *wonder* whether they are doing the right thing. Feelings of anger, outrage, horror, or resentment may follow in turn. I may *intend* to help the cat or try to prevent further harm from occurring. Very likely, I will express these thoughts and feelings in some way—even if only by means of a gasp or a shout. At times, I might even come to *doubt* my own moral beliefs about the situation and whether the action is indeed morally wrong and my own beliefs, morally correct. These examples illustrate the diversity of moral mental states one can have and their corresponding verbal and non-verbal behaviors. In discussing this variety, it is important to keep in mind two distinct aspects of mental states: attitude and content. I will discuss each of these in turn before addressing further issues raised by moral mental states.

1.1.1. Mental Attitude

This diversity of moral mental states—or rather, mental attitudes[7]—is similar to the range of attitudes one may adopt toward non-moral mental content. I can, for example, think, believe, judge, wonder, doubt, hope, or wish that it is raining at the moment or that the table in front of me is red. In all of these cases, the content of

---

[7] Rosenthal (2005) employs the term 'mental attitude' in roughly the same way as Frege uses 'sense' to differentiate the "cognitive significance" or "mode of presentation" of the intentional content in question.

the mental state remains fixed on a particular situation or concept, yet the stance one adopts toward that content varies. This attitudinal difference is usually demarcated by a difference in the linguistic expression of those mental states. J. L. Austin discusses this in terms of illocutionary force:

> To perform a locutionary act is in general, we may say, also and *eo ipso* to perform an *illocutionary* act…[such as]
>
> asking or answering a question,
>
> giving some information or an assurance or a warning,
>
> announcing a verdict or an intention,
>
> pronouncing a sentence,
>
> making an appointment or an appeal or a criticism,
>
> making an identification or giving a description,
>
> and the numerous like. ([1962] 1974, 98–99)

In each of these cases, the wording of the locutionary act and its illocutionary force reflect a difference in the attitude of the mental state expressed in that locutionary act. If I *doubt* whether it's raining, I may ask, "Is it raining?", whereas if I *believe* that it's raining, I may assert, "It's raining." Similarly in the case of moral expressions, it is (roughly) clear from the locutionary act whether I am condemning someone for an action, asserting a moral belief, or asking another for a moral opinion, and the

illocutionary force of each of these corresponds to my mental attitude of intention, belief, wonder, and so on.[8]

Attitudinal differences are also particularly important in respect of the verbal and non-verbal behaviors they incline people to perform. If I, for example, *believe* that it's raining, I am much more likely to express, "It's raining," than I would be if I merely *wondered* whether it was raining. Similarly, I would be more inclined to take an umbrella with me if I *believed* that it was raining than I would be if I *doubted* that it was raining. The conventions for these behaviors are not *so* regular that we can always tell what a person is thinking from what he/she is saying or doing—the person could, after all, be silent or lying or acting in some way counter to how he/she would otherwise act while in such a mental state.

Despite this *general* problem of ambiguity in the conventions governing mental states and their corresponding non-verbal behaviors, it is quite clear that, in the case of *moral* mental states, there is a strong, perhaps necessary connection between those states and non-verbal behaviors. As Stevenson put it, "'goodness' must have, so to speak, a magnetism. A person who recognizes *X* to be 'good' must *ipso facto* acquire a stronger tendency to act in its favour than he otherwise would have had" (1937, 16). So if I have a belief that torture is wrong, it's simply not enough for me to hold that belief and stand by while someone is being tortured. It may not even be enough for me to simply say, "Stop!" and yet do nothing more to

---

[8] Interestingly, Austin thought analysis of illocutionary acts was crucial for the task of metaethics: "we shall not really get clear about this word "good" and what we use it to do until, ideally, we have a complete list of those illocutionary acts of which commending, grading, &c., are isolated specimens—until we know how many such acts there are and what their relationships and inter-connexions are" ([1962] 1974, 164).

prevent the torture from occurring. Quite the contrary, we treat the holders of moral beliefs as properly criticizable for not *acting* in accordance with those beliefs. We will explore further the close connection between moral mental states and non-verbal behavior in the next chapter. For the moment, let us agree that moral mental states *tend* to issue in non-verbal behaviors and we indeed *expect* them to do so with more regularity than non-moral mental states.

Though moral mental states seem to exhibit a wide range of possible mental attitudes, (meta)ethicists have almost always focused on attitudes of credence (e.g., belief, judgment, thought) rather than more uncertain ones (e.g., wonderings, doubts). Perhaps this is because the nature of moral experience is better characterized by deeply held moral convictions than by disaffected musings on the nature of morality; we seem to encounter assertions of settled moral attitudes more frequently than we encounter Socratic individuals who are open to questioning many of their moral views. Perhaps, also, philosophical interest in truth and epistemological justification incline us toward mental states of assertion and certainty rather than more agnostic pursuits. Hence, there is a preoccupation in the literature over whether moral mental states are truth-conditional and whether moral statements may be analyzed semantically in terms of truth. Although discussions of learning and acquisition of moral wisdom, with their attendant features of wonder, doubt, ambiguity, even humility in the face of complex subject matter, is found in the literature—and, in my view, an important part of the overall task of ethical theory—it is encountered much less frequently, and often in

normative theory rather than in metaethics.[9] Moral beliefs and judgments are mainstay.

Though this fetish, by itself, gives no grounds for eliminating moral doubts and wonderings from our present discussion, we have good reason to do so because of the strong connection between moral thought and action, which only accompanies credence-based attitudes. Moral wonderings and doubts do not directly guide actions, nor do we expect them to do so. If I merely *wonder* whether torture is wrong, I have no decisive reason to oppose it. In fact, one might claim it would even be irresponsible of me to act on a moral wondering or doubt in which I had no (or even very little) credence. Consider here the range of historical abuses that have been committed simply because people have weak or virtually non-existent adherence to certain religious or moral codes. While absolute certainty of belief may be too strict a standard, it does seem that one should have some minimal standard of belief in a moral principle before acting on it. In other words, a mental state, in order to be a *moral* mental state in the relevant respect, must have a partially credence-based mental attitude. Moral doubts and wonderings fail to meet this standard and so fail to require us to act—though we may have good reason to investigate things further and come to a settled opinion on the matter. For this reason, it makes sense to treat them separately from the familiar moral mental states of beliefs, judgments, thoughts, and so on.

---

[9] This position is especially well reflected in work on literature and moral knowledge. For examples, see MacIntyre ([1984] 1986, esp. 215–25), Meyers (2004), Nussbaum (1990, esp. 3–49), and Wilson (1983).

In what follows, then, I will use 'moral mental states' to designate moral mental states with the familiar cluster of credence- or commitment-based attitudes that would normally be expressed by remarks with assertoric or commanding illocutionary force. This approach should not be taken to prejudge the question of whether moral mental states are truth-conditional. It is important to take seriously the debate between cognitivists and noncognitivists about the truth-conditions (or lack thereof) of moral beliefs and expressions of moral beliefs. As we will see, truth-conditions are one of *many* metrics of evaluation that enter into discussions of moral mental states. But even if the content of these states is not, strictly speaking, true or false, one can use such terms in a looser sense to mean recommended or not recommended. Even noncognitivists can agree that there are better and worse plans to follow. Thus, we might think of the content of these mental states as being able to be *treated* as true or *rejected* as false. We can act *as if* charity is good and refrain from torture *as if* it were wrong, even if the content of our thoughts about those actions are not, strictly speaking, true or false. Having addressed issues of mental attitude for moral mental states, we may now briefly consider their content.

1.1.2. Moral Content

In the foregoing discussion, the moral mental states in question have primarily concerned the forbidden, permissible, or obligatory nature of certain actions. A person may believe he/she should not eat meat, think that abortion is forbidden, and so on. These moral valences form the core of moral theory and its principles.

However, many authors have drawn attention to so-called "thick" moral terms. Bernard Williams describes these as "union[s] of fact and value…[where t]he way these notions are applied is determined by what the world is like (for instance, by how someone has behaved), and yet, at the same time, their application usually involves a certain valuation of the situation, of persons or actions" (1985, 129).[10] Common examples of such terms include "treachery," "promise," "brutality," and "courage." These thick terms, Williams argues, are usually shaped by particular cultures, and terms encountered in one culture are not always found in others. Thus, thick terms pose a special problem for any broader discussion of moral mental states because the particular moral content in question may be culturally relative and generalization across cultures may not be possible.

In what follows, I focus primarily on thin moral terms, such as "good," "bad," "right," "wrong," and "permissible," in my discussion of moral mental states. This is not because I think the distinction between thin and thick terms is unintelligible or because I think that thick moral terms do not figure in to moral mental states. On the contrary, I suspect they occur quite frequently across the entire range of moral mental states. At the formal level, however, it seems possible to resolve thick moral terms into their descriptive elements and their evaluative components. In other words, any thick moral concept may be treated as a concatenation of a non-moral description and a thin moral concept. Consider these examples:

CHARITY = (GIVING RESOURCES TO OTHERS) + (OBLIGATORY/PERMISSIBLE)

---

[10] For similar discussions, see Anscombe (1958) and Foot (1958, 1959).

GRATITUDE = (BEING THANKFUL FOR ANOTHER PERSON'S ACTION) + (OBLIGATORY)

BRUTALITY = (PERFORMING EXCESSIVE VIOLENCE) + (FORBIDDEN)

COWARD = (A PERSON LACKING RESOLVE) + (FORBIDDEN/PERMISSIBLE)

Though there may be disagreement about exact correctness of the descriptive portion of the formula or the particular strength of the valence in question (i.e., permissible or obligatory), there is *general* agreement on the meaning of the term and whether it is something to be recommended or guarded against. Since these details will be subject to vagaries of meaning and interpretation anyway, we need not concern ourselves with *exact* definitions of the thick moral terms and agree simply on the enthymeme presented above.

The meanings of thin moral terms present fewer problems. Simon Blackburn, for example, gives helpful definitions of them as follows:

> deductive relationships between norms can be studied by thinking of ideal or relatively ideal worlds in which the norms are met. If we have here the basis for a logic, it extends to attitude. For *H*!*p* can be seen as expressing the view that *p* is to be a goal, to be realized in any perfect world. A world in which ~*p* is less than ideal, according to this commitment. The contrary attitude *B*!*p* would rule *p* out of any perfect world, and corresponding to permission we can have *T*!*p*, which is equivalent to not hooraying ~*p*, that is, not booing *p*. (1988, 508)

Blackburn's attitudes of hooraying, booing, and tolerating *p* correspond to what we might regard as the obligatoriness, forbiddenenss, or permissibility of (doing something that brings about) *p*. These three properties find familiar application in

the normative and applied literature that presents principles about the duties of charity, moral horror of torture, the permissibility of abortion, and so on. Though more could be said about the nature of moral mental content, further discussion will be reserved for the following chapter. At this point, we have a clear account of the thin moral terms and a working method for resolving thick moral terms into their thin and descriptive components.

1.1.3. Further Issues

Moral mental states, like all mental states, bear important connections to extra-mental reality. In Harman's example, the moral mental states we considered were caused, at least partially, by the event of the cat being ignited. In addition, there were behaviors, both verbal and non-verbal, that may have followed from the moral mental states of the observers. In discussing these connections to extra-mental reality, it will help to adapt Wilfrid Sellars' (1974) language concerning "language-entry" and "language-exit" transitions. For our purposes, we may consider "moral-mental-state-entry" transitions, principally the ways in which moral mental states come about and the role that the world plays in that process; we may also consider "moral-mental-state-exit" transitions, namely the behaviors that follow from moral mental states.[11] The most obvious case of mental-state-exit transitions are speech acts expressing those states. But with moral mental states,

---

[11] Sellars also discussed intra-language—or, in our parlance, "intra-mental-state"—transitions by which one moved inferentially from one mental state to another. There are interesting issues here about how moral mental states are connected to non-moral mental states. Among them is the question of how non-moral observations about a situation (e.g., seeing a cat being ignited) are connected to moral observations about that situation (e.g., judging that it is wrong). Some of these questions will be addressed in discussing performativist accounts in Chapter 5.

more than any other mental states, it is crucial to explain how those states are connected to volitions. Philosophical interest in moral mental states may be unique for its focus on the ways in which moral judgments are connected to and prompt action. In what follows, I will give a brief overview of three issues involved in entry- and exit-transition considerations—moral reality, moral action, and moral discourse—with further discussion to occur in subsequent chapters.

1.1.3.1. Moral Reality

Moral mental states, as a sub-species of all thoughts, are about something. But what? Some theorists, among them moral realists, believe that moral mental states represent an independent moral reality and that reflection on our moral experience reveals the existence of and provides evidence for such a reality. Others, including constructivists and expressivists, deny these claims and sometimes deny the existence of any such reality. Although adjudication of this debate requires careful analysis of moral mental states and the theories in question, some general points are worth noting.

Unlike normal cases of sensory states, moral mental states need not be caused by the external world. I can have a thought about something that does not, never has, and never will exist. Moral mental states, in particular, frequently involve future states of affairs that we hope or fear will come to pass. When I reflect morally, I may think about the badness that would result if I break a promise, or the rights that might be violated if certain privacy restraints are not upheld, or the virtue that lies in helping those in need. These cases, and those involving *past*

misdeeds, generosities, and the like, are ones in which the referents of my moral judgments do not, strictly speaking, exist. Only *very* rarely do people make the judgment that would be expressed as, "What is happening right now is wrong." Moral *reflection* almost always involves thoughts about the past or plans for the future and often concerns ways we do not want the world to become. Even if one adopts a possible worlds account on which past and future worlds actually exist, the content of moral mental states will very often refer to worlds that are not the actual world. I will have more to say on this issue and its attendant problems in Chapter 3.

Still, one might think that moral mental states are caused by the external world in a broader sense. Take, first, a non-moral example: if I have a thought of a unicorn, perhaps that thought has been "caused" (in a loose sense) by, say, horses and horns.[12] On this account, as long as the constituent parts of a mental state have been caused by some experience of the world, any combination of them, however complex, might similarly be said to be caused by the external world. This is certainly a doctrine good empiricists should endorse, but it will not do for present purposes as an explanation of the causation of moral mental states.

First, it is not clear that the theory succeeds even in the simplest cases. Upon seeing a horse, for example, I may have a thought of a spring day when I was nine years old. Such experiences are not uncommon and display the highly idiosyncratic connections between experiences and the thoughts that attend them. Consider W.

---

[12] This is the Lockean concept of complex ideas in all its unvarnished form. Still, the model will do for present purposes.

V. O. Quine's (1960) *gavagi* example in which a field linguist attempts to learn the native word for rabbit. Though he hears them utter "gavagi" each time they see a rabbit, he will never be certain whether the word refers to the whole animal, its undetached parts, rabbit-time-slices, etc. It is no more clear that any robust and interpersonal patterns could be discovered that would show a clear causal connection between extra-mental *moral* reality and moral mental states.

A second concern comes from the complexity of moral thoughts. Such thoughts are highly abstract and often concern theoretical constructs, such as obligations or justice, that are simply not detectible by ordinary sensory mechanisms. So even if a situation such as seeing a cat ignited *did* cause moral mental states in *all* observers (which we should doubt), it's not clear that all observers would have *exactly* the same moral judgment. Consider, for example, an act-utilitarian observer and a Kantian witnessing the same situation and both judging the action to be wrong. The concept WRONG will have slightly different content for each of the two observers. The act-utilitarian's conception will have to do with the pain caused to the cat, while the strict Kantian's judgment will, more indirectly, be concerned with how such a situation would cause changes in behavior toward persons, which are the only objects of value in his/her view. This example demonstrates the sense in which reality underdetermines the thoughts one can have about it. The content of two identical judgments caused by the same situation—or, rather, what *appear* to be identical judgments and can only be shown to be different by some amount of further questioning—may differ very slightly but to a degree that makes it implausible that experience directly causes

determinate mental content. And some observers will not see any moral issue in play at all!

These reflections on moral mental states should suffice to show that there is no unique causal route from any purported state of affairs to moral thoughts about them. In fact, the class of theories that has come closest to positing a kind of moral sense-perception, intuitionism, has almost always—or at least most plausibly—held that this mechanism involves reasoning processes and reflection rather than sensory ones involving direct causation.[13] A causal account of intentional states, if one is available at all, will encounter special difficulties with moral mental states. Having settled the issue of intentional-state-entry transitions, let us explore the other side of the issue: intentional-state-exit transitions.

1.1.3.2. Moral Action

Moral mental states, as we have already seen, are intimately connected with action, even if such action is only the performance of a certain kind of speech act. This characteristic, however, fails to distinguish them from mere wants, plans, or even wishes—all of which involve bringing about specific states of affairs in the world. *Moral* mental states, however, aim to bring about the *right* states of affairs and to prevent or destroy worse ones. For this reason, the success conditions for moral mental states may seem to parallel the success conditions for many non-moral

---

[13] For a helpful delineation of varieties of intuitionism, see Williams (1995, esp. 182–84). Williams distinguishes between accounts on which intuitive knowledge is likened to mathematical knowledge and those on which it is likened to perceptual knowledge. Pritchard (1912) is usually regarded as a perceptual intuitionist, though Baldwin (2002) argues that 'perceptual intuitionism' is best replaced with the term 'particularist' and that Prichard is actually committed to act-types, making him a kind of generalist. Contemporary intuitionists clearly fall on the rational side. See Audi (2005) and Huemer (2005).

mental states, which involve veridicality. I can, for example, have a better or worse perception of a person's face in the sense of my being able to accurately describe that face or discriminate it from other similar faces. To take a more theoretical example, I can have a better or worse belief about a particular chemical reaction in the sense of my being able to predict the effects of the reaction or explain how the reaction works in a way that is consistent with my other scientific beliefs. Similarly, one might argue that I can be correct or incorrect about my duties and the degree of their stringency. I may be correct in believing that I owe someone gratitude, or I may be incorrect in thinking that my duty to keep a promise to a friend trumps some small kindness I could perform for a stranger.

However, the primary metric of evaluation for moral mental states cannot *simply* be whether those states are correct or incorrect. If they were, moral beliefs and statements would be in *exactly* the same category as scientific beliefs and statements as mere descriptions of the world. As such, they should either be confirmable through proper investigation or, if they were not, we would be forced to discount them as systematically false or meaningless and perhaps even to eliminate them from our thought and talk.[14] Wittgenstein's "Lecture on Ethics" makes such an overture: "Ethics so far as it springs from the desire to say something about the ultimate meaning of life, the absolute good, the absolute

---

[14] The biggest mistake of error theory seems to be its assumption that moral statements have only mind-to-world satisfaction conditions. On Mackie's (1977) view, for example, moral statements are systematically false because they purport to reflect stance-independent moral facts and properties when, in fact, there are no such entities. On this point, emotivists might agree. But emotivists believe that the role of moral beliefs has little to do with their mind-to-world satisfaction conditions anyway; what's more important are the world-to-mind conditions whereby moral attitudes and statements may bring about certain changes in the world. So even if moral mental states were systematically false, they might be able to succeed in other, important ways.

valuable, can be no science. What it says does not add to our knowledge in any sense" ([1939] 1997, 70). Stevenson agrees with this rejection of the analogy between ethics and science, but rescues the role of moral statements by according them another purpose: "the 'goodness' of any thing must not be verifiable solely by use of the scientific method. 'Ethics must not be psychology'…Doubtless there is always *some* element of description in ethical judgments, but this is by no means all. Their major use is not to indicate facts, but to *create an influence*. Instead of merely describing people's interests, they *change* or *intensify* them. They *recommend* an interest in an object, rather than state that the interest already exists" (1937, 16–17). Moral mental states are important because their success conditions involve *prescriptive* force—that is, their ability to directly recommend or forbid certain courses of action and to guide or goad others to act accordingly. Thus, when I say, "Torture is wrong," it's not enough that I might be correct that torture should be banned; what also matters is that torture not be practiced.

Given the widely accepted view that moral mental states involve both some standards of correctness and some standards of success in action, the task for any metaethical theory will be to accommodate both kinds of success conditions. In successive chapters, we will evaluate the extent to which various accounts succeed or fail in reconciling the two.

1.1.3.3. Moral Discourse

In the previous section, we explored the connection between moral mental states and non-verbal behaviors. But moral mental states also issue in a wide range of

verbal behaviors that have formed a cornerstone of metaethical analysis. As we have seen, linguistic expressability is a hallmark of moral mental states and, indeed, of many mental states. The exact connection between moral mental states and language, however, has proved a contentious issue.

In the tradition of Frege and Russell, many theorists have treated moral statements as expressing propositions that exhaust the intentional content of the mental states that occasion those statements. On this analysis, the intentional content of any statement can be set off using a 'that-' clause embedded in a statement that reflects the appropriate mental attitude. Thus, John's saying, "Murder is wrong," may be recast as "John believes *that* murder is wrong," and Susan's worry, "Stealing may not be permissible in this case," may be reformulated as "Susan doubts *that* stealing is permissible in this case." Since propositions are the primary bearers of truth, this account entails that all moral mental states (strictly speaking, moral mental content) are capable of being true or false.

This account is controversial. Ayer, for example, brands moral utterances as "unanalyzable, inasmuch as there is no criterion by which one can test the[ir] validity….they are mere pseudo-concepts" ([1946] 1952, 107). In his view and that of others, moral mental states lack truth conditions because there is no method by which truth-values could be assigned to moral statements. Accordingly, treating moral statements and their attendant moral mental states as expressing propositions with truth conditions seems to beg the question in favor of moral realism, which holds among its central tenets the bivalence principle that every moral statements is determinately true or false.

The propositional account can be retained, however, if we expand the satisfaction conditions for propositions along the lines discussed in the previous section. There is no reason, in principle, why our notion of propositions could not be expanded to cover *other* kinds of satisfaction conditions, such as success or failure in action. Thus, we might talk of a proposition being satisfied or unsatisfied, fulfilled or unfulfilled, and so on.[15] Though the formal apparatus for drawing inferences from and quantifying over these propositions is less developed, attempts at non-truth-functional logics and multi-variable logics—which include, notably, deontic logics—reflect the fact that regimentation is not limited to truth-functional statements. As a result, it is possible, in principle, to imagine propositional calculi that cover different kinds of satisfaction conditions.

For the present, we may take away the more banal point that the intentional content of statements is roughly equivalent to the intentional mental states that they express. Saying that it's raining expresses my thought that it's raining, thanking someone expresses my gratitude, telling someone that rain is likely expresses my expectation that it will rain, and so on. As David Rosenthal puts it, "Every sincere speech act expresses an intentional state with the same content, or nearly the same, as the speech act and a mental attitude that corresponds to the speech act's illocutionary force" (2005, 264). My statement, "It's raining," expresses my belief that it's raining with an assertoric force that corresponds to my mental

---

[15] Perhaps some would prefer to call this kind of content "dispositional content" and treat it as occurring in parallel with propositional content in many intentional states, including moral mental states. But where posits are concerned, I favor fewer, not more, so I prefer to regard this suggestion as one about two species of propositional content.

attitude of believing, and similarly for other states. When I say, "I doubt if it's raining," I express an intentional state with the mental attitude of doubting and a content involving the state of the weather. Expressions have "nearly" the same content as the intentional states they express because it is always possible that one's mental states outstrip one's vocabulary, or that one misuses one or more words in attempting to express a mental state, or that the meaning of terms varies ever so slightly, even for speakers of the same language. As Quine (1969) pointed out radical translation begins at home.

This point suggests that one method for investigating the intentional content of moral mental states is investigating the intentional content of moral statements. While this strategy has a long and central history in metaethics, it is also possible that authors have been too preoccupied with the syntax and semantics of English-language moral statements. Very crude ordinary-language analysis fails to account for the fact that moral terms in some languages may be more or less fine-grained in the sense that they may not *fully* capture the intentional content of speaker's thoughts. Some languages have richer resources for expressing wishes or commands in ways that might make moral statements appear less descriptive than they do in others.[16] The "English-language fallacy," as one might call it, would be drawing faulty conclusions about the nature of mental

[16] These languages would probably surpass English in their number and variety of irrealis grammatical moods, which are aimed at changing states of affairs in the world, particularly other people's attitudes and plans of action. Some relevant moods might include the imperative/prohibitive (used to express commands, requests, and prohibitions), optative (used to express hopes, wishes, or commands with a subjunctive flavor), jussive (used in Arabic to express pleas, hopes, wishes, etc.), and cohortatitve (used to express pleas, hopes, wishes, etc. and interchangeable with the jussive in Latin).

states from syntactic and semantic features of expressions of those states in the particular language of English. Though expressions may be our best source of information about intentional states, we should be cautious about just how much we can conclude from them.

1.2. Conclusion

Metaethicists, in studying of the foundations of value, are inevitably drawn back to considerations of moral mental states: the thoughts, beliefs, judgments, hopes, desires, and other mental states in virtue of which our moral discourse has its meaning and through which we effect changes in the world. In considering these states, they must address both the sense in which moral mental states may be responsible for accurately reflecting purported moral reality and the sense in which moral mental states are responsible for occasioning actions and their effects upon the world. These considerations touch on perennial metaethical questions of moral reality, moral action, and moral discourse. It is to those questions we must now turn as we give careful consideration to the nature of moral mental states and their implication.

Chapter 2

Moral Mental States

> We possess impressions with *physical* content. These exhibit to us sensuous qualities located in space. Out of this sphere arise the conceptions of colour, sound, space, and many others. The conception of good, however, has not its origin here. It is easily recognizable that the conception of the good like that of the true, which, as having affinity, is rightly placed side by side with it, derives its origin from concrete impressions with *psychical* content.
>
> —Franz Brentano, *The Origin of Our Knowledge of Right and Wrong*

In the last chapter, we saw that several central questions in metaethics turn on the issue of moral mental states—those thoughts, beliefs, judgments, and similar mental states that occasion verbal expressions and non-verbal behaviors, such as trying to prevent certain things from happening or facilitating others. These questions include the existence and nature of moral reality, the relation between mentality and moral action, and the intentional content of moral expressions. We also agreed that states such as moral doubts and wonderings were importantly different from recognizable moral mental states in respect of their failure to incline individuals toward action by comparison with beliefs or convictions. I noted that metaethicists have paid far too little attention to the issue of moral mental states, and I alleged that existing metaethical theories fail to fully account for them.

In this chapter, I want to get clear on the exact nature of moral mental states and how their features constrain what counts as a plausible metaethical theory.[17] Among the questions to be considered here are: What is a moral mental state? How are moral mental states different from non-moral mental states? Where do moral mental states fit in the overall taxonomy of all mental states? These questions concern both the *content* and the *structure* of moral mental states; we are interested not only in the distinctive character of moral mental content but also the kinds of mental states in which that content occurs, including (possibly) sensations, thoughts, and volitions. In the course of this discussion, I present and defend four propositions about the nature of moral mental states. These propositions will serve as criteria for assessing the adequacy of existing metaethical approaches—externalism, internalism, and performativism—and for constructing alternative ones.

I begin by considering methods for individuating moral mental states—specifically, methods for demarcating them from non-moral mental states. After rejecting two common approaches, I argue for a functional account on which moral mental states are picked out by the distinctive role they play in one's overall mental economy. This will establish a few basic criteria with which to later evaluate metaethical theories. In specifying this role in detail, I present the remaining propositions.

---

[17] For a discussion of the epistemological implications of moral mental states, see Morrow's (2009) account of "moral appearances."

2.1. Individuating Moral Mental States

A preliminary problem in discussing moral mental states is determining what separates non-moral mental states from moral mental states. Both kinds of states can involve some of the same content or be about the same event. I can make a moral judgment about lighting cats on fire just as much as I can make a scientific judgment about the rates of the chemical reactions involved. Both, moreover, are subject to some standards of confirmation, albeit different ones according to which metaethical theory one adopts. What, then, makes the former judgment moral and the latter judgment non-moral? In this section, I will address questions of individuation along three different lines. As I will argue, the first two face serious problems at the metaethical level, and a functional account, properly construed, offers the best prospects for demarcating the special province of moral mental states.

2.1.1. Substantive Accounts

The most common approach to individuating moral mental states turns on what the states themselves are about. Call this a "substantive" approach because it attempts to individuate with respect to the *content* of the mental state in question. A mental state is a *moral* mental state in virtue of its involving rights, utility, character, interpersonal relations (Joyce 2006, 65–66), or luck-corrective policies (Rawls [1971] 1999, 86) (as you like). What differentiates the moral judgment about lighting cats on fire from scientific ones is that, in the former, my representation of the situation involves aspects of rights, utilities, and so on that are absent in the

latter. Normative accounts vary, of course, according to which of these content(s) they endorse. Where a virtue theorist will separate moral from non-moral mental states according to whether those states involve certain habits of character, a Rawlsian will do so (at least where justice is concerned) based on whether the content of the state involves arranging the basic structure of society in a way that distributes rights and burdens fairly.

A notable subset of substantive accounts is formal accounts on which moral mental states are defined by the correspondence of their content to a particular structural criterion. William Frankena gives a helpful sampling of these views as follows:

> According to the former [wider formal concept], an AG (action-guide) is a morality (a moral AG as opposed to a nonmoral AG) if and only if it satisfies such formal criteria as the following, regardless of its content:
>
> (A) *X* takes it as prescriptive.
>
> (B) *X* universalizes it.
>
> (C) *X* regards it as definitive, final, over-riding, or supremely authoritative.
>
> According to the second, narrower concept, *X* has a morality, or moral AG, only if, perhaps in addition to such formal criteria as A and B, his AG also fulfills some such material and social condition as the following:
>
> (D) It includes or consists of judgments (rules, principles, ideals, etc.) that pronounce actions and agents to be right, wrong, good, bad, etc., simply because of the effect they have on the feelings,

> interests, ideals, etc. of *other* persons or centers of sentient experience, actual or hypothetical (or perhaps simply because of their effects on humanity, whether in his own person *or* in that of another). Here 'other' may mean "some other" or "all other."
>
> On this conception, a morality must embody some kind of social concern or consideration; it cannot be purely prudential or purely aesthetic. (1966, 688–89)

In addition to these examples, Frankena cites Kant as proposing "that a rational man can adopt a moral AG only under the presupposition that it is ultimately rational for him to live by it" (695). Though these formal accounts are not described as specifying any particular content—and, indeed, a chief objection to Kant's view was that it amounted to an empty formalism[18]—they indeed concern the *content* of the mental state and not, say, the instantiation of the state itself as a thought as opposed to, say, a perception or a wish. They require that the mental content of the moral mental state meet a particular set of conditions (e.g., that it prescribe or be universalizable or overriding), which is no different, in principle, from the condition we just examined: that the content involve rights, utilities, character, and so on.

Although substantive accounts are prevalent (and perhaps implicitly assumed by most ethicists), they suffer from several problems. First, at the metaethical level, they seem to beg the question in favor of particular normative theories. In adopting one view of moral content over others, we rule out by *fiat*

---

[18] For an overview of the "empty formalism" objection, see Lo (1981).

entire classes of possible moral judgments. This technique has been used all too frequently to dismiss normative theories or moral epistemologies that diverge from the author's favored views. A few examples should bear this point out:

- Kant's loathing of the "serpentine windings of this immoral doctrine of expediency" reflect his own view that "expediency" or utility is simply not the foundation of morality ([1795] 1891). "A good will," he insists, "is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end, but only because of its volition, that is, it is good in itself…" ([1785] 2002, 4:394). This definition, however, immediately excludes act-utilitarian theories on which one is permitted to maximize expected utility even if it violates the formal test of the categorical imperative.[19]
- More recently, Jeanette Kennett has deployed a Kantian account of moral judgment to dismissing concerns that psychopaths, who are amoral and rational but lack affect, offer empirical evidence against rationalism and for sentimentalism: "It is not until sympathy fails and the man, for the first time, deliberates on whether he nevertheless is required to help those in need, concludes that he is so required, and overcomes his "dead insensibility" to act in accordance with his judgement, that his judgement and actions count as moral. The critical step is deliberation about what, if anything, he has normative reason to do in the situation" (2006, 79). In Kennett's view,

[19] My own view is that rule utilitarianism is the most plausible form of utilitarianism and does not, in fact, violate the categorical imperative because doing so would actually produce less utility over the long run.

psychopaths aren't making *moral* judgments at all because they are not deliberating and self-regulating their behavior in the proper way. As a result, she claims, they do not constitute evidence against the rationalist doctrine that the rational choice is the moral one or evidence for sentimentalism to the effect that those lacking affect are incapable of making the moral choice.

- One final example: In a famous paper, G. E. M. Anscombe indicts Bishop Butler, who "exalts conscience, but appears ignorant that a man's conscience may tell him to do the vilest things" (1958, 530). But what licenses Anscombe to claim that any act dictated by a person's conscience is wrong? Our judgments about the rightness (wrongness) of acts will always depend on normative assumptions about what is permitted (forbidden) by ideal moral theory. So the objector will need to adopt a certain moral theory to assess the act in question. But then she is on the horns of a dilemma: If the she adopts theory *T*, she will judge the act to be right, just as the theory prescribes. In this example, if the moral worth of some act prescribed by conscience theory is assessed in terms of conscience theory, that act will turn out to be *right*, not wrong; it must be the case that any act prescribed by a theory is right in view of that theory. So to get the objection off the ground, the objector must adopt some theory *other* than *T* that contradicts *T* by issuing in a counter-prescription about the act in question. But this is precisely what the objector cannot do. In using some theory that contradicts

T, she assumes that *T* is inadequate. This claim should be the conclusion of the argument, not part of the supporting premises.[20]

These rather diverse examples illustrate the common problem of substantive accounts: in adopting a particular view of moral content on a substantive account, one thereby excludes all other views and the theories that employ them. It is, of course, possible that theories will converge on similar prescriptions. The Kantian and utilitarian can both agree that torture is wrong. But the content of the moral mental state in the minds of the two theorists is entirely different, for the meaning of RIGHT (WRONG) for one theorist is quite different than the meaning of RIGHT (WRONG) for the other. What is wanted in metaethics, though, is a normative-theory-neutral characterization of moral mental states, and substantive accounts are ill-suited for this purpose.

A second (and related) problem is the fact that substantive accounts rule out the *possibility* of making wrong moral judgments or holding incorrect moral beliefs on too many occasions. Consider the strict Kantian who sees the utilitarian not as wrong in his/her views but rather as not making moral judgments at all! This formalist is forced to say that any judgment that does not conform fully to the categorical imperative simply isn't a moral judgment. There's no getting it wrong, as it were, in such a case; one either applies the correct decision procedure or one applies some other procedure, which *ex hypothesi* leads to a judgment that is not a *moral* mental state. Sharon Street makes this move in a recent article on

---

[20] For a similar objection to this methodology, see Singer (2005).

constructivism: "to explain this idea of entailment [from the practical point of view on constructivist accounts], we need only make observations about what is constitutively involved in the attitude of valuing or normative judgment itself—identifying those things such that if one fails to do them, one is not making a *mistake* of any kind, but rather not recognizably *valuing* at all" (2010, 367).

To put this point a bit differently, the formalist must enquire as to whether a person purporting to make a moral judgment knows how to play the "moral game," as it were, by following a rule that involves correspondence of that moral judgment to certain formal criteria. One can inquire, then, if it is possible to make a mistake in the course of playing this game. As Wittgenstein pointed out, "if everything can be made out to accord with that rule, then it can also be made out to conflict with it" ([1953] 1999, §201). In other words, there should be a possibility of error as long as there is a possibility of correctness for any rule-governed behavior. These standards of correctness and error are answerable to the wider community, which has, as Saul Kripke put it, "justification conditions for attributing correct or incorrect rule following to the subject" (1982, 89). The problem here is that there is no agreement in the widest possible community (of all moral theorists) on what counts as success or failure in making a moral judgment (on a substantive view). There is only more local agreement among subcommunities of formal substantive theorists (e.g., Kantians), and there is no reason to expect that those standards should be applied more broadly in ways that rule out arguably genuine moral mental states—not as being incorrect but as not being moral judgments in the first place because they have nothing to do with the favored criteria.

Non-formalists may fare better with this problem, but they are not immune to it. A utilitarian, for example, could assert that a comrade had made a wrong judgment by omitting some utility value or making some error in calculation. However, in a utilitarian's eyes, moral judgments about rights are not, strictly speaking, *moral* judgments unless they involve utilities. If a Kantian makes a "moral" judgment that involves absolutely no utilities, the utilitarian will see him/her as doing something *other* than making a moral judgment—applying a decision procedure or algorithm, perhaps. The point is that on all substantive accounts, there will be cases in which the chosen view of content categorizes some judgments as non-moral because their content does not conform to the given standard. The problem, of course, is that the utilitarian and the strict Kantian are not, in fact, talking past each other. They *disagree*, and substantive accounts are sometimes unable to account for this disagreement.

Even setting aside problems of neutrality, a third criticism of substantive accounts is their failure to define moral content in a sufficiently clear manner. Consider familiar questions of distinguishing between morality and etiquette or mere mob mentality. Both types of rules may be widely held among members of a given group or culture, and both may be enforced (to varying degrees). Kurt Baier partially distinguishes moral rules from etiquette according to the stringency with which the rule is enforced and the feelings that violators experience in and after breaking it:

> If infringers of the rule are said to be immoral, wicked, wrong-doers, evil, morally bad, or some term implying one of these, and they are treated

> accordingly, then the rule is supported by the specifically moral pressure. Whatever may be the precise treatment meted out to those we think we rightly say are immoral, evil, wicket, etc., it is plain that we tend to condemn them, dissociate ourselves from them, perhaps would want to see them punished. It is evidence that the rule is part of the morality of the group if rule-breakers feel guilty and experience remorse. It is evidence that the rule is not part of the group morality if group members feel merely regret or pleasure when infringing it. (1954, 109–10)

To Baier's credit, one might say, he also includes a number of formal criteria (e.g., that a rule not be self-frustrating, self-defeating, or unteachable) in this analysis. But as Philippa Foot famously pointed out, the rules of etiquette can be quite similar to the rules of morality, even down to their (categorical) bindingness on agents regardless of their own inclinations, preferences, desires, and so on: "we find this non-hypothetical use of 'should' in sentences enunciating rules of etiquette…where the rule does not *fail to apply* to someone who has his own good reasons for ignoring this piece of nonsense, or who simply does not care about what, from the point of view of etiquette, he should do" (1972, 308). Foot was roundly attacked for and ultimately retracted this view,[21] but her argument stands, and it brings home the very real problem that substantive accounts have in demonstrating that they have, in fact, characterized the mark of the moral.

---

[21] This lively debate occurs in Foot (1972, 1974, 1975), Frankena (1974), Holmes (1974, 1976), Phillips (1977), and McDowell and McFetridge (1978). There is lively work at present in the empirical moral psychology literature on this distinction. For a helpful overview, see Turiel (1979, 1983); Turiel, Killen and Helwig (1987); Smetana (1993); and Nucci (2001). For critical remarks on this research programme, see Kelly, Stich, Haley, Eng and Fessler (2007).

An important exception here is Catherine Wilson's "semi-essentialist" account of morality as a system of advantage-reducing rules: "Moral rules occupy a sector of the normative realm, just as sofas and chairs occupy sectors of the category 'furniture'. As there are 'good' and 'less good' exemplars of sofas and chairs as well as items that are intermediate between 'sofa' and 'chair', so there are good and less good exemplars of moral rules, as well as rules that are intermediate between prudential and moral rules, or rules of decorum and moral rules" (2004, 12). This account is distinct from the substantive accounts we have examined in two respects. First, it makes no attempt to draw a sharp distinction between moral and non-moral content and, in so doing, insulates itself from the third objection I gave. Second, its permeability prevents it from being used in the question-begging ways I outlined under the first two objections to substantive accounts. Since it makes no fine distinction between moral and non-moral content, it cannot be marshaled in arguments to the effect that a questionable judgment or expression is not a *moral* judgment or expression because it fails to have certain determinate content. I think substantive accounts would do well to assimilate this semi-essentialism. I leave open the question of whether moral content indeed concerns advantage-reduction with reference to others. On my own view described in Chapter 6, I do think that certain moral mental states could be primarily self-regarding, and it is possible that these states concern advantage reduction. If so, Wilson is happy to take these on board: "The semi-essentialist need have no objection to including rules mandating, say, women's haircovering, as moral rules, provided they are not taken without

further explanation to be examples of central or focal moral rules and provided their advantage-reducing feature can be made apparent" (16).

2.1.2. Mongrel/Cluster Accounts

According to what I will call "mongrel" or "cluster" accounts, moral mental states do not exhibit enough underlying uniformity to describe in any great degree of rigor. This may be either because, as we explored in the last section, moral content is not sufficiently homogeneous to describe with any generality[22] or because moral content occurs in a wide variety of mental situations whose unique aspects shape the nature of that content in a way that resists general description. On this latter account, one might argue that mental content and mental attitude are inseparable, as it were, in the sense that mental attitude shapes the very nature of (moral) mental content. Consider again the diversity of moral experiences one can have regarding a single action: I may *observe* an act of murder and find it to be repulsive and wrong. I may *think* further and, upon reflection, decide that, indeed, such actions are wrong. Or I may even *intend* to prevent murders by becoming a police officer or social worker. In each of these mental states, the moral content remains fixed on murder, but the attitude of the states varies widely. Some skeptics about metaethics may find such diversity impressive and conclude that moral mental states do not admit of rigorous, systematic analysis.

---

[22] This position is similar to moral particularism. Although particularism has been developed in response to normative issues, particularists often claim that the complexity of moral situations bars generalization at the level of moral rules. For further discussion of particularism, see Hooker and Little (2000).

This argument for mongrel/cluster accounts fails for several reasons. First, there are many examples of other clustered or complex phenomena for which we have general, systematic theories. Solidity is one such example; although all solids exhibit the same macroscopic properties with respect to their solidity, the phenomena of solidity occurs in different ways depending upon the chemical structure of the material involved. There is no reason to think that simply because moral content occurs across mental states with a wide variety of attitudes it is impossible to give a systematic account of the underlying moral mental content.

Second, if we consider only mental states, it is clear that the same content can occur unproblematically across various attitudes. I can think about cats, have sensory states about the one I have, wish that I had others, intend to get them, and so on. There seems to be no special difficulty raised by *moral* mental content across various kinds of mental states above and beyond problems raised by mental content generally. And if we consider the connection between mental content and verbal expressions, it seems obvious that intentional content remains fixed, as it were, in some cases though mental attitude and illocutionary force may vary. To return to the example of murder, one would, in fact, use the word 'murder' to express the various distinct moral mental states of believing, judging, thinking, and intending actions about or pertaining to murder. We can only adopt a mongrel account of content on pain of rejecting the very plausible assumption that words mean (roughly) the same things across various expressions.

While mongrel/cluster theorists may be right in stressing the limits of a systematic theory of moral mental states, it would be worthwhile to attempt such a

theory as far as possible and advert to the mongrel/cluster account only in the face of absolute failure. With that in mind, let us turn to the third method for characterizing moral mental states, which will be the method employed in this dissertation.

2.1.3. Functional Accounts

Functional accounts pick out moral mental states according to their role in one's overall mental economy. Although those states may have certain similarities in content, those similarities are, strictly speaking, irrelevant to determining whether the state is a moral mental state; *whatever* state plays the given role in one's overall mental economy counts as a moral mental state in virtue of its functional role.

The most common class of functional accounts are those that make moral mental states hinge on specific emotions, such as guilt (Gibbard 1990) or empathy (Nichols 2004).[23] On these views, all moral judgments must involve specific emotions in order to count as moral judgments. The problem with such accounts is twofold. First, it is hard to know how to separate moral mental states from other mental states involving the specified emotion. If there are no additional criteria besides the emotion, it will be impossible to tell feelings of, say, guilt from moral judgments. The two are surely different, and cases of abnormal psychology, such as psychopathy, offer evidence that a person could recognize guilt and report feelings of guilt and yet not exhibit characteristic moral behavior as a result.[24] Second, the

---

[23] Nichols claims that affect, particularly empathy, is contingently connected to morality.

[24] Haidt (2001) credits Hume with recognizing that "a person in full possession of reason yet lacking moral sentiment would have difficulty choosing any ends or goals to pursue and would look like what we

strong influence of culture on emotions and their expression makes it difficult to identify a set of universal emotions felt by all peoples that would form the basis for identifying moral mental states in all of them. The proper conclusion to draw might be that moral mental states *themselves* vary across cultures and no systematic account is possible. But, as with the mongrel account, we would do well to forestall that conclusion as long as possible and instead attempt to develop a functional account insofar as one is possible.

The first task for such an account would be determining the class(es) of mental states to which moral mental states belong. According to the standard taxonomy, mental states may be some combination of

- sensory states, the inputs of cognition that begin with stimulation of sensory organs and usually give rise to perceptions;
- qualitative states, which have a distinctive "feel" about them and are often taken to be conscious;
- intentional states, more highly conceptual episodes that are often expressed in language; or
- volitional states, the immediate mental precursors of action—mental tryings.

Mental states may also be *hybrid* in the sense that a single mental state may have

---

now call a psychopath" (816). Accordingly, one could know *that* people usually feel guilt in certain situations and even say *that* one feels guilt but not *actually* feel it or exhibit characteristic moral behaviors. Such examples do not disprove the link between morality and feelings of guilt. The point is that it's extremely hard to know whether psychopaths actually *have* feelings of guilt based strictly on their verbal behaviors. For further discussion of the connection between psychopathy and emotions, see Blair (1995) and Nichols (2002).

aspects of two or more kinds of mental states. The most common example of hybrid states is emotions, which often have intentional, qualitative, and volitional aspects.[25] Not only can I be angry in general—a qualitative state of pure rage—but I can, more often, be angry *about* something, such as your stealing my wallet—which has intentional aspects. Hybrid states are distinct from the mongrel account discussed in the previous section in that the hybrid classification applies to a single state and the mongrel account is a view about the nature of all moral mental states based on the occurrence of a variety of *different* moral mental states.

In the rest of this chapter, I am going to argue that moral mental states are hybrid states with both intentional and volitional components. While it often happens that moral mental states have a qualitative element as well, this is not, strictly speaking, necessary for a state's being a moral mental state. This classification of moral mental states reflects the standard (meta)ethical literature that discuses these states chiefly in terms of beliefs, judgments, thoughts, concepts, and so on—all with volitional elements. But even though nearly all (meta)ethicists would probably accept this classification, it is no trivial assertion. As we will see, there are several features of intentional states that make them unique among other mental states. Among these features is their capacity to be about things that do not exist. This is, on the one hand, crucial for *moral* mental states, since they often concern states of affairs that we hope to bring about and others we wish to prevent from occurring. On the other hand, this feature entails that moral mental

---

[25] I will discuss emotions in some detail in §2.3.2.

states are not necessarily about existent things, a point that will prove especially problematic for moral realists. If we begin by merely assuming that moral mental states are intentional–volitional states, we might beg important questions against externalism from the start. So it will be useful to consider why moral mental states are not some *other* kind of mental state, namely sensations or qualitative states.

2.1.3.1. Why Moral Mental States are Not Sensations

Sensations are exclusively connected to some sensory modality; sights, tastes, smells, touches, and hearings all correspond to dedicated organs that function as input mechanisms for downstream perceptions, thoughts, and so on. Moral mental states, by contrast, seem to have no such mechanisms and instead originate in later cognitive processes. Recall Harman's example of seeing children light a cat on fire. First come the visual, auditory, olfactory, and possibly tactile (heat) sensations of the situation; then come perceptions and judgments to the effect that the children are pouring the gasoline and striking the match and that the cat is on fire; only later does moral content enter in the form of a judgment that the action is wrong.

Recent empirical work using neuroimaging similarly suggests this general upstream perception – downstream judgment model of stimuli and moral mental states. William D. Casebeer and Patricia S. Churchland, in particular, draw attention to the relations between brain regions associated with moral judgments (e.g., frontal cortex and pre-frontal cortex) and those structures that evolved much earlier, such as the brainstem, which regulates "breathing, blood pressure, arousal, thermoregulation, shifts in behavioral state from wakeness to sleep to dreaming,

integration of signals from the all-pervasive interoceptive system that carries signals about a host of features of the inner body, and coordination of inner drives (for food, sex, oxygen, etc.) with perceptions (e.g., flee now – do not feed)" (2003, 175). The more recently evolved frontal areas—found almost exclusively among social animals—allow for more complex motor planning and decision-making, which are critical for moral judgments. As Casebeer and Churchland note, the visual, auditory, and other modal "pathways presumably provide perceptual signals relevant to guiding PFC operations," which seem to include moral judgment (180). "Moral cognition," as they point out, "is part of a broader network of understanding molded by multi-modal interactions with the world. Life is usually not limited to a single dimension of stimulation, but embedded in a rich context" (187).

If moral mental states were sensations, we should expect three things to be true. First, we should expect to find such states earlier in cognitive processes in "older" areas of the brain, just as we do in the case of non-moral sensations. Recent empirical work, however, places moral mental states in more downstream brain areas. While it might be the case that researchers are simply looking in the wrong areas or somehow ignoring an important part of moral judgment, Casebeer and Churchland claim that "the empirical work has illuminated some reliably involved brain systems," and it seems that the opponent here has the burden of producing empirical support for his/her view (172).

Second, if moral mental states were sensations akin to sights, tastes, and smells, we should expect there to be some dedicated sense organ attuned to

properties of rightness, wrongness, and the like. And not only is there no such organ (that we know of), the closest correlates of such a mechanism posited in the literature—intuition and moral sense—fail to offer compelling evidence that moral mental states are primarily sensory states. If an intuitive faculty exists on par with other sensory organs, we should expect dedicated biological and psychological mechanisms to mediate it. Quite the contrary, intuition is always taken to occur in the brain, which performs a vast array of functions, and is usually described in terms of thought or reflection, which are intentional states through and through. There is no reason to think that a moral sense organ exists, much less reason to think that moral mental states are simply sensory states.

Third, if moral mental states were sensations in full standing, we should expect them to occur, at least sometimes, independently of other modes of sensation. We can, for example, isolate the visual system from the auditory system by noting the brain areas that are simulated in the presence of one mode of sensations rather than another (i.e., sights vs. sounds). If, however, we imagine subtracting away the sensory modalities of sight, sound, smell, and touch from Harman's example, it's simply not clear that a moral judgment will occur at all. (Even textual or narrative studies of moral judgment involve some kind of sensations, whether auditory or visual.) This, again, points to the fact that moral mental states rely on upstream processes in cognition, including sensations and perceptions of the situation, rather than actually *being* direct sensations themselves.

Sensations, moreover, are ill suited to explain one important feature of moral mental states: their capacity to be about things that do not exist. Sensations,

under conditions of sobriety, sanity, and favorable environment, always imply the existence of some object in that environment. Of course, one can be wrong about certain sensations (e.g., colorblindness), but the existence of *some* object that occasions the sensation is what distinguishes sensations from hallucinations in which one seems to sense something that is *not* there. However, it should be clear that there are a variety of moral mental states that are not occasioned by one's immediate environment. For example, consider a victim of apartheid who imagines a better world in which all races are treated equally. There are no objects in his/her environment that would occasion such thoughts of equality, and yet he/she can perfectly well imagine it without being said to hallucinate. This is because moral mental states are at least in part intentional states, which can be about things that do not exist, rather than sensations of one's immediate surroundings. Let us consider, then the case of qualitative states.

2.1.3.2. Why Moral Mental States are Not (Always) Qualitative States

Qualitative states are often discussed in connection with sensations, such as the feeling one has upon seeing the particular colors of a sunset. If we have reason to be skeptical that moral mental states were sensations, we might be similarly doubtful that qualitative states—sensations once removed—will fare any better. It is possible, however, to talk about the qualitative aspects of memories, wonderings, imaginations, intentions and so forth. While memories may be explained along Hobbesean or Humean lines as decaying sensations, it seems harder to accommodate these other attitudes into the model of sensations. More likely, they

are, at least in part, intentional states or volitional states. The important question here is whether qualitative states ever occur *independently* of all other kinds of mental states (or whether they always occur in conjunction with, say, intentional states) and, if so, whether moral mental states are best described as qualitative states.

If qualitative states occurred independently, they should not be expressible in language. Expressability is a hallmark of intentional states, not qualitative ones. There are, however, two ways in which qualitative states may figure in language. The first is when a qualitative state is a hybrid state with intentional content, where the intentional content is what is expressed in language. The second way occurs when qualitative states are reported, not expressed, in language. Reporting and expressing differ in respect of their truth conditions: an assertion that reports is true just in case the speaker is in the mental state reported by the statement, while an assertion that expresses is true just in case its intentional content is satisfied by the world. In the case of reporting a qualitative state, a speaker may have a purely qualitative state and another (intentional) state about that qualitative state, where the latter—not the former—is expressed in language. In this case, the presence of a qualitative state is reported rather than directly expressed. Consider, for example, the difference between "I think I'm having a reddish sensation" and…well, it's just not clear what expression of a purely qualitative state would be like. The expression of a hybrid state with qualitative and intentional aspects might be "I'm seeing a particular shade of red," but it's not clear how one would elicit an understanding of that state in listeners without employing shared linguistic

concepts and thereby expressing the state's intentional content.

We *do* count on the fact that moral judgments, principles, values, etc. can all be expressed in language. This point should be obvious enough to the moral realist, who thinks that moral statements are determinately true or false (which are logical properties of sentences) in virtue of how the world is, as well as the constructivist, who aims at some objective standard of moral assessment. Emotivists and expressivists, though far less inclined toward views about the truth-conditions of moral statements, still place a premium on moral language and indeed must do so, since pragmatics play a central role in their analysis of moral statements. At the very least, moral mental states *can* be put in propositional terms—even if those propositions are systematically false or lack truth conditions—which should be sufficient for demonstrating the intentional aspects of those states. So it's clear that all metaethicists are committed to some view about the expressability of moral mental states in language and, to that extent, seem to endorse the view that moral mental states are always, in part, intentional states and not merely qualitative ones.[26] And while it's certainly possible that qualitative states occur in conjunction with moral mental states—probably quite frequently—this hardly seems necessary in all cases. It's quite possible (and the hope of some virtue theorists) that right action can simply become a habit, something that does not

---

[26] In discussing "the moral point of view" and emphasizing the role of moral experience, some internalists seem to collapse the intentional properties of moral mental states into qualitative ones. It is hard to know whether such expressions are merely loose talk on the part of some metaethicists or whether they reflect a more substantive, if unacknowledged, theoretical commitment. It is interesting to note in this connection that Kant and Thomas Nagel, both internalists, are also phenomenalists who place a premium on qualitative experience in their discussions of consciousness. For further discussion of phenomenalism in connection with these metaethical views, see §4.4.

always need to be planned or thought about or deliberated but simply flows from one's character.

2.2. Moral Mental States as Intentional States

In examining the theses that moral mental states are sensations or qualitative states, we have discovered that moral mental states must be, in part, intentional states because only intentional states can be about things that do not exist and only intentional states (or rather their intentional content) are expressible in language. This classification has important consequences for metaethics.

"Intentionality," John Searle tells us, "is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world" (1983, 1). For example, I may have a thought *directed at* a cup of coffee or *about* your shirt, or I may have a desire *for* your sandwich. More contentiously, a map may be about a place, a painting may be about a certain landscape, or even a sentence may be about its content.[27] Though we rejected the view that moral mental states are to be defined in terms of their content, it seems safe enough to say, generally, that moral mental states are about actions, events, objects, institutions, and people in the real or ideal worlds, specifically whether and to what extent they are to be promoted, encouraged, protected, etc. or avoided, abolished, guarded against, etc.

---

[27] There is significant debate whether these are cases of intrinsic or original intentionality on a par with the intentionality of mental states, or whether they are merely derived or secondary cases that depend on the former to cause them or to interpret them in such a way. Notably, Dennett (1978, 1987) rejects the distinction and advances his own view of "instrumental" intentionality. On the status of sentences, see Haugeland (1981), Searle (1980, 1983, 1992), and Fodor (1987).

Moral mental states, as intentional states, exhibit a variety of features that are important to metaethical considerations. Among them are intentional inexistence, their ability to be about things that do not exist, and various directions of fit, their ability to be satisfied (or not) along a variety of metrics. In this section, I will apply these considerations to some of the questions we examined in the previous chapter: their connection to moral reality, their connection to action, and their function in moral discourse. In each case, the intentional properties of those states will serve to explain why moral mental states exhibit the features we already noted.

2.2.1. Intentional Inexistence

Franz Brentano, who revived the Scholastic notion of *intentio*, meaning to point (at) or aim (at) or extend (toward), believed that intentional states were unique in their capacity to be about things that do not exist: "This intentional inexistence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves" ([1874] 2009, 68). The most controversial part of Brentano's thesis is its apparent reference to mental objects—objects that need not exist in the external world but are the subjects of intentional mental states. We need not endorse this ontology to appreciate his point that the mere fact *that* I have an intentional state about something does not entail that that something exists.

Quine (1960) made essentially this same point differently in discussing the

opacity of intentional states. As he argued, intentional clauses do not allow substitution of co-extensional concepts without changing the truth-value of their larger contexts.

Tom believes that Cicero denounced Catiline

and

Catiline is Tully

may both be true, and thus 'Catiline' and 'Tully' will refer to the same person in the common tongue, but nevertheless we cannot derive that

Tom believes that Cicero denounced Tully

via a modus ponens substitution instance. The real-world identity of Catiline and Tully outstrips the intentional content of Tom's thoughts about Catiline. Though Quine's examples of referential opacity are usually discussed in terms of aspects, guides, or representations of an object *qua* something, they appeal to the phenomenon of intentional inexistence, for they bar substitution of real-world entities across intentional contexts. Indeed, Quine himself discusses opacity in connection with intentionality when he concludes that intentional vocabulary cannot not be reduced to non-intentional (e.g., extensional) vocabulary.[28] Accordingly, we have no reason to expect that the aboutness relation—whether it

---

[28] Quine went on to say that we could either accept the "indispensability of intentional idioms and the importance of an autonomous science of intention" and thereby reject a physicalist ontology, or we could accept physicalism and renounce the "baselessness" of intentional idioms and the "emptiness" of a science of intention (221). Many influential figures, including Field (1978), Fodor (1983, 1987), and Dretske (1981), have rejected the distinction, trying to marry a physicalist view of the world with intentional realism. For what it's worth, I think Quine's dilemma is tenable as long as we recognize that its second horn is not so bad. We could, following Dennett, regard all cases of intentionality as mere ascriptions of intentionality, or we could advert to Sellars' (1962) notion of the "manifest image," which serves as a contrast to the scientific image of mankind and includes such things as values, agents, persons, and perhaps even intentionality.

is about whole objects or mere aspects—*must* be about things in the real world.

Intentional inexistence and referential opacity have far-reaching consequences for metaethics. As I have already argued, moral mental states need not be about actions, events, objects, institutions, or people that actually exist. Previously, I explained this feature in terms of the distinction between actual and ideal states of affairs and the difficulty of explaining the origins of moral mental content on a strictly causal model. We can now add to this list the feature of intentional inexistence. Moral mental states, in virtue of their status (partially) as intentional states, can be about their contents without implying the existence of those contents. I can form a moral belief about a situation that is not occurring at present, never has in the past, and possibly never will even in some distant future. This ability, of course, is crucial for imagining other ways the world might be and making judgments about what an ideal (or more ideal) world would look like. On the other hand, it does make the task of determining the nature of any extra-mental reality much harder. There is no simple route from the occurrence of moral mental content to any extra-mental moral reality. This point will be considered at length in discussing externalist accounts in Chapter 3. For now, let us agree on the first criteria for metaethical theories in light of the nature of moral mental states:

C1 The presence of a moral mental state does not by itself imply the existence of some extra-mental moral reality, including moral facts and properties.

One may be able to provide additional arguments (probably abductive) to the effect that an extra-mental moral reality is the best explanation available for moral

mental states. However, that point will have to be argued. The importance of (C1) is that from the simple fact of there being moral mental states at all, we cannot draw the further conclusion that there exists any moral reality. As I will argue in Chapter 3, some arguments for moral realism have run afoul of this criterion by arguing from the phenomenology of moral mental states to the existence of a stance-independent moral reality.[29] Such arguments should be rejected in light of the property of intentional inexistence that moral mental states exhibit.

2.2.2. Directions of Fit I: Mind-to-World Satisfaction Conditions

Searle (1983) discusses intentionality in terms of "conditions of satisfaction," a phrase coined by Austin ([1961] 1979, 188).[30] On Searle's view, intentional states such as thoughts, beliefs, hopes, wishes, and desires are things that can go right or go wrong, i.e., be true or false, veridical or nonveridical, fulfilled or unfulfilled. These conditions of satisfaction vary in "direction of fit" along two dimensions: some

---

[29] It is worth noting that Moore's open question argument—which is, as Thomson (2006) says, the foundation of twentieth-century metaethics—turns on the issue of intentional inexistence. In *Principia Ethica,* Moore inquires as to the definition of 'good'—not what things are good (i.e., *the* good), but the nature or essence of goodness itself ([1903] 1993). He notes that even though naturalistic properties such as desire or pleasure co-occur with goodness, it is an open question whether these things *are* good, meaning that the terms are not synonymous and the properties, non-identical. Thus, he concludes, goodness must be non-natural and any attempt to reduce it to a natural property is a naturalistic fallacy. To recast this discussion a bit: our thoughts about the good are about something, but even though we regularly find naturalistic properties in the objects in which we find good, it is a logical error to identify good with these properties. The fallacious argument moves from the naturalistic features that all good objects possess to the claim that their goodness is identical to their common naturalistic features. The flaw in this argument, which Moore correctly identified (but failed to put in these terms) is that from the fact that a person thinks an object is good, nothing follows about the existence of goodness, much less that goodness is identical to a natural property. One can argue, of course, that goodness' being identical to such a property is the best explanation of why the person thinks something is good (or other ampliative arguments of this sort), but as I have been saying, the point remains that there is no direct argument from a person's moral mental states to the existence of something (e.g., a property of goodness which his/her states are about). Moore used this principle to argue against naturalism, but it applies more broadly.

[30] Austin does not discuss intentional states *per se*, but rather uses "direction of fit" to discuss the difference between fitting a name to an item and fitting an item to a name.

intentional attitudes have a direction of fit that is "mind-to-world," where the intentional attitude is responsible for matching the independently existing world; others have a "world-to-mind" direction of fit, where the intentional attitude is responsible for trying to bring about a certain state of affairs in the world. Borrowing an example from Anscombe ([1959] 2000, §32), Searle discusses a man with a grocery list who tries to make the world (in the form of purchases) fit the list, which acts as an order or a desire. Here, success is determined by how well the world has been made to match one's intention (i.e., mind). By contrast, a detective following the man and noting the items he purchases makes a list that functions in the opposite direction; the detective's list succeeds or fails depending on how well it represents the world. The shopping list has a world-to-mind direction of fit; the detective's list has a mind-to-world direction of fit. In each case, the satisfaction conditions reflect the direction of fit for the attitude of the intentional state.[31]

At first glance, the very term 'mind-to-world' direction of fit can be rather misleading, for it seems to imply that we are talking about the actual world. But there are plenty of other cases of intentional states that are not about the actual world and that still have mind-to-world satisfaction conditions. Counterfactuals are

---

[31] Searle also discusses a third class of intentional states such as being sorry or feeling gratitude, which have a "null" direction of fit and "propositional contents that may or may not be satisfied" (103). My being sorry depends on certain beliefs about you being wronged, about my being responsible, and so on. These beliefs, like all beliefs, have a mind-to-world direction of fit. If I learn that I was mistaken about causing you some harm, I similarly learn that being sorry is inappropriate. Put generally, null states have satisfaction conditions that are derived from the satisfaction conditions of mind-to-world states that constitute them. Moreover, states such as being sorry or feeling gratitude also seem to have a world-to-mind direction of fit that involves, say, apologizing or expressing one's gratitude, whether verbally or non-verbally. As I discuss in §2.3.1, the performance conditions and corresponding world-to-mind satisfaction conditions for these states may be indeterminate, but that does not show that the conditions are absent. Most likely, the cases Searle has in mind as "null" states are hybrid states, the satisfaction conditions of which can be resolved into their mind-to-world and world-to-mind components.

prime examples. When I say that such and such would have happened had I not prevented it, I am not talking about the actual world but my statement nonetheless seems answerable to something—probably to the nearby possible world in which I did not intervene. The mind-to-world satisfaction conditions for many moral mental states may be in quite the same category. The satisfaction conditions that obtain (or not) when I say that mass genocide *would* be an atrocious wrong map not onto the actual world (pray not!) but onto the nearby possible world in which the plan is carried out, or onto the (class of) ideal world(s).[32]

The bottom line of mind-to-world satisfaction conditions seems to be that there is something approaching objective standards of evaluation for moral mental states—that the state could be tested in some way or verified or at least examined by other thinkers or observers who could reach a similar judgment about it. This point is usually not in doubt among (meta)ethicists; even though theorists may disagree on exactly *how* things should go, they do seem to agree that there is *some* metric of assessment beyond the single individual. Thomas Nagel calls this the "objective pretensions" of moral discourse, the sense in which agents see themselves "as acting for an objective reason, and promoting an objective end" (1986, 96–97) and the sense in which "the requirements of objectivity can be regarded as a condition on *whatever* values one holds" (97).

Given the widely accepted view that moral mental states involve both some standards of correctness and some standards of success in action, let us agree that

---

[32] I leave it open here whether there is one ideal world or many ideal worlds, equal in respect of their idealness but varying such that they are not identical.

C2 Moral mental states have both mind-to-world and world-to-mind satisfaction conditions.

The challenge for moral realists in Chapter 3 will be showing how mind-independent moral facts and properties will enter into cognition in a way that motivates moral action. The challenge for constructivists in Chapter 4 and expressivists in Chapter 5 will be showing that their theories can give some account of the mind-to-world satisfaction conditions, including standards of correctness and confirmation for moral beliefs. Having settled these issues about the intentionality of moral mental states and their mind-to-world satisfaction conditions, let us now turn to their volitional aspects and world-to-mind satisfaction conditions.

2.3. Moral Mental States as Volitional States

Moral mental states, more than any other kind of intentional states, are connected to non-verbal behavior. In the last section, we saw that moral mental states may have both mind-to-world and world-to-mind satisfaction conditions. The latter are closely connected to the volitional aspect of moral mental states, which involves bringing about certain states of affairs in the world, including changes in other peoples' attitudes. It is worth pausing for a moment to consider whether moral mental states could be purely intentional states (as opposed to being intentional–volitional hybrids as I claimed earlier). Since we have already dismissed the possibilities that moral mental states are, necessarily and in part, sensations or

qualitative states, we need only consider the connection between intentional states and volitional states with reference to moral mental states.

We can examine this connection by again considering the question, What is the connection between having a moral mental states and trying to act in some way? In what follows, I will take 'volition' to mean simply a kind of trying and 'volitional state' to denote the conceptual episode that *ceteris paribus* (under favorable conditions) leads to a person's performing a certain action. On this treatment, motivation will amount to a volitional state—motivation is, after all, a propensity to act in certain ways—and the connection between moral mental states and their (associated) volitional states could be captured in several different ways:

T1 Moral mental states are always, in part, volitional states.

T2 Moral mental states are sometimes, in part, volitional states.

T3 Moral mental states always co-occur with volitional states.

T4 Moral mental states sometimes co-occur with volitional states.

Since thesis (T1) is the one I wish to defend, I will focus here on eliminating (T2)–(T4) as possibilities, beginning with the question of whether a hybrid model (represented in (T1) and (T2)) or a co-occurrence model (represented in (T3) and (T4)) is more plausible.

On a co-occurrence model, the satisfaction conditions for moral mental states will not turn out as they should. If we separate the moral mental state from its attendant volitional state, the *volitional* state will have satisfaction conditions that involve actually bringing about some change in the world. The state will be

satisfied if and to the extent that a certain situation is brought about. The moral mental state, on the other hand, is *ex hypothesi* purely intentional and may have certain satisfaction conditions involving correctness—being right about, say, a certain action having the moral property of wrongness (although this is a point of contention among different metaethical theories). But the moral mental state will not have any success conditions involving change in the world; those will belong entirely to the co-occurrant volitional state. This individuation of the states clearly conflicts with the demand that moral thinkers are properly criticizable if they are not also *doers* in some sense. Again, it is not enough that I simply hold a moral belief or make a certain judgment; we expect, moreover, that people *act* on their moral beliefs and judgments. This clearly signals that we take certain conditions involving action to be part of the satisfaction conditions of the moral mental state itself, not a separate volitional state. (How strange it would be to say that *I* (or should I say my brain?) failed to produce the proper volitional state!)

Before dismissing the co-occurrence model, it is worth noting that one could argue that moral mental states do have satisfaction conditions involving action, but only insofar as they are responsible for producing *other* volitional states that lead to actions. In other words, one might say moral mental states are responsible for motivating us by producing the right volitional states. Here, the satisfaction conditions for the pure volitional state would involve success in action, while the satisfaction conditions for the moral mental state would turn on whether it, indeed, produced that volitional state or not.

This account is unattractive for several reasons. First, whether a mental state

cascades to cause other mental states is largely an involuntary matter. If we recognize a split between moral mental states and volitional states leading to actions, a person would be properly criticizable if his/her moral mental states did not produce the right volitional states (i.e., if his/her moral mental states didn't cause the person to form intentions to do something). But what would responsibility amount to here in the absence voluntary control? Perhaps we could say that if someone didn't *want* to form the intentions, he/she would be at fault. But this seems to ignore the difficult and sometimes undesirable demands of morality. We don't usually fault agents for not wanting to be moral; after all, they can have natural desires and inclinations that are quite opposed to what is moral. We fault them for not *acting* morally *in spite of* their other desires, inclinations, and preferences. But that is just to say we fault them for not acting on their moral mental states. To interpose another volitional state in the process and try to split the difference in terms of satisfaction conditions is to have "one state to many," to adapt Bernard William's (1981, 18) famous phrase.[33]

Second and more importantly, it would be nearly impossible for most observers to tell whether a moral mental state had succeeded or failed on this account because the satisfaction conditions of the moral mental state turn on its causing another mental state. We can, of course, assess this second volitional state based on whether the agent succeeds in his/her action. But failure of action does not imply that the second volitional state was absent and that the moral mental

---

[33] It should be noted, of course, that Williams marshaled this phrase against motivational externalism, the view that motivation lies beyond an agent's own commitments, wants, goals, etc.—in short, beyond the agent's own mental states.

state similarly failed. There are a number of complications (e.g., bad luck) that can arise in action that would prevent my intention from succeeding. This account requires simply that the moral mental state produce the right volitional state, and there is very little evidence publicly available on whether I have the right volitional state or not. Trying and failing because of bad luck may be extensionally equivalent to not trying at all. Absent advanced neuroscience (and, presumably, real-time monitoring of people's minds), we have no way of knowing whether the moral mental state has produced the second volitional state. This would have disastrous effects for assessment of other agents and the important roles that criticism, shaming, and gossip seem to have for social stability, training others in cultural norms, and circulating information about other peoples' reputations and trustworthiness.[34] For practical reasons of normative assessment, it makes sense to assess agents directly by keeping success in action as part of the satisfaction conditions for moral mental states themselves, while making appropriate exceptions for moral luck.

Having dismissed the co-occurrence model, we can now consider thesis (T2), the view that moral mental states are hybrid states, but only sometimes. This weak condition holds for an important range of non-moral mental states. My belief that it's raining or that the ice is thin or that I need groceries seems to incline me toward action (e.g., bringing an umbrella, taking care where to walk, going to the store),

---

[34] For background on the special role of gossip in human evolution and psychology, see Dunbar (1996); Kniffin and Wilson (2005); Mesoudi, Whiten and Dunbar (2006); Hess and Hagen (2006); Sommerfield, Krambeck and Semmann (2007), Sommerfield, Krambeck and Milinski (2008); O'Gorman, Wilson and Sheldon (2008); and Scheuring (2010).

and it is arguably the case that these states have satisfaction conditions that involve doing something in the world, whether that be performing some kind of non-verbal behavior or even, say, issuing a verbal warning to someone about to step out onto the frozen lake.

The difficulty of determining the satisfaction conditions in non-moral cases is, in my view, derived from an epistemological problem involving performance conditions of the behavior in question. When we see someone perform a certain piece of non-linguistic behavior, we cannot always (or perhaps, even very regularly) infer that the person had a particular thought (i.e., that a certain thought was a prerequisite for performance of that behavior). I may take an umbrella because I think it's going to rain, but I may just as well take one to avoid the sun or because I want to return it to someone or because I just like the look of it. Any one of these actions can be occasioned by several different volitional states, and lacking any one-to-one correspondence, we cannot conclude on the basis of such an action that a person indeed has such and such volitional state. (Cases of verbal behavior are quite different; as long as a person is being sincere, we can pretty well tell what he/she is thinking.) Since we cannot always know what a person was thinking in performing some non-linguistic behavior, our standards of evaluation for states that sometimes incline behavior and sometimes do not are correspondingly indeterminate and, hence, there is no consensus on whether the satisfaction conditions for certain non-moral mental states indeed turn on acting in some way.

By contrast, we *do* have clear expectations where moral mental states are concerned. Moral beliefs *do* typically and are expected to occasion a wide range of

nonverbal behaviors, including activism, charity, intervention, advice giving, and so on. At the very least, we expect that a person who has moral beliefs that forbid torture actually refrain from administering torture. We may expect, further, that he/she try to bring it about that no torturing occurs—either perhaps by persuading others not to torture or by directly intervening in situations to prevent torture from occurring. Anyone who professed to have a belief against torture and failed to do any of the following would surely be seen as failing in some respect, or perhaps even as not genuinely holding that belief. This condition holds not just for some moral mental states but, indeed, for *all* of them.

If the condition did not hold, there would be no puzzle concerning cases of akrasia, or weakness of will, in which an agent acts contrary to his/her own better (moral) judgment. If thesis (T2) were true, it would be possible to interpret the akrasiac as having token moral mental states with satisfaction conditions that he/she failed to satisfy through action.[35] But the true puzzle of akrasia is that for *any given* moral mental state the akrasiac has (but not necessarily *all*), he/she fails to act, which runs counter to the satisfaction conditions for that moral mental state—*any* moral mental state. Let us conclude, then, by rejecting (b) and treating the foregoing discussion as further support for (a), the thesis that all moral mental states are, in part, volitional states with their attendant satisfaction conditions.

---

[35] Indeed, one wonders how to figure out *which* moral mental states have those satisfaction conditions on (b). Is it in virtue of particular contents with states about, say, torture having those conditions but thoughts about benevolence lacking them? Is it how strongly the agent endorses purported moral mental state in question? Are the satisfaction conditions tied to contents or cognizers, with all their variations? Could my thoughts about torture have those conditions and yet yours lack them? How would we determine, for any particular agent, whether his/her token moral mental state had those conditions? I do not see much hope for a systematic account along the lines of (b).

In the next two sections, I want to clarify two issues involved with the volitional aspect of moral mental states. In the first of these sections, I consider, in detail, the world-to-mind satisfaction conditions of moral mental states, including what counts as success in action. In the second section, I return to the demarcation problem and attempt to mark out the domain of moral mental states, insofar as is possible on a functionalist account.

2.3.1. Directions of Fit II: World-to-Mind Satisfaction Conditions

In §2.2.2, I argued that moral mental states have world-to-mind satisfaction conditions that involve changing the world in certain ways. It is important to note that these satisfaction conditions may be quite broad: In some cases, I may be responsible for acting directly to bring about (or prevent) a situation. Examples here may include both positive duties (e.g., giving to charity, helping strangers in need) and negative duties (e.g., refraining from acts of violence, tolerating others' lifestyles). In other cases, I may be responsible for changing the views of others and thereby influencing events in some ways. Consider here the important role of activism in bringing to light situations of injustice and effecting change at the most structural levels of society. In this respect, it might even turn out that in some cases, mere verbal behavior is sufficient for satisfying the moral mental state in question—especially if the task is so large that it requires group action to be accomplished. The point is that there may be varying levels of demand on moral cognizers in terms of the world-to-mind satisfaction conditions for their moral mental states, and there can also be varying levels of success.

There are, no doubt, cases in which moral mental states, as a matter of fact, do not incline people toward action. So rather than offering a description of all moral mental states here, we are discussing, to a certain extent, what moral mental states *should* be, or rather what they should do. But to say they should be something is just to say that there are satisfaction conditions about the state that can be fulfilled or not fulfilled. We can acknowledge that someone has a defective moral mental state in respect of his/her having no attendant behaviors. What makes the state defective is precisely the fact that part of its satisfaction conditions are unfulfilled. Note also that these conditions can be satisfied to varying degrees. I can successfully *intend* to put my hand up in the sense that the mental precursors of my hand's raising are there and, absent paralysis, amputation, or some physical or chemical restraint, my hand would go up. Searle gives an example here drawn from Williams James: "a patient with an anesthetized arm is ordered to raise it. The patient's eyes are closed and unknown to him his arm is held to prevent it from moving….[W]e can say of the patient that his experience is one of *trying* but *failing* to raise his arm. And the conditions of satisfaction are determined by the experience; he knows what he is trying to do and he is surprised to discover that he has not succeeded" (1983, 89). Notice, further, that the matter is not even as simple as my arm raising or not, since my arm can go up to varying degrees: a tremor, a halfhearted flail, a proud vote. We need not preoccupy ourselves here with the exact degree to which one's hand must raise for us to say that one has succeeded. The point is that the satisfaction conditions—in this case, what one would have to do to fully raise one's hand and the descensions in degree

downward from that point—*are* clear. The varying degrees to which satisfaction conditions may be fulfilled in the case of moral mental states raises interesting questions of moral praise and punishment.

As we saw earlier, merely having a moral mental state and even attempting to act upon it may not be sufficient to fully meet the world-to-mind satisfaction conditions for that state. Despite my best attempts to help others, a variety of factors can intervene (e.g., restraint, incapacity, disability, lack of sufficient skill or knowledge, bad resultant luck). The sufficient level for success here seems to vary according to different normative theories. Act utilitarians, for example, will take nothing less than successful results and will measure success in proportion to the amount of utility produced by the attempted action. Rule utilitarians might be a little more lenient here, allowing that partially or even completely unsuccessful attempts at action are still important for producing a good society overall. Kant and his followers, importantly, are impressed by agents' intentions or their moral motives, noting that results can often go awry:

> Even if, by a special disfavor of fortune or by the niggardly provision of a stepmotherly nature, this will should wholly lack the capacity to carry out its purpose—if with its greatest efforts it should yet achieve nothing and only the good will were left (not, of course, as a mere wish but as the summoning of all means insofar as they are in our control)—then, like a jewel, it would still shine by itself, as something that has its full worth in itself. Usefulness or fruitlessness can neither add anything to this worth nor take anything away from it. ([1785] 2002, 4:394)

Intuitively, what is wanted in specifying the world-to-mind satisfaction conditions for moral mental states is some kind of handicap for moral luck.

Nagel says moral luck occurs "where a significant aspect of what someone does depends on factors beyond his control, yet we continue to treat him in that respect as an object of moral judgment" (1979, 26). Nagel outlines four types of moral luck:

- resultant moral luck ("luck in the way one's actions and problems turn out"),
- circumstantial moral luck ("the kinds of problems and situations one faces"),
- constitutive moral luck ("the kinds of person you are, where this is not just a question of what you deliberately do, but your inclinations, capacities, and temperament"), and
- causal moral luck ("luck in how one is determined by antecedent circumstances"),

which might bear on the world-to-mind satisfaction conditions of moral mental states (28). Resultant moral luck *directly* impedes—in the course of action—one's success in achieving these satisfaction conditions, while constitutive and causal moral luck seem to *antecedently* limit one's prospects for meeting those conditions (e.g., genetic or congenital disability, predilection for erratic or risky behavior, mental disorder). Circumstantial luck seems least related to these conditions but could possibly put one in a situation in which one acts in some way other than how one would have acted.

The phenomenon of moral luck is a complicated one that bears on our assessment of whether an agent has met the world-to-mind satisfaction conditions of his/her moral mental states. This question, however, seems to be normative in nature, and I raise the problem of moral luck here only to highlight the puzzles involved in world-to-mind satisfaction conditions. The point remains that those conditions belong to all moral mental states however they are cashed out by different theories.

2.3.2. Practical Reason and the Consistency of Volitions

On our current functional account, moral mental states, as a species of hybrid intentional–volitional states, are in the same category as emotions and other mental states with intentional and volitional elements and their corresponding mind-to-world and world-to-mind satisfaction conditions. What we need to consider here is what, in addition to these conditions, serves to distinguish moral mental states from all other mental states. Let us take up consideration of the emotions and the respects in which, functionally, they differ from moral mental states.

Robert M. Gordon presents a very useful analysis of the intentionality of emotions in saying, "the intentionality of emotional sentences is derivative from the intentionality of sentences that ascribe beliefs and wishes" (1973, 36). On his view, to say that someone is angry is to say that he/she is angry *about* something, which plays a significant role in predicting how a person will act and planning how to act toward them: "With no idea what she is angry about, we can only wait and

see what her anger will motivate her to do. Nor do we know what if anything should be done to 'make amends,' and who should do it" (1987, 23). Gordon's treatment diverges from many accounts on which emotions are purely volitional states. He cites as an example Robert C. Solomon's view that "Emotions are judgments and actions, not occurrences or happenings that we suffer. Accordingly, I want to say that emotions are choices and our responsibility" (Solomon 1973, 40). In drawing crucial attention to the intentional features of emotions, Gordon points to the ways in which the mind-to-world satisfaction conditions bear on whether one is in an emotional state or not. In cases of error, the emotion often dissipates: once I learn that you did not steal my wallet, I am no longer angry at *you*.

But what about cases of success in which the mind-to-world satisfaction conditions of the emotional state *are* met and, say, you indeed stole my wallet? The volitions that may be part of this emotional state include my trying to get it back, seeking help from the law, being prepared to use physical force, purchasing weapons to use in this regard, attempting to get others to attack you for it, and perhaps even plotting some kind of revenge. Many of these and other possible actions are not *moral*, and this difference is precisely the point I am attempting to draw out here. Emotions incline us to various actions that moral mental states would never allow. Morally speaking, it may be the case that you owe me something (e.g., my wallet, interest, additional compensation) and perhaps I even have a right or a legitimate claim against you for such things. But, in the case of restoring the balance sheet, as it were, I can't go so far as to violate more stringent rights of yours (e.g., your right to be free from bodily harm), cause you unnecessary

pain and suffering, and so on. What explains the difference between these two sets of volitions—those engendered by the emotions and those belonging to moral mental states?

Undoubtedly, someone will say that the difference between the two is that the former case of just the emotion is self-regarding, while the latter, the moral case is other-regarding. My emotions concerning the stolen wallet concern only my own self-interest, while the moral mental state forces me to take account of others' interests as well. (As stated, this hypothesis is oversimplified, since I can have an emotional state when someone steals *your* wallet, especially if you are close to me. Still, the unrefined example will serve our purposes here.) While this description of the difference may be quite correct, it draws us back to a substantive account where moral mental states are demarcated with reference to their contents—here, that they concern the interests of others and not just those of the person who has the moral mental state. For reasons already discussed, it would be preferable to find a functional account that explains this gulf between moral mental states and emotional states.

With that in mind, let us return to the notion that functional classifications of moral mental states specify the role that they play in one's overall mental economy. On this view, moral mental states will be among various other mental states one has in the course of cognition. In particular, the volitional state that a person has in virtue of having a moral mental state will be among other volitional states he/she has, including wants, desires, preferences, intentions, goals, emotional states, and other moral mental states. One respect in which emotional states and

moral mental states differ is that the former seem insensitive to one's other volitional states, while the latter take some stock of them. If I am angry about your stealing my wallet, my rage may overtake me and cause me to do any number of things, some of which may frustrate my others plans and wishes, such as keeping an appointment or being a non-violent person. A moral mental state, by contrast, would somehow take account of these other volitional states and constitute a judgment, belief, attitude, etc. that I should do certain things but refrain from others. The process of moral thought and reflection involves some kind of practical reason across one's other volitional states. As part of this planning, I may come to learn of new preferences, desires, etc. of which I was unaware, discount some of them upon further analysis, elevate others, and reach a prudent judgment, on the whole, of what is best to do.

One may perform better or worse in this process of practical reasoning. I may be particularly adept at understanding and ranking my volitional states, as it were, and arriving at moral mental states. On the other hand, I may be presented with particularly difficult sets of volitional states to reconcile or I may exhibit some flaw in practical reasoning that makes this task extremely difficult. I may not succeed on every occasion, but it should be possible, in principle, to say whether a person has done better or worse in reaching a moral mental state that is consistent with his/her other volitional states.

Let us add to our list of propositions the following:

C3 Part of the world-to-mind satisfaction conditions for moral mental states involve consistency with one's other volitional states (that are

not ruled out by moral mental states).[36]

Total volitional state consistency will be among the world-to-mind satisfaction conditions for moral mental states, along with actual success in action, because acting on an inconsistent set of volitional states is likely to result in failed attempts and frustrated long-term plans. If I have a moral mental state that involves acting charitably toward others and yet I allow my feelings of resentment toward some and envy of others to guide my actions (even some of the time), my behavior overall will tend in opposite directions: I may give the money but feel self-loathing as a result. Or I may act stingily but feel guilt. Or I may just be so internally divided that I give some money—not enough to satisfy the demands of charity but too much to keep my miserly attitudes at bay. Indeed, divided total volitional states are tragic phenomena.

What of the opposite? Suppose a person is lucky and his/her particular set of volitional states is inconsistent, but the results somehow turn out just fine. It is rather hard to imagine such cases, and one doubts that they occur very frequently. Nevertheless, I am inclined to say that we need not be concerned about this case of good luck. Returning to our discussion of moral luck, nearly all correctives concern bad luck—take, for example, the moral agenda set by Rawls' veil of ignorance, which is designed to rule out the possibility of discrimination against the *worst* off. It is unclear, normatively, that we need to be concerned about robbing people of

---

[36] I add the parenthetical above because in some cases the content of certain moral mental states will rule out one's consistently holding some other non-moral mental state. I take the moral mental state to "trump" the non-moral mental state—strictly speaking, I take the satisfaction conditions of the moral mental state to require that one not adopt the non-moral mental state on pain of inconsistency and failure in meeting those satisfaction conditions.

their *good* fortune—unless, of course, that is necessary in the course of compensating for other's bad luck (e.g., redistributive taxation).

It is a complicated question which of my volitional states I should retain and which I should reject as a matter of resolving the total set.[37] It is always possible for a rogue volitional state to come upon a person, even after he/she attempts to extinguish such states in light of a moral mental state he/she endorses. It seems that such cases should not count against the success of the moral mental state in achieving the right fit, especially because the agent has ruled it out in the process of practical reasoning. Recalcitrant states, however, may signal a strong preference for something (e.g., a former lover), and failure to take account of this deep attraction *would* count against the agent's success in determining the right thing to do. In addition, there are interesting questions about whether short-term or long-term goals are to be generally preferred as part of this process and which among each is better. I leave these as areas for further inquiry.

Humans are similar enough that there will be some common core of volitional states that we all share.[38] To this extent, the exercise of practical reasoning *for us* will be roughly similar, even though our particular sets of volitional states will exhibit variation based on different preferences, plans, and so on. This common core is sufficient to get the normative project off the ground, and specific metaethical views about the confirmation of moral mental states—that is,

---

[37] For further discussion of this issue, see Morrow (2009).

[38] Sen and Nussbaum's capabilities approach is highly useful in this regard. Nussbaum includes as basic capabilities: life; bodily health; bodily integrity; senses, imagination, thought; emotions; practical reason; affiliation; other species; play; and control over one's environment (1999, 41–42).

correctness in their mind-to-world satisfaction conditions—will further constrain what moral advice normative theories can offer. In Chapter 6, I will present my own account, which includes, notably, the empirical thesis that, in enough cases, self-regarding and other-regarding interests coincide. In my view, the conceptual divide between self-regarding actions and other-regarding actions is a false dichotomy; many moral actions and their attendant moral mental states lie in overlapping territory. In any case, other metaethical frameworks add further constraints to this functional view of moral mental states I have developed here.

2.4. The Methods of Metaethics

Before concluding these introductory remarks, I want to preview the metaethical views in focus in this dissertation as well as the problems they face with respect to moral mental states. I think it is helpful to begin by sketching these metaethical methodologies in the broadest possible outlines.[39] These will, of course, correspond to prominent theories in the current literature: moral realism, constructivism, and expressivism. Discussing them more broadly, however, will bring these accounts into full contrast with each other and make clear the larger frameworks available for understanding moral mental states and their connection to extra-mental reality. In my view, none of these frameworks is yet adequate for addressing the full range of metaethical questions that can be raised, but

[39] In describing these views, I employ the terms 'externalism' and 'internalism'. While these views have rather obvious connections to epistemic externalism and internalism, I stress that metaethical *methodologies* are under consideration here—methodologies that touch on metaphysical, semantic, psychological *and* epistemic questions—not simply questions about moral knowledge and justification.

surveying them and their related problems does help in constructing an alternative account presented in Chapter 6 of this dissertation.

2.4.1. Externalism

The first metaethical methodology to be examined may be called metaethical methodological externalism ("externalism" or "externalist" for short) because it attempts to answer metaethical questions by looking to the external world: Do moral facts and properties exist? If so, what kind of facts and properties are they? How do they enter into the thoughts and volitions of moral agents? Does recognition of them entail that subjects are disposed or motivated to act rightly? If so, what is the psychological explanation of this process? And so on. On the externalist view, metaethics is akin to science, since the confirmation of normative claims depends on proper observations about what exists—morally speaking—and so, this view is most commonly found in realist accounts on which moral facts and properties are said to exist stance-independently.[40] As Richard Boyd says, "Many philosophers would like to explore the possibility that scientific beliefs and moral beliefs are not so differently situated….We may think of this task as the search for a conception of 'unified knowledge' which will bring scientific and moral knowledge together within the same analytical framework in much the same way as the positivists' conception of 'unified science' sought to provide an integrated treatment of knowledge within the various special sciences" (1988, 183–84). The major prospect for externalism seems to be its potential for ruling out certain

[40] For reasons I discuss in §3.1, 'stance-independence' is a better term than the more common 'mind-independence'.

classes of normative theories according to what kind of moral facts and properties they countenance. If, for example, externalists could determine that 'good' referred to property *A* and not property *B*, that would instruct us, as moral agents, to promote and protect things with *A* and possibly to discourage and destroy things bearing *B*. Similarly, knowing something about *A* might tell us, as agents, when and where and how to detect it, and that, too, would give us some practical guidance about what to do.

The major challenge for externalists is showing that there are, indeed, procedures by which we can detect stance-independent moral facts and properties. Given the phenomenon of intentional inexistence, there is no guarantee that our thoughts about morality represent real-world moral facts and properties, if they exist at all. Moreover, it is not clear that externalist accounts preserve the directly action guiding nature of moral mental states. Merely *discovering* that a particular (class of) action(s) has the property of rightness or wrongness does not seem to entail that I am motivated to act in any particular way, any more than discovering facts about a particular wavelength of light or the chemical properties of a particular substance. At the very least, the externalist owes us a robust story of how discovery of moral facts and properties is connected to action. These and other difficulties for externalism will be explored in Chapter 3.

2.4.2. Internalism

The second approach to metaethical questions may be called metaethical methodological internalism ("internalism" or "internalist" for short) because it

attempts to answer metaethical questions by looking at the content of moral judgments and extracting normative principles from them. On this view, the very structure of moral reasoning—that is, the perspective one takes up in asking moral questions—contains within it the grounds necessary for answering moral questions. For internalists, these answers are not a matter of discovery or observation but rather emerge from reflection on one's experience from the moral point of view. Internalism is most commonly found in constructivist accounts according to which moral agents develop principles of practical reason through the moral concepts of rights and persons that engage moral questions in the first place. The ontological commitments of constructivism are a complicated matter and will be discussed in some depth in Chapter 4. For the present, it is enough to point out that moral truths are stance-dependent for constructivists in the sense that they require the existence of moral agents equipped with basic moral concepts who undertake the process of constructing these facts and properties in the very process of moral reasoning. As Christine Korsgaard puts it: "our use of the [moral] concept when guided by the correct conception *constructs* an essentially human reality—the just society, the Kingdom of Ends—that solves the problem from which the concept springs. The truths that result describe that constructed reality" (2003b, 117).

Internalist accounts are notable for their close connection to moral experience and moral practice. Constructivists view the Kantian tradition as addressing both normative and metaethical questions and in a way that collapses the two, as it were. Sharon Street claims, "the traditional metaethical project of

reconciling normative discourse with a naturalistic worldview turns out to be a substantive one" for which the constructivist is well equipped (2010, 380). By undertaking the moral point of view and reflecting on the nature of right actions, we in turn construct the framework needed for addressing questions at the metaethical level. This unified approach to metaethical and normative pursuits is appealing for many, and David Solomon (2005) and others hail Rawls' constructivism in *A Theory of Justice* as a "sea change" (352) that rescued ethics from conceptual analysis and brought about a revival of normative questions, along with the advent of applied ethics.

The chief difficulty for internalists has been achieving objectivity in their account of normative judgments—in effect, making sense of mind-to-world satisfaction conditions. There is robust evidence, both anecdotal and anthropological, that different moral observers sometimes have different moral reactions to the same situation. Since internalists address metaethical questions from the standpoint of individual experience, they must somehow account for this difference and, along with the externalist, provide criteria for separating reasonable or true judgments from faulty or biased ones. For *even if* all observers at a given moment actually converged on the same moral judgment, it's simply not obvious that that judgment would count as reasonable or correct—consider the history of American culture, which has been largely racist, sexist, and homophobic.

Both worries could be addressed by giving a robust account of the mind-to-world satisfaction conditions for moral judgments. Solutions to these problems have largely been pursued along Kantian lines, but, as we will see in Chapter 4,

there are significant problems with this approach. As I will argue, it's even possible that constructivism collapses into a kind of externalism and suffers from some of the same problems as moral realism.

Thus characterized, externalism and internalism clearly seem to be competing theories. Realism and constructivism, however, have not always been regarded as such in the literature, mainly because they are seen as attempting to answer *different* metaethical questions. Whereas realists, following Moore, have often focused on the nature of moral language, particularly its referential and truth-conditional aspects, constructivists have pursued conformational questions about normative claims, particularly whether such claims can be objectively verified. In my view, these linguistic and epistemic pursuits are ultimately aimed at the same goal of informing downstream normative theories, and, as we have just seen, their methodologies are indeed opposed; we may safely treat them as direct competitors in the metaethical field.[41] Nonetheless, they do not exhaust the full spectrum of past and present theories.

2.4.3. Performativism

A third alternative, which has been varied but persistent in the literature, may be called metaethical methodological performativism ("performativism" or "performativist" for short) since it holds that normative statements, rather than referring to moral facts and properties or reflecting objective moral reality, instead express attitudes of approval or disapproval on the part of speakers and attempt

[41] Indeed, it is worth noting that realist-constructivist accounts are extremely rare, if they have been offered at all. For a possible example, see Ross (2004).

to persuade others to act accordingly. This view has been found among emotivists, prescriptivists, and expressivists in the twentieth century. These three classes of performativism have differed slightly in their semantic treatment of moral statements: for emotivists, moral statements have primarily dynamic use, aimed at influencing actions and bringing about like responses in others; for prescriptivists, moral statements issue imperatives that are binding on all those in similar circumstances; and for expressivists, moral statements express an agent's approval or disapproval of actions against a background system of norms. These accounts seem to follow Austin's discussion of nondescriptive speech acts, on which "the uttering of the sentence is, or is part of, the doing of the action, which…would not *normally* be described as, or as 'just,' saying something" (5). Performativists regard all moral statements as nondescriptive speech acts, where the primary function of the speech act is to express the speaker's attitudes or to bring about some change in the world, rather than to describe any moral reality.

For these reasons, nearly all performativists reject the ontological commitments of moral realism, and almost none straightforwardly embraces the objective or Kantian/transcendental views of constructivists.[42] More broadly, they resist the aims of externalists to describe stance-independent moral facts and properties, and they resist the ambitions of internalists to build an objective account of moral judgment, starting from the standpoint of the individual's experience in making those judgments. For these reasons, performativists are hard-

[42] For examples of these deviations, see Horgan and Timmons (2006) and Gibbard (2003).

pressed to account for the mind-to-world satisfaction conditions for moral mental states—that is, to account for the sense in which some moral mental states may be better than others.

This problem has most often been expressed as a semantic point. If moral statements do not describe anything in the real world or ideal worlds, they are not, strictly speaking, truth-conditional. My statement, "Murder is wrong" is equivalent to a command, "Don't murder," which can be obeyed or disobeyed but cannot be true or false. This analysis of moral language makes it hard to understand how non-truth-conditional moral statements would function in logical arguments or how they would impact the overall semantics of statements such as "I am a philosopher and murder is wrong," contain truth-conditional parts and non-truth-conditional parts.[43] Moreover, if moral statements have no truth-conditions—or, to put the point a bit differently, if there is no objective standard for moral judgments—it is unclear whether there is any point in trying to resolve moral disagreements or why we should criticize those whose views seems narrow, biased, or just plain wrong, or, indeed, whether it is even possible rationally to scrutinize the moral judgments of others. Performativists, particularly expressivists, have attempted a variety of responses to these charges while still trying to avoid ontological commitments to stance-independent moral facts and properties. As I will argue in Chapter 5, contemporary quasi-realist accounts fail in this respect collapse into a kind of externalist account.

---

[43] This is essentially the Frege-Geach point, which demands that the meaning of terms (and their truth-conditions) remains the same across various instances, including in formal arguments and embedded clauses. For further discussion of this point, see §§3.2.3, 5.2, and 5.3.

2.5. Conclusion

In this chapter, I attempted to construct a working notion of moral mental states and to establish criteria according to which existing metaethical frameworks—externalism, internalism, and performativism—may be assessed. I argued that a functional account of moral mental states, which took notice of the special role that such states play in our overall mental economy, was most appropriate. I rejected substantive approaches for their question-begging exclusion of certain judgments and statements that appear to reflect moral mental states, and I criticized mongrel accounts for failing to take seriously the possibility that disparate moral mental states could be systematized.

On my own functional account, moral mental states are intentional–volitional hybrid states with corresponding mind-to-world and world-to-mind satisfaction conditions. The former reflect the phenomenon of intentional inexistence and demand an explanation of why some uses/applications of moral concepts/terms are appropriate; the latter reflect the directly-action-guiding aspect of moral mental states and demand both consistency with one's other volitional states and an explanation of how moral mental states lead to success in action, making suitable discounts for moral luck.

With this framework in place, we can now examine in greater detail each of the metaethical methodologies I described in this chapter. In each case, I will argue that externalist, internalist, or performativist accounts fail to adequately address the constraints on moral mental states established in this chapter. Following that, I will present my own interactionist account in Chapter 6, which incorporates some

of the strongest elements of these other views while still fulfilling the established criteria.

Chapter 3

Metaethical Externalism, or Morals by Observation

> I have endeavoured to write "Prolegomena to any future of Ethics that can possibly pretend to be scientific."
>
> —G. E. Moore, *Principia Ethica*

Metaethical methodological externalism draws heavily from science in its approach to moral experience. Externalists believe there is a stance-independent moral reality out there to be discovered; we simply need to attend carefully to aspects of our experience to appreciate and systematize it. The tools and methods of the ethicist's investigation of ideal states of affairs will naturally differ from those of the scientist's investigation of the physical world, but the approach to the two areas of inquiry is essentially the same. Richard Boyd claims that "moral beliefs and methods are much more like our current conception of scientific beliefs and methods (more "objective," "external," "empirical," "intersubjective," for example) than we now think" (1988, 184). On his view, the case for scientific realism underscores the powerful attraction to what he calls the "unity of science" (217) (what I here call "externalism"), and this, in turn, cries out for application to the moral realm. In the scientific case, this method is compelling because

> of the extraordinary role which theoretical considerations play in actual (and patently successful) scientific practice. To take the most striking example, scientists routinely modify or extend operational "measurement" or "detection" procedures for "theoretical" magnitudes or entities on the basis of new theoretical developments. This sort of methodology is perfectly explicable on the realist assumption that the operational procedures in question really are procedures for the measurement or detection of unobservable entities and that the relevant theoretical developments reflect increasingly accurate knowledge of such "theoretical" entities. Accounts of the revisability of operational procedures which are compatible with a non-realist position appear inadequate to explain the way in which theory-dependent revisions of "measurement" and "detection" procedures make a positive methodological contribution to the progress of science. (188)

Correlatively, Boyd expresses his optimism that in the moral realm, we can achieve success in discovering a set of true moral beliefs—*greater* success, in fact, than in the scientific realm. As he notes, related advances in history and economic theory cannot help but add to our moral knowledge (205), and we seem to have evolved robust and reliable psychological and cognitive mechanisms that aim at the good: "Locke was right that we are fitted by nature for moral knowledge (in both seventeenth- and twentieth-century senses of the term) in a way that we are not so fitted for scientific knowledge of other sorts" (209).

While many externalists share Boyd's enthusiasm about moral progress, they offer more qualified accounts of their methodology. Peter Railton (1989) quotes Nagel in arguing that "it begs the question to assume that this [scientific] sort of explanatory necessity is the test of reality for values" (Nagel 1986, 144). Railton himself claims that rightness is a matter of "what is instrumentally rational from a social point of view," where instrumental rationality plays the explanatory role in our theories of social interaction that scientific explanations play elsewhere (1986, 200). Russ Shafer-Landau takes an even more guarded stance. "Morality," he argues, "is essentially a matter of regulating and assessing the activities of *agents*" (2003, 102). As a result, it considers those agents' mental states of planning, intending, and so forth, and "it is questionable whether an ontological standard that requires a showing of explanatory efficacy quite independently of such attitudes is really as neutral as it purports to be." Though these externalists may disagree that moral beliefs and explanations must pass the *same* tests as scientific ones, they do not abandon the general framework of externalism. Shafer-Landau again: "When it comes to ideal epistemic agents…the important point is that their impeccable accuracy reflects an acquaintance with a truth not of their own making. Their perfect knowledge is not constitutive of moral facts, but rather reflects an awareness of what is there, awaiting their discovery" (17).

In some cases, externalists are not so optimistic. In fact, some externalists deny the existence of moral reality altogether. John Mackie gives perhaps the strongest statement of error theory in saying, "although most people in making moral judgments implicitly claim, among other things, to be pointing to something

objectively prescriptive, these claims are all false" (1977, 35). For Mackie, both the cultural variation found in moral codes and the "queerness" of what would have to exist in order to make moral claims true show that objective moral beliefs couldn't *possibly* be satisfied. Nevertheless, like all externalists, he takes them to be aimed at something in the world prior to human thought and judgment but, finding no such entities, draws a skeptical conclusion where other theorists find hope.

I will not dwell on the merits of error theory here. I have more positive views of moral mental states, which I present in Chapter 6. If those arguments succeed, or perhaps if they are acceptable to Mackie (who is, in fact, a relativist himself), so much the worse for hard-line error theory! The main task of this chapter will be examining moral realist accounts on which observers discover moral facts and properties that exist and have their natures independently of our intentional stances toward them. I begin with a brief overview of realists' accounts of moral reality in §3.1. This survey will be necessarily varied, since many such accounts exist, but I will try to draw out the common themes of realism in the process, especially with reference to externalism. Following that, §3.2 examines three common arguments for realism in light of the notion of moral mental states I developed in Chapter 2. I conclude that none of them succeeds in establishing stance-independent moral facts and properties, but in §3.3, I consider what would follow if they did. In §3.3.1, I explore the notion of observation contained in realist accounts; I offer no critical remarks there, merely an account of what such access would look like and how different accounts of access explain the difference between naturalistic and non-naturalistic realisms. In §3.3.2, I consider the links in the

opposite direction by taking up the case for motivation on realist accounts. I argue there that realism has little chance of presenting a plausible view of motivation, particularly because of constraints imposed by their accounts of observation. I conclude with several remarks on why I think any externalist will have difficulty in presenting a fully plausible view of moral mental states.

3.1. The Nature of Moral Reality

Moral realists believe there is a class of moral facts that makes our moral statements true (depending on the statements), and this class is constituted by moral properties that exist independently of whether and how we think about them. Put a bit differently: there is a moral reality out there to be discovered. We attend to it, our moral mental states and expressions purport to be about it, and we get things right insofar as our thoughts and language correspond to that reality. Let us, then, take moral realism to consist of the following thesis:

MORAL REALISM There is a moral property *m* picked out by 'good' that is instantiated in each element of a set of possible worlds $\{W_{i-j}\}$ which includes the actual world α. The instantiation of *m* in an object or event constitutes the moral fact F about the goodness of that object or event. And so on for 'right' and all other moral predicates and their corresponding moral properties and facts. When an observer O makes an assertoric moral statement about an object or event, that

> statement is true iff the object or event is *m* or (equivalently) iff F.

A few points here bear explaining.

First, it is crucial that the moral realist have in mind a *set* of possible worlds in which moral properties are instantiated. Not only do we make judgments about the actual world, we often make moral statements about what the world *should* be like or what it would be like *if* something else had happened. These ideal and hypothetical imaginings take us beyond the reach of the actual world to other possible worlds, and the realist must be able to say these moral statements are true (or false). Now it might be the case that this set contains only the actual world. If you're a Leibnitzian and you think this world is as ideal as it could possibly be, that's absolutely fine because, on your view, the set $\{\mathbf{W}_{i-j}\}$ contains at least one member. Even so, you should be able to imagine other possible worlds less ideal in which some moral properties are instantiated and be able to talk about them.

Second, it must be the case that α is a member of this set, otherwise the theory will amount to error theory. Sometimes realists like to talk as if merely establishing truth *conditions* for moral language is enough to demonstrate the truth of moral realism. Geoffrey Sayre-McCord, for example, formulates moral realism as "embracing just these two theses: (1) the claims in question, when literally construed, are literally true or false (cognitivism), and (2) some are literally true….[I]n the account I offer, there is no mention of objectivity or existence, no mention of recognition transcendence or independence, no mention of reference,

bivalence, or correspondence. And this is a virtue" (1988, 5–6). On this position, the statement

E1 "If anyone commits an act of torture, it is wrong"

is true if merely the following fact obtains:

F1 In world $\mathbf{W_1}$, torture instantiates the moral property of wrongness.

As long as torture is wrong in at least one world (and the rightness or wrongness of actions does not vary across the set of worlds in question), my statement will wind up being true. Now consider another statement

E2 "This act of torture is wrong!"

The facts that could make E2 true include

F2 In α, torture instantiates the moral property of wrongness

F3 In $\mathbf{W_1}$ and α, torture instantiates the moral property of wrongness

F4 In $\{\mathbf{W}_{i-j}\}$, torture instantiates the moral property of wrongness

The difference between E1 and E2 is this: Only F2–F4 commit us to the existence of torture and its wrongness in the actual world. One must be committed to at least this much to count as a moral realist because the hypotheticals E1 and F1 are consistent with error theory. One can imagine Mackie, in a sarcastic mood, saying, "I agree with the realist about what it *would* take for moral claims (about this world) to be true, but, unfortunately, they're always false because there are no such things as moral properties in this world." Error theorists and moral realists can agree at the hypothetical level about what ideal worlds look like and what would follow *if* moral properties were ever instantiated. They disagree precisely on the point of *whether* such properties are ever instantiated in the actual world. How dreary it

would be for the realist to tell us that we can indeed imagine possible worlds in which there are moral facts and properties but *we* happen to be situated in a world outside of that set and, in our world, there are no moral properties! Realism is and must be the claim that there is at least this moral fact: on at least one occasion, at least one moral property is instantiated in this world.[44] Of course, this is the minimum requirement for moral realism; one hopes (for the realist's sake) that there are many good moral properties instantiated on many occasions.

Finally, it must be the case that observer O's statement has assertoric force in order to fall under the realist schema. Questions, commands, promises, etc. are all well and good (and the realist should have some story about how to account for them and which are appropriate), but they do not admit of truth or falsity. Commands, for example, may be satisfied (or not) depending on whether people obey them. Promises bind (or not) depending on the agents' intentions in making them and their subsequent actions. Only *assertions* may be true or false, so it must be the case that O is trying to say something about the moral properties of an object or that a moral fact obtains in order for the realist to say that O's statement must be queried against the set of moral facts and properties.

Having established these points, we can now consider the nature of moral reality according to the realist. The formulation above makes it clear that moral

---

[44] This is a problem for realist interpretations of Kantianism. I read Kant at certain points as giving us the conditions for moral judgment and action if they exist at all. When presented with counterexamples, many Kantians are happy to retreat to the (empirical) thesis that moral judgment and action happen rarely, perhaps never at all. (For examples of these concessions, see §2.1.1.) I don't think the latter case is sufficiently realist, for it gives only hypothetical conditions for moral judgment and action that are consistent with error theory. The same could be said of any Platonic theory that fails to claim that moral properties are ever instantiated; non-naturalism is fine (relatively speaking), as long as moral properties are sometimes instantiated in this world.

properties exist independently of the mental states of observers and, indeed, antecedent to the act of observation. Wrongness, for example, is a property of an act of murder whether or not anyone observes this fact, and the property is instantiated in the action before the act of observation takes place (I can't observe something that's not there). Nothing in this thesis commits the realist to the view that moral properties are *always* instantiated or that moral facts are eternal. It is possible that nothing is right or good until sentient creatures come on the scene. In such cases, those creatures may be involved in (but do not play a role in constituting) the properties in question. Consider a moral property *m* that causes pleasure. It may be the case that nothing is *m* until sentient creatures are around because it is a constitutive part of *m* that it causes pleasure in them. Thus, *m* is a "mind-dependent" property. This fact does not in any way imply that those creatures determine what is *m* or even the general nature of *m*. The underlying factors that result in *m* may be in place long before such creatures appear, but until they appear, (the potential) *m* does not have the effect that is a necessary condition of its being *m*. To put this more broadly, any psychological property or property that essentially involves a psychological property (e.g., pleasure) will not be instantiated until there are minds situated in the right way to perceive it. The property is still independent of those minds in the sense that it does not arise from the intentional states of the creature; it does not depend on the creature's *stance* as a condition of its existence or its nature (i.e., the particular features it has).[45]

---

[45] The attempt here is (roughly) to distinguish secondary or response-dependent properties from constructivist properties. Milo (1995) makes the canonical distinction between mind-dependent and

Accordingly, moral realists claim that moral properties are stance-independent, though they may admit that some such properties (e.g., being the object of desire) are mind-dependent.[46]

The exact nature of moral properties is a matter of some dispute among realists. Naturalists claim that these properties are identical to non-moral properties. On Boyd's account, goodness is identical to the cluster of physical properties that constitute important human goods and the homeostatic mechanisms that unify them. (*Very* roughly, homeostatic mechanisms ensure that particular properties are mutually supporting within a cluster, such as properties involving nutrition and exercise for the property of health.)

On non-naturalist accounts, moral properties are irreducible. In a well-known example, Nicholas Sturgeon (1985) claims that Hitler's depravity depends on non-moral facts about him inciting anti-Semitic sentiments, ordering Jews to concentration camps, and instigating their execution. But Sturgeon goes on to say that those facts alone, without the additional *moral* fact of his depravity, are not

---

stance-dependence properties, where the former include mental states as essential components while the latter are instantiated only if they are the object of an intentional stance. This formulation is used by Shafer-Landau (2003) and many others, but I think there are problems with it. Depending on the creature, an act of perception may include concepts and *eo ipso* involve intentional content. In this case, there is no difference between saying a property supervenes on the creature's mental state and that it supervenes on being an intentional object of the creature's mental state. Intuitively, the difference is between my passively perceiving something and my actively determining something. I think the best way to cash this out is to split the difference between the existence of the property and the effects of the property. In both cases, the property doesn't exist until a mind is around to interact with it. But in the case of "stance-dependence," that mind determines the nature of that property. Another way of saying this is that the property is a logical construct out of the creature's intentional mental state, which is just to say that the mental state explains why the property exists. In the other case of mere "mind-dependence," it is just the opposite; the property serves to explain why the mental state exists. Given that the stance-dependent/stance-independent distinction cuts at the same joints as my account and that the former is standard nomenclature, I am happy to employ it in this section and later ones (esp. §4.1.1). Strictly speaking, we would do well to substitute "explanation of mental states" and "logical construct out of mental states," respectively.

[46] For the contrasting position of stance-dependence, see §4.1.1.

sufficient to explain his behavior or our beliefs about it. Moral facts and properties turn out to be irreducible, *sui generis* properties. Shafer-Landau advances the same general view, arguing that non-natural moral facts and properties set standards for the normative statuses of actions, agents, and the like (2003, 102).

Though these realists advance different views of moral properties, it should be clear that according to all of them, moral properties are extremely complex. They consist in or supervene on varied biological and social factors or on aspects of people's characters and rules for action. In many cases, they involve the presence of sentient creatures and are extremely disjunctive. Consider, for example, a realist version of utilitarianism on which goodness is whatever causes pleasure for the greatest number of sentient creatures. This property will depend on a myriad of causal powers and psychological states, and its instantiation will occur across extremely different events. In fact, some of the most interesting (and unaddressed) metaethical questions for the realist may be how to interpret such normative statements as "rightness is whatever produces the most good for everyone involved" or "the virtues constitute a unity." I leave this as an area for further inquiry. For the moment, it is worth pointing out that the complexity of the realists' moral properties may be why so many have stressed the disanalogy between science and ethics. In science, it is at least possible to design instruments or tests to detect most properties under controlled experimental conditions. The process can be repeated in successive cases to establish the robustness of the property. In the moral case, there are no microscopes or laboratories that can similarly detect moral properties, even if they do exist. They are spread across a variety of confounding

situations and bind together seemingly disparate things, from neutrons to neurons. This disanalogy, it must be stressed, does not undermine the externalist's claim that general methodology for approaching science and ethics is the same; it only brings home the fact that the particular methods for examining each must be different.

Having appreciated several of these subtleties of realism, we can now consider the chief arguments for the position. As I will argue, none of them are successful, but that does not preclude us from considering what realism would look like if it were true (and the deeper problems that attend it), and I will take up that consideration in §3.3.

3.2. Arguments for Moral Realism

Moral realists claim that moral thought and talk represents a reality that we can discover—a moral reality of stance-independent facts and properties—and that moral remarks are true just in case they say that this reality is how it, in fact, is. This account, they claim, reflects our everyday experience, which gives us evidence to support this overall metaethical stance. Moreover, they claim, observations about the surface grammar of moral language, embedded contexts, and truth-preservation in logical arguments all support realist claims about moral properties and moral language.

In the following sections, I argue that none of these considerations is adequate to establish the truth of moral realism. In particular, §3.2.1 argues that realist views of moral experience, specifically what can be concluded from it, are

seriously misguided in light of the feature of intentional inexistence we examined in Chapter 2. Following that, §3.2.2 takes up consideration of moral language. I claim that the realist has seriously overstated the degree to which all moral expressions fit the assertoric model required by the realist account, and even those that do are not proof of the full battery of realist claims. Similarly, §3.2.3 offers brief reflections on embedding and truth-preservation, arguing that even if the realist is correct about them, they fail to establish robust realism.

3.2.1. The Transparency of Moral Experience

Many realists argue that their account is reflected in the very nature of our moral experience. In the course of making moral judgments, holding moral beliefs, and so on, we can't *help* but think there exists an independent moral reality to which our moral mental states were answerable. Simply put, we are *all* realists in the course of our moral experience. Jonathan Dancy puts forward this phenomenological argument for realism as follows: "taking our experience at face value, we judge it to be experience of the moral properties of actions and agents in the world. And if we are to work with the presumption that the world is the way our experience represents it to us as being, we should take it in the absence of contrary considerations that actions and agents do have the sorts of moral properties we experience in them. This is an argument about the nature of moral experience, which moves from that nature to the probable nature of the world" (1986, 172). As Dancy makes clear, this argument for realism proceeds from the nature of moral experience to the existence of moral reality. This trajectory is found in other forms

of realism, including scientific realism, which moves from observational experience to the claim that scientific entities exist mind-independently.

Arguments for *scientific* realism are often more sophisticated than Dancy's "face value" argument for moral realism. Michael Devitt, for example, claims that realism provides not only our best explanation of our observational experience but also of cases where we manipulate the purported entities to produce a result (the abductive argument) and of cases where the world turns out to be as theories predict it to be (the argument from theoretical success) ([1984] 1991, 113–14). Whether or not these claims work in the scientific case, they do not apply to the moral one. Setting up a "moral experiment" with controlled moral properties and seeing how agents respond won't tell us anything about those properties or whether they even exist because the agents could just be immoral.[47] Moreover, it's not as if we can *apply* moral theory in the same way we can apply scientific theory: moral theory doesn't build bridges or send rockets to space. As Wilson points out, "If the parallel with theories in natural science were to hold, the true moral conjectures would be the ones we have selected, through a determinate process of inquiry, for their puzzle-solving ability, their empirical adequacy, and for enhancement of our capabilities—prediction, control, manufacture—that they offer….[But t]here are no readily-identifiable expert communities who possess and

[47] One might argue that immorality is just a lack of sensitivity to moral properties, no different in principle than, say, deficiencies in vision or other sensory mechanisms that might plague perception of scientific entities. This response won't do because someone could be receptive to moral properties (i.e., know what the right thing to do is) and yet act counter to that. Immorality and moral blindness issue in extensionally equivalent non-verbal behaviors.

disseminate moral information" (2010, 102).[48] Applied ethics is not the same as applied science in the sense that the applied sciences (e.g., engineering) provide confirmation of scientific theories by means of practical results. While it's true that the world may become better if we discover a true moral theory—lives may be saved, unnecessary suffering, eliminated, etc.—there is no way of assessing these "good" results *independently* of the moral theory itself. Whatever the theory recommends will be "good," and implementing the theory will, under favorable conditions, bring about "good" results in a trivial sense. But there is no means of assessing the results independently of what the theory (or at least *some* theory[49]) recommends, and, as a result, applied results cannot count as confirmation on pain of circularity.

Gilbert Harman (1977) attacks more basic arguments for moral realism, noting that moral facts and properties are not necessary to explain token moral judgments on the part of observers. Harman's example, as we have seen, is a person who watches children pour gasoline on a cat and ignite it. The observer sees a certain pattern of light on his/her retinas and has knowledge of various background theories about people, cats, gasoline, burning, and so on. The observer may also have certain beliefs about rightness and wrongness that contribute to his/her moral judgments about the situation. As Harman points out, we do not need to make any further assumptions about the existence moral facts to explain

---

[48] Wilson goes on to argue that moral realism is not necessary for an account of moral progress.

[49] As I argued in §2.1.1, assessing the results of one moral theory by means of another begs the question against the target theory. Better assessments would involve consistency and coherence with respect to both the prescriptions of the theory itself and to larger philosophical commitments.

these judgments in the observer, merely some claims about his/her psychology or "moral sensibility."

Harman contrasts this case of moral observation with another case in which a physicist sees a vapor trail in a cloud chamber and thinks, "There goes a proton." In this case of scientific observation, we must assume (for the purpose of explanation) the existence of a proton, which caused the vapor trail observed by the physicist, in the chamber. Harman concludes: "Observational evidence plays a part in science it does not appear to play in ethics, because scientific principles can be justified ultimately by their role in explaining observations….Conceived as an explanatory theory, morality, unlike science, seems cut off from observation" (1977, 9). Thus, no moral observations count as evidence for moral theories because moral facts need not be posited to explain the occurrence of moral observations.[50]

It is worth noting here that the most serious objection to Harman's argument *presupposes* the existence of ethical facts and then goes on to explain their role in ethical observations: Nicholas Sturgeon (1986) claims that explanations of moral beliefs that reference moral facts are *better* than competing explanations that make no such reference. On his view, the fact of Hitler's moral depravity—over and above his instigation and oversight of the degradation and deaths of millions of people—provides the best explanation of why we judge him to be morally reprehensible. "Once we have provisionally assumed the existence of moral facts,"

---

[50] Morrow (2009) points out that explaining token moral judgments is only half of the explanatory task; we must also explain why we make moral judgments at all, and this story, which probably includes a fair bit of evolutionary psychology, may require a modest amount of moral realism. I believe I have incorporated the relevant aspects of moral realism into my own interactionist account without taking on the unwanted baggage of moral realism.

Sturgeon claims, "they *do* appear relevant, by perfectly ordinary standards, to the explanation of moral beliefs and of a good deal else besides" (1985, 57). Whether or not Sturgeon's claim is true (and many have doubted it), the question at present is whether there are good arguments in support of moral realism. Assuming the existence of moral facts from the start begs the question and provides no independent ground for setting metaethical disputes. We should conclude then, with Harman, that explanations of the intentional content of moral beliefs require no posits about moral facts and properties.

The problem involves more than just standards of explanation. To appreciate this, we can consider Michael Tye's claims about the "transparency" of experience. According to this thesis, introspecting one's own experience gives one a direct awareness of the external world, not (just) the inner qualities of that experience.[51] Tye says: "When you introspect your visual experience, the only particulars of which you are aware are the external ones making up the scene before your eyes. You are not aware of those objects *and* a further inner object or episode. Your awareness is of the external surfaces and how *they* appear. The qualities you experience are the ones the surfaces apparently have. Your experience is thus transparent to you" (2002, 139). As Tye points out, nothing in this argument presupposes the veridicality of experience. One can be wrong about one's visual experience, but then one is wrong, in a transparent way, about the way the world is and not (just) one's own experiences. Tye's argument pairs nicely with

---

[51] This view is contrasted with early twentieth-century sense-datum theories on which one is only aware of intermediate *sensa*, not the world itself.

an externalist methodology: to access claims, we only need to look at the world, and we can do this by looking through our own experience of it, as it were.

There are two problems in applying transparency arguments to support of moral realism. First, the lesser problem is that arguments for transparency almost always turn on cases of direct perception, such as vision. In those cases, there are sensations that *are* directly caused by things in the world. This makes it easy to claim that cats cause CAT-thoughts and that in having cat thoughts, we are having thoughts about the things in the world that cause them. Moral mental states, as we saw in Chapter 2, are completely different. They are more complex, ranging over disjunctive objects, events, and properties, and they are not as closely related to sensations as non-moral mental states are related to sensation. How, for example, are Rawls' reflections on the nature of justice directly related to sensations (or even examples of justice in the actual world)? The analogy with perception largely fails in the case of moral mental states and so goes transparency.

The second, deeper problem turns on the issue of intentional inexistence. As we saw in Chapter 2, moral mental states are intentional states *par excellence* (though they are more than just pure intentional states). As such, they exhibit the feature of intentional inexistence: they can be about things that do not exist. From the mere fact that I have a thought about virtue or the nature of good, nothing follows about the existence of those things in the actual world.[52] This means that

---

[52] Notice that this problem applies, too, to scientific observations, which are doubtless theory laden and involve heavy use of concepts with intentional content. Intentional inexistence may threaten scientific realism and any other kind of realism. I suppose there is some story the realist can tell about the origins of intentional content such that it will turn out that *some* things must exist mind-independently

any appeal to moral experience will be insufficient to establish the existence of moral reality. The realist may be able to marshal other arguments in support of his/her views about moral reality (as scientific realists often do for theirs), but I have no idea what those are, given the disanalogies between scientific entities and moral entities we have examined. The point here is that intentional inexistence *guarantees* that *no* amount of reflection on moral experience will *ever*, by itself, suffice to establish the existence of the stance-independent moral facts and properties to which the realist is committed.

This objection is disastrous for moral realists. The other arguments for moral realism we will examine aim only to establish that moral language is truth-conditional. But this claim, by itself, is consistent with error theory and too weak to ground the full realist thesis. After all, even Mackie thinks moral statements are true *or* false; he just happens to think they're all false. The point at which he and the moral realist diverge is whether at least one such statement is true of the actual world. Without some argument for the existence of moral reality, the realist thesis is lost, and in light of intentional existence and the failed argument from moral experience, I do not see how they can resuscitate realism.

3.2.2. The Surface Grammar of Moral Language

In addition to arguments about the nature of moral experience, moral realists also claim that their theory reflects everyday talk about moral dimensions of actions,

---

in order for there to be intentional content in the first place. Davidson (1977), to my mind, gives such an argument, but does not specify *which* things exist, only that *some* must (or rather, that massive error about the world is unintelligible).

events, characters, and institutions. Moral language, they claim, aims at true descriptions of the world. Thus, statements such as "Murder is wrong," "Torture is cruel," and "Charity is good" all purport to be true statements about moral reality. According to externalism, the method of verification for these statements involves determining whether they, indeed, reflect stance-independent moral facts and properties in the actual and ideal worlds. David Brink summarizes this argument as follows: "The syntax of moral discourse is descriptive. Moral judgments are expressed in the declarative mood and treat moral predicates as modifiers of noun and verb phrases. As such, they appear to express propositions ascribing moral properties to persons, actions, and institutions; acceptance of these propositions seems to take the form of belief. We ought to give up a cognitivist construal of moral discourse only if cognitivism can be shown to involve unacceptable commitments" (1997, 9). Accordingly, the illocutionary force of moral statements is assertoric, the attitude of moral mental states is belief (or veridicality more generally), and moral talk and thought are answerable to the external world.

This descriptivist model[53] of language has been dominant in philosophy since at least the time of Descartes, but the case for non-descriptive language has also been made, most notably in J. L. Austin's ([1962] 1974) discussion of performative utterances. Austin argues that there is a "descriptive fallacy" in understanding *all* utterances as descriptions of states of affairs in the world or

---

[53] Call "descriptivism" the view that moral remarks purport to represent states of affairs in the world and, as such, should be understood as truth-conditional assertions of belief about moral facts and properties. Descriptivism, when combined with the claim that some moral facts and properties exist in the actual world, yields moral realism.

statements of fact. In fact, many utterances are performative in nature (e.g., stating wedding vows, making promises, placing bets) and must be understood on a non-descriptivist model. These speech acts do not aim at being true, but rather at being binding, successful, obeyed, and so on. Correspondingly, the mental states that these speech acts express have world-to-mind, not mind-to-world satisfaction conditions. Attention to the descriptive fallacy is often a starting point for performativist views on which moral language aims not to describe but rather to change the world in certain ways. We shall return to these views in Chapter 5.

For the moment, it is worth noting that many moral sentences are not assertoric statements but rather commands, wishes, injunctions, and the like. Consider, for example, the sentences:

E3 "Develop your talents!" (Kant [1785] 2002, 4:423)

E4 "Do whatever your conscience tells you!" (Butler [1726] 1827, Sermon II)

E5 "Feminists: Embrace the language of liberalism!" (Nussbaum 1999, 56)

The surface grammar of each of these is non-assertoric; they are imperative/prohibitive (expressing commands, requests, and prohibitions) or optative (expressing hopes, wishes, or commands with a subjunctive flavor). As such, they do not seem to admit of truth or falsity. That is not to say there are *no* standards for assessment; such remarks may be vague or precise, wise or foolish, worthy or unworthy of being said, and so on. The point, however, is that they are not, strictly speaking, true or false. Admittedly, they may *reflect* statements that are genuine assertions:

E3′ "It is right to develop your talents."

E4′ "The right action in any situation is determined by one's conscience."

E5′ "The liberal notion of all people as free and equal is good for feminists."

But the earlier non-assertoric remarks are distinct from these latter examples. The realist may attempt to say that (E3)–(E5) can be reformulated along the lines of (E3′)–(E5′) to admit of appropriate analysis. But this revisionary account would undermine the realist's own claim that realism does justice to ordinary language by reading the semantics of moral remarks right off of their surface grammar. In effect, reformulating non-assertoric remarks into assertoric ones undermines the realist's argument about surface grammar. If the realist is allowed to give a revisionary account of *some* moral sentences, how can he/she then turn around and criticize other accounts, including performativism, for being wrong-headed in *their* revisionary view of moral language?

Performativists try to accommodate all moral language to a non-assertoric model. So they are naturally prepared to handle (E3)–(E5) but stumble over (E3′)–(E5′). In a parallel way, realists are poised to analyze (E3′)–(E5′) but have trouble accommodating (E3)–(E5). The two theories are symmetrical and opposed, and there is no reason antecedently to think that one theory is better able to handle moral language than the other. The realist may be prepared to handle a larger number of moral *utterances* than the performativist (assuming most moral utterances *are* assertoric), but quantity is irrelevant here. We are considering two *kinds* of moral sentences—descriptive/assertoric and non-descriptive/non-assertoric—and the

work of analyzing both will have to be done, one way or another.

In any case, claims about surface grammar and the truth-conditions of moral language (i.e., cognitivism) are insufficient to net the realist his/her theory. As I have been arguing, error theorists, too, think that moral language is truth-conditional. The crucial task for realism is defending the claim that moral facts and properties actually exist, and as I argued in the previous section, the outlook is not good. So even if realists were right about the surface grammar of moral language, that alone would not establish realism.

3.2.3. Embedded Contexts and Truth-Preservation

According to moral realism, ordinary moral statements are assertoric and truth evaluable. As we have seen, this is a sharp contrast from non-descriptivist theories, such as performativism, on which moral remarks are not (always) assertoric and do not have truth conditions. In "Assertion," Peter Geach presents an argument about meaning that has been taken as an important criticism of performativist theories and further support for realism. Citing what he says is a point of Frege's, Geach claims that the content of a thought (or the meaning of its correlative expression) should be the same whether or not it is offered in a context of assertion. He gives the following example:

> If doing a thing is bad, getting your little brother to do it is bad.
>
> Tormenting the cat is bad.
>
> *Ergo*, getting your little brother to torment the cat is bad.

> The whole nerve of the reasoning is that "bad" should mean exactly the same at all four occurrences—should not, for example, shift from an evaluative to a descriptive or conventional or inverted-commas use. But in the major premise the speaker (a father, let us suppose) is certainly not uttering acts of condemnation: one could hardly take him to be condemning just doing a thing. (1965, 463–64)

Geach's complaint about performativism is that it violates this requirement by making the meaning of moral remarks descriptive in some contexts (e.g., the major premise above) and evaluative or conative in others (e.g., the minor premise).

Geach's criticism raises two separate but related problems. First, there is the embedding problem. In the major premise of his example, the clause "getting your little brother to do it is bad" has been embedded within a larger sentence—the same thing that happens in *that*-clauses, such as "He knows that the earth is around" and not just "The earth is round." The truth conditions of the embedded and non-embedded sentences differ, and it seems that only descriptivists, including realists, can give a straightforward analysis of these statements because only they treat *all* parts of moral sentences as having truth conditions. On other accounts, including non-descriptivism and performativism, sentences with embedding are semantically complex, for they contain truth-conditional and non-truth-conditional parts.

A similar and second problem arises with logical inference. Since the major premise above, on a non-descriptivist or performativist account, contains truth-conditional and non-truth conditional parts, it is not clear that the argument is

sound or even *valid*, since the meaning of 'bad' seems to vary across its occurrences. On a descriptivist/realist account each premise has a truth-value and the meaning of 'bad' remains the same, making it possible to easily evaluate the argument for validity and soundness.

Neither of these problems strikes me as particularly impressive. In response to the first, it is possible to imagine embedded contexts that cause no problems with semantic evaluation. For an asserted or descriptivist sentence *p*, ⌜It is true that *p*⌝ will save truth-value for any substitution of *p*. For a conative sentence *c*, ⌜It is commanded that *c*⌝ will save "command-value" for any substitution of *c* (and so on for other illocutionary forces). So the problem with embedding is not *embedding*, syntactically speaking, but really the *fit* between the embedded and embedding clauses. Where the two are complementary (e.g., true–truthy, commanded–commanding), no problem arises. Moreover, Blackburn (1988) has given a compelling account of the "logic of the attitudes" that is widely recognized as solving Geach's worry about how to interpret moral arguments, logically speaking.[54]

The genuine problem, as I see it, is how to deal with *hybridity* in systems of semantic evaluation. How shall we analyze the semantics of a statement such as "Murder is wrong and snow is white," in which one part of the conjunct has truth conditions and the other part does not.[55] Similarly, how do truth-conditional and

---

[54] We saw this in §1.1.2.

[55] In presenting their norm-expressivist account, Horgan and Timmons (2006) develop a formal language in which each sentence is the product of inserting an open or closed nonsentential formula

non-truth-conditional premises (or their parts) figure into the validity and soundness of logical arguments? Consider the example:

Snow is white or *Damn it all to hell!*

Don't damn it all to hell!

Ergo, snow is white.

Admittedly, this is a strange argument—although one can almost imagine Moore saying this about his hands! Nevertheless, we can all see the standard logical form that it trades on, and we all recognize that there is something correct about it—and not just because snow *is* white. These are problems for realists and non-realists alike. It's simply implausible to suggest that *all* sentences (or even all *moral* sentences) are assertoric. There are genuine commands and other kinds of speech acts, and even if we don't find them in the philosophical literature very often, *we* can at least talk that way.

The realist has the market on semantic evaluation only insofar as he/she oversimplifies facts about speech acts and assimilates everything into an assertoric model. As I argued in the last section, it's equally implausible for non-realists to do the same with their models. The overarching point is that problems of semantic evaluation affect us all, and Geach's preoccupation with assertion and

---

into the bracketed slot of an is- or ought-operator. These two operators correspond to different species of belief separated by their distinct ways of mentally affirming a core of truth-evaluable cognitive content. A complex sentence containing both is- and ought-operators is true just in case the partial valuations are true. In developing this language, Horgan and Timmons seem to meet my concern about the semantics of hybrid statements. However, even they concede: "The fundamental semantic principles governing the operator **O**[ ] are quite weak. One could consider strengthening them in various ways, even while retaining their *if/then* form. But we doubt that there is adequate theoretical motivation for doing so. Also, a general reason to avoid stronger principles is the need to avoid various well-known deontic paradoxes" (294 Note 9).

performativist "failings" is one of those problems, not a definitive argument for moral realism.

3.3. How, Really, to Be a Moral Realist

If the preceding sections are correct, none of the standard arguments succeed in establishing moral realism. But what if they did? In this section, I want to consider what moral realism would actually look like, especially with reference to moral mental states. This topic is largely undiscussed in the literature; most authors focus on the cognitivist aspects of moral realism or the nature of moral facts and properties (e.g., naturalism vs. non-naturalism). On my general approach, we need to consider prior questions about the connection between the mind and the world before we can go on to discuss downstream semantics or even metaphysics. Accordingly, I am going to consider what realists *could* say about moral mental states, both in the mind-to-world direction and the world-to-mind directions that I discussed in Chapters 1 and 2.

3.3.1. Observation, Direct and Indirect

Moral realism, as a species of externalism, puts a premium on observation and discovery of pre-existing reality. In §3.2.1, I argued that intentional inexistence imposes severe limits on what lessons the realist can draw from the "observations" of moral experience. These worries extend a bit deeper in the moral case than in common sense cases or even scientific cases of perception. In ordinary cases, if conditions are normal and I am awake, sober, and not hallucinating, the intentional content of my perceptual state probably does refer to some object(s) in the world.

Moral mental states, however, frequently involve *future* states of affairs that we hope or fear will come to pass. When I reflect morally, I may think about the badness that would result if I break a promise, or the rights that might be violated if certain privacy restraints are not upheld, or the virtue that lies in helping those in need. These cases, and those involving past misdeeds, generosities, and the like, are ones in which the referents of my moral judgments do not, strictly speaking, exist—that is, they do not exist in the actual world.

This "non-present" aspect of moral judgments should invite serious doubts about what metaphysical claims can be read off of the supposed causal connection between things in the world and representational mental states. Notice that this worry extends a bit beyond Harman's claim about token moral observations and concerns a persistent feature of moral judgment that is rooted in future planning.[56] Even if we assume that intentional mental content is caused by certain non-intentional features of the world and even if we assume that moral judgments, in principle, could be caused in exactly the same way, there is still reason to doubt that *actual* moral judgments entail any metaphysical conclusions because future states of affairs do not cause anything to happen in the present. In short, it is even harder to derive claims about moral ontology from moral mental states, given the objects of these states, which are not about the actual world.

---

[56] In the last paragraph, I claimed that many moral judgments concern past actions. I think that, as an empirical fact, we usually make these judgments only when they concern either future punishment or reward, or the influence that past actions will have on future ones. Even retributivists should be happy with this claim, for in recommending punishment based on past action and for no other reason than the fact that wrongdoing occurred, they are nevertheless recommending a *future* action. In seven studies by Caruso (2010), participants judged wrong deeds in the future more negatively, and good deeds in future more positively, than equivalent behavior in the equidistant past. Participants also thought that unfair actions in the future deserved more punishment than unfair actions in the past.

Maurice Mandelbaum makes a very helpful distinction between three kinds of moral judgments, each of which raises its own problems:

- "direct judgments of moral rightness and wrongness," "made by an agent who is directly confronted by a situation which he believes involves a moral choice on his part;"
- "removed judgments of moral rightness and wrongness," "made by an observer on the conduct of another person (or by an agent on his own past conduct);" and
- "judgments of moral worth," "judgments of praise or blame which are directed toward specific traits of character, or which concern the total character of a person considered as a moral being" ([1955] 1969, 45).

Mandelbaum's distinction between actions and characters need not concern us here, since we are interested *generally* in the assignment of purported moral properties to objects and events, and a person's character or total character may count as an object or event. The important point is that a person may have a moral mental state about an occurrent object or event or a non-occurrent object and event (e.g., something he/she just hears about, or something he/she imagines).

To bring out this difference between direct and removed judgments and give it a more externalist flavor, we can employ the device of possible worlds. On this account, a judgment about the past or the future, the ideal or the hypothetical will be a judgment about a possible world in which certain objects "exist" or certain events "occur." Accordingly, only a small number of our moral judgments will be

about the actual world (e.g., situations such as Harman's cat); the vast majority of them will be about other possible worlds.

Mandelbaum's distinctions can be recast as two kinds of moral observations that the realist must countenance:

DIRECT OBSERVATION O observes *m* instantiated in an object or event in α

INDIRECT OBSERVATION O observes *m* instantiated in an object or event in **W**, where $\mathbf{W} \neq \alpha$

Observations about past, future, hypothetical, or counterfactual states of affairs will all count as "indirect," while observations about immediately occurring objects and events will count as "direct." With this framework in place, I want to consider how the realist identifies the moral properties in question across possible worlds.

Presumably, cases of direct observation are not enough. Even if observer O has occasion to make many observations of torture and other actions such that he/she can reliably say what cruelty is and when it occurs, he/she may want to know more about possible future cases or unobserved ones (e.g., would torture in a ticking-bomb case[57] count as cruel? what about torture followed by an amnesia-

---

[57] Shue (1978) gives a helpful account of what justified torture might look like: The proposed victim of our torture is not someone we suspect of planting the device: he *is* the perpetrator. He is not some pitiful psychotic making one last play for attention: he *did* plant the device. The wiring is not backwards, the mechanism is not jammed: the device *will* destroy the city if not deactivated….The torture will not be conducted in the basement of a small-town jail in the provinces by local thugs popping pills; the prime minister and chief justice are being kept informed; and a priest and a doctor are present. The victim will not be raped or forced to eat excrement and will not collapse with a heart attack or become deranged before talking; while avoiding irreparable damage, the antiseptic pain will carefully be increased only up to the point at which the necessary information is divulged, and the doctor will then immediately administer an antibiotic and a tranquilizer. The torture is purely interrogational. Most important, such incidents do not continue to happen. There are not so many people with grievances against this government that the torture is becoming necessary more often, and in the smaller cities, and for slightly lesser threats, and with a little less care, and so on" (142). In light of the disanalogy between such cases

inducing drug?) to understand more fully what cruelty is and when it occurs. On the advice of the realist, O can query other possible worlds in the set $\{\mathbf{W}_{i-j}\}$ to increase his/her observational evidence by indirect observations. O may ask what *past* cases of cruelty looked like or whether a certain kind of behavior *would* count as cruel.

These appeals to past, future, and hypothetical worlds all appeal to other possible worlds and the objects, events, and moral facts properties that "exist" there. This reference to possible worlds raises interesting questions about the meaning and reference of moral terms according to realists. Once the realist fixes on a particular moral property *m* given observations of that property in the actual world, how is it that he/she accesses that property in other possible worlds to increase the set of his/her observational evidence? What guarantees that the *m* O identifies in α is identical to the *m* O identifies in **W**, given the complexity of moral properties and the diverse situations in which they occur?

It is worth noting that this question about how moral properties are picked out cross-worlds is different from the question of how observations of moral properties can be *correct*. Most realists agree that "observation" in the sense relevant for moral realism involves moral reason and reflection. Boyd, for example, looks to reflective equilibrium to fine-tune our moral intuitions (1988, 207),[58] while

---

and actual conditions, Shue argues for a (defeasible) presumption against the moral permissibility of torture.

[58] Boyd does think a large part of our moral knowledge (i.e., our knowledge of homeostaticly clustered moral properties) will be experimental (203), akin to the social sciences (206), but he recognizes that "*some* of what must count as observation is observation of oneself, and *some* of the sort of self-observation involved is introspection" (206). This data, he thinks, must be subjected the process of reflective equilibrium to purge them of their theory-dependent influences.

Shafer-Landau averts to reason (2003, 166–69). Realists do not claim that we literally point to possible worlds as evidence for moral claims (as we can, for example, point to the actual world in cases of scientific dispute). Instead, a moral assertion advanced within a realist framework will be answerable to and scrutinized by the community of moral inquirers, who, possessing reason themselves, will review the claim and make their own "observations" through reason and reflection or defer to others in their judgments. *This* level of response—observation or deferment—*is* exactly the same as the scientific case in which the scientist, upon being presented with a scientific claim, can either run an experiment and generate more observations or trust in the observations of his/her fellow scientists. Some observers are better than others. Some fabricate data; others are more assiduous in their claims and considerations. The point is that realism establishes a *framework* for this process; it does not *guarantee* correct results. We must try to understand, then, how observation functions in this framework, and our immediate question is how observers identify moral properties *across* possible worlds.

We can imagine this process working two different ways: (1) survey all possible worlds for all actions, events, characters, and institutions and see which of them are *m*, or (2) examine the possible worlds in which *m* is instantiated and see what actions, events, characters, and institutions it is there instantiated in. But which of these more accurately reflects the claims of moral realists?

As Kripke tells us, possible worlds are what we take them to be: "'Possible worlds' are *stipulated*, not *discovered* by powerful telescopes" (1972, 44). In this sense,

we determine what will happen in the possible world in question according to how we stipulate it to be.[59] If that possible world is a nearby world, we are constrained in stipulating the possible world in question: we take the actual world as a baseline, as it were, and stipulate the respect(s) in which our nearby possible world differs. So the realist, although "observing" these possible worlds, is, in fact, stipulating the conditions of the world under discussion (e.g., that the object or event be there, that conditions are roughly similar to the actual world). This claim suggests that the second strategy—stipulation and investigation—is the strategy employed by moral realists. In making indirect judgments, observers stipulate worlds in which *m* is instantiated—*m* based on the extension of *m* in the actual world. This act of stipulation ensures that *m* in α is identical to *m* in **W** because (roughly) the home team determines reference.[60] Having guaranteed via stipulation that the moral property of interest is instantiated in the world(s) under examination, the observer need only examine the actions, events, characters, and institutions in that world to see whether they are *m* or not. To take one example, suppose O fixes the referent(s) of 'cruelty' based on observation of torture, cat burning, etc. in the actual world. To investigate the nature of cruelty in greater depth, the observer imagines other possible worlds in which cruelty occurs and examines the objects and events in which it occurs there. In doing so, the observer

---

[59] Salmon (1996) has a helpful article in which he explains that Kripke is not arguing that we have any peculiar ontological or epistemological status regarding possible worlds *per se*, only that where possible worlds or anything else is concerned, the discussants may stipulate the entities under discussion. Similarly, the constructivist here stipulates the possible world under discussion and the moral mental states of the idealized subjects in question.

[60] For further discussion of cross-world reference, see Putnam's classic papers "Meaning and Reference" (1973) and "The Meaning of 'Meaning'" (1975).

increases his/her knowledge of cruelty, which he/she can then employ to prevent other cases of cruelty in the actual world before they come about.

Assuming all of this apparatus is sound and that observers indeed have reliable access to possible worlds, two further questions arise: What about moral properties that aren't instantiated in α, thereby denying the realist grounds for referring to *m* in other possible worlds? and Does the extension of *m* ever change in light of this observational evidence? Both of these questions aim at the issue of moral progress. We can imagine new situations arising that would make the actual world better than it is—and in ways we haven't seen yet! Suppose that, in the future, we are able to extend human life by several centuries. Our ideas about what matters in life will no doubt change, for it will be quite intelligible to say that we are different people over the course of our lives.[61] Moreover, we can imagine our own moral concepts improving by the processes of reason and reflection in ways that take us beyond the ways that we currently apply moral terms. What shall we say about such cases and world stipulation?

The first question is a moot point. If I tell you there are other moral properties that we don't know about yet that exist outside of this world, there is no way for you to examine them apart from my description of the worlds in which they occur. In other words, the person who purports to have *otherworldly* moral knowledge gets to stipulate the world(s) under discussion. If these properties are never, have never been, and will never be instantiated in the actual world, one is

---

[61] For examples of the shift imposed on personal identity by such changes, see Lewis' discussion of Methuselah ([1976] 1983, 65–7) and Parfit's discussion of "immortals" (1984, 290, 304–5).

tempted to say, "Great! Who cares?" If, however, they are properties that will eventually come on the scene here, such as something involved with my longevity example above, then we will have to wait and see whether the properties that get instantiated are indeed identical to the purported properties, and once they are instantiated, it's much more likely that we'll fix the extension of the moral predicate according to experience of them in this world (which is no doubt richer than our mere imaginings of possible worlds), and then the process will go on exactly as I have described.

The second question is much more interesting. Could it be that we begin inquiry with a certain stock of examples of rightness, but, by reasoning and reflecting (in this case, querying other possible worlds), we discover new examples, which, in turn, increase the extension of our application of the term? It's plausible (perhaps even historically accurate) to think that people in the past didn't know that human rights extended in full to humans of certain races, genders, nationalities, etc. Fortunately, liberalism and the whole movement of the enlightenment have revealed that such people were in the extension of our moral predicates about human rights all along, even if we didn't know that. (Notice that I didn't say that they *weren't* in the extension of our predicates all along—I said we didn't *know* that they were.) This underscores the stance-independence of the moral properties. They are instantiated or uninstantiated all along *regardless* of what we think about them. Our concepts, of course, and the known extensions of our moral predicates need not line up with this—and probably don't, since we don't (yet) have a finished moral science. The upshot of realism is that by imagining other

possible worlds in which those properties are instantiated, we can learn more about them and come to refine our own concepts and our stance on this world.

But how is it that observers are able to imagine other possible worlds and stipulate the nature of moral properties in them without prejudging the question of which objects and events have those properties there? The problem is that *m* is likely to contain some implicit assumptions about the extension of *m* based on its occurrence in α, so how *could* it be that carrying this over to another possible world via stipulation would show anything new or different about the extension of *m*. To put this a bit more pointedly, why is the realist's "observational" evidence about possible worlds anything stronger than the claim, "Because I said so"? Our "findings" as a result of indirect observations are limited to conceivability, and surely conceivability is limited to our experience, at least to some degree. Others can challenge us, to be sure, but the results of indirect observation would be pseudo-empirical at best (perhaps "derivative" would be a better term) because they would depend on the incomplete knowledge gained from direct observations of the actual world.

Consider, for example, Rawls' claim that utilitarianism will violate principles of fairness by not taking "seriously the distinction between persons" ([1971] 1999, 24) and allowing "the greater gains of some…[to] compensate for the lesser losses of others" and violating the liberty of a few for "the greater good shared by many" (23). On a realist reading, Rawls is claiming that unfairness, losses, and violations are offset and made right in an ideal utilitarian world by the greater good of the man. These must be limited to what he takes utilitarianism, fairness, rights, and liberties

to be and what physical conditions he imagines other possible worlds to have. The utilitarian can challenge Rawls' findings by claiming that the possible worlds under consideration are not the right ones to be examining. The utilitarian may respond (correctly in my view) that a world in which those things happen is *not* the utilitarian ideal.[62] A better world—the one that utilitarianism actually calls for—is one in which rights and liberties are respected. The point is that such things are open questions for discussion, especially where realist stipulation is concerned.

The procedure by which the realist stipulates the conditions of the possible world under examination will determine what kind of realism he/she endorses. Naturalists claim that the conditions of the world we stipulate are determined by the microphysical structure in this world in which moral properties consist or on which they supervene. In other words, naturalists, in stipulating a possible world in which goodness is instantiated, *eo ipso* stipulate certain things about the microphysical structure of that world in virtue of the structure of *this* world and the ways in which good is instantiated *here*. This is just to say that possible worlds for the naturalist must be "nearby" in the sense that their microphysical structure must be the same as in ours with respect to what is or what determines goodness. If goodness supervenes on certain pleasurable responses in humans, then the possible worlds of inquiry for the naturalist must contain sentient creatures with similar sensory mechanisms and cognitive processes to us, as well as similar causal

---

[62] I am indebted to David Morrow for this point.

relations that obtain between the objects that instantiate goodness and these creatures.

Non-naturalists may have a wider berth because rather than stipulating a certain microphysical structure, they need only stipulate certain logical or conceptual relations between objects and events on the one hand and moral properties on the other. For example, they need only stipulate that any case of murder is a case in which wrongness instantiated, and so on for other objects, events, and moral properties. This is not to say that non-naturalists are always discovering *necessary* moral truths. For example, it may be that 'murder', on a non-naturalist reading, appeals to certain contingent facts about murder in the actual world and that the non-naturalist carries those facts over to other possible worlds by considering only a subset of all possible worlds in which murder occurs—namely the subset in which the contingent facts about murder that obtain in this world (e.g., that it robs something of life) also obtain. If there are possible worlds in which people spontaneously resurrect each time they are killed, then murder in those worlds (if it can be said to be practiced there) will probably not fall under the non-naturalist's sphere of interest, and those possible worlds will be excluded from the inquiry. What the non-naturalist should be able to say is that with respect to a restricted subset, $\{\mathbf{W}_{i-j}\}$, of the set of all possible worlds in which murder occurs, $\{\mathbf{W}_{h-k}\}$, it is "necessarily" true that murder is wrong. I think there are important limits that need to be addressed when it comes to this kind of conceptual analysis. As I noted earlier, it's not clear that our moral (or even our non-moral) concepts are correct in the first place. So taking those as bedrock from which to examine logical

or conceptual connections to other concepts seems a bit spurious. Moreover, there seems to be no method of correcting these concepts independent of the process of conceptual analysis itself. At least the naturalist can appeal to microphysical structures of the worlds in question and attempt to fine-tune his/her moral concepts accordingly. Of what corresponding elements can conceptual theorists avail themselves?

Though doubts remain, it should be clear that the realist has *some* story to tell about what moral observation consists in and how it is aided by and functions in the possible worlds implicit in indirect moral judgments. While realists have more work to do, we at least have a working understanding of the means by which the realist determines whether the mind-to-world satisfaction conditions of any moral mental state have been met. Though I doubt that realists can construct a robust account of these conditions given that possible worlds are mere stipulations and not full empirical evidence, we should set this problem aside and consider instead how world-to-mind satisfaction conditions would operate in realist accounts.

3.3.2. Motivation and Possible Worlds

In Chapter 2, I argued that moral mental states have world-to-mind satisfaction conditions that involve certain behaviors, even if those behaviors are, in some cases, mere performances of speech acts. Here, I want to consider whether moral realism can accommodate these features of moral mental states. In general, the question we need to consider is, How do moral properties interact with an

observer's volitional states, including the volitional aspect of moral mental states in a way that satisfies the constraint that moral mental states are successful or unsuccessful, in part, in respect of whether they issue in certain behaviors?

To begin, accounts that make reference to mentality in their description of moral properties fare better here, *prima facie*, than those that do not. On a stance-dependent view of moral properties, those properties depend crucially on the intentional stance of the observer, and, accordingly, it is easier to imagine objects and events being connected to the agent's own plans, drives, and ambitions. At the very least, the realist must make clear how purported moral facts and properties figure into agents' practical reasoning about what to do. Christine Korsgaard, for example, says that "we have normative problems because we are self-conscious rational animals, capable of reflection about what we ought to believe and do. That is why the normative question can be raised in the first place: because even when we are inclined to believe that something is right and to some extent feel ourselves moved to do it we can still *always* ask but is this really true? and must I really do this?" (2003b, 46–47). On her constructivist account (to be examined in the next chapter), moral facts and properties are logical constructs out of a subject's judgment of what, from the practical point of view, he/she ought to do. That makes morality a direct solution to the normative problem.

The difficulty for the realist lies in the stance-*independence* of his/her purported moral facts and properties, which place those facts and properties outside of the concerns, preferences, and plans of agents. Stance-independence may play well with cognitive, mind-to-world satisfaction conditions—say, those

that accompany straightforward moral beliefs—but they seem cut off from conative, world-to-mind satisfaction conditions of moral mental states.

Before examining two possible solutions to this problem, I want to highlight what cannot count as a solution. There is a very general sense in which many or perhaps even all perceptions prompt observers to act in certain ways: I see an unknown object coming towards me and I avoid it; I'm cold and I see a fire and I move towards it; you yell, "Watch out!" and I take note of my surroundings. This pattern is explained by the fact that our everyday perceptions act as inputs to cognition, which often has behavioral outputs. But this is the same black box that I was just complaining about. In the moral case, we want to know *how* stance-independent moral facts and properties act as inputs in ways that make agents *wrong* if they don't act on them in appropriate ways. The realist does not need to show that we *must* be moral—surely people can fail to respond properly to the demands of morality. The realist needs an account of moral mental states that is robust enough to show that we are *wrong* (as in "have failed," not "are incorrect") if we don't act in the proper way based on the moral facts and properties we observe.

The moral realist can try to explain this in one of two ways. Call the first approach the "causal powers strategy." On this tract, the moral realist can claim that the causal powers of moral properties are such that they induce not only mental states with mind-to-world satisfaction conditions but also states that have volitional components and world-to-mind satisfaction conditions. We can, of course, fail to satisfy those conditions, but this explanation locates the source of those conditions in the moral properties themselves. These causal powers are no

more mysterious than the general sense in which non-mental things can occasion and act as inputs to mental states, and something similar has been advanced by a variety of naturalists from Epirucus to Hobbes. In fact, Hobbes seems to put the point nicely in saying "endeavour, when it is toward something which causes it, it called appetite or desire….And when the endeavour is fromward something, it is generally called aversion" ([1651] 1994, VI.2). On this mechanistic account, objects in the world impinge upon us, resulting in our motions toward certain things and away from others. This is no different than saying that the properties of some objects have certain pulls or pushes with respect to our volitions and motivational states, which is exactly what the realist hoped for on the causal powers account.

This process runs afoul, however, in considering possible worlds. As we saw in the previous section, moral realists depend on a notion of possible worlds to talk about indirect moral judgments concerning past, future, hypothetical, and ideal states of affairs. We must be able to imagine worlds that are different from and hopefully better than this one and learn things about morality from them; otherwise, we're stuck with the limited observational opportunities of this world, and it's less clear that we could learn how to make moral progress. The problem for the realist is that once we shift over to possible worlds, it's extremely doubtful that the causal powers of moral facts and properties *there* would have any direct effect *here*. They may have indirect effects through the vehicles of cognition—this second strategy that we will explore momentarily—but that was not the proposal put forward by the realist on this response. On the causal powers strategy, moral

properties are supposed to have the ability to directly motivate agents to act. Once we add in possible worlds, the realist loses the crucial causal links between properties and observers' mental states. No one seems to grant that there is such a thing as cross-world *causation*, merely epistemic access.[63] So even if the causal power account works for moral properties in the actual world, it will fail in the more numerous cases of other worlds.[64]

An alternative strategy would be to grant that observers in the actual world merely have epistemic access to moral facts and properties in other possible worlds, but that in apprehending those facts and properties, they acquire the proper motivation. This response claims that moral facts and properties enter into mental states with mind-to-world directions of fit (e.g., beliefs) and that *other* mental states or other *aspects* of the same mental state enter in to modulate that mental content to secure motivation. Call this approach the "mentalist" strategy: in

---

[63] Suppose we understand causal powers in the actual world ($\mathbf{W_1}$) as bringing about the change from $\mathbf{W_1} \rightarrow \mathbf{W_2}$. This view, combined with a view about backwards causation, could allow that a (future) possible world ($\mathbf{W_2}$) exerts causal forces on observers in this world ($\mathbf{W_1}$) that brings about changes in this world such that the possible world is brought about. I am deeply skeptical about cases of backwards causation, so I leave this as a task for future moral realists.

[64] At the very least, it seems as though different people are drawn to different things, and it's not clear that there *is* a single property of goodness that exists stance-independently across all good things. Put slightly differently, given the diversity of tastes and pursuits observed across humans, it seems that the good is, to some extent, stance-dependent. More specifically, while we might all agree that murder is bad and undifferentiated pleasure is good, it's very unlikely we will come to such agreement regarding the relative value of specific goods. Is it really determinate whether one should become a doctor or an engineer because one profession is intrinsically better than another? This underscores the extent to which the good arises in interaction with certain creatures. It is not something simply foisted upon them; it is something relative to their goals, intentions, and general characteristics, which is just to say that it is stance-dependent. Boyd thinks that some such framework could still count as realist: "Instead of offering, for the discipline(s) in question, a single theory of epistemic contact and a single theory of error derived from one of the plausible alterative theoretical conceptions, the realist should be thought of as offering a family of such pairs of theories, one pair grounded in each of the alternative conceptions. She then should be thought of as arguing that these alternative theories of epistemic contact and of error participate sufficiently in a relationship of (*partial*) *mutual ratification* sufficiently deep that an adequate realist philosophical package can be grounded in the disjunction" (1988, 220).

effect, moral facts and properties cause moral beliefs, then desires come in to produce action. Rather than phrasing this process in terms of beliefs and desires, a rationalist response would add the framework of practical reason to secure the same result, but practical reason is no less mental, as it were, than a mental state of desire; it is a mental process instead of a state, but the key fact is that something mental is acting on a moral mental state with intentional content that is about some moral fact or property. So Hume and practical reason theorists wind up being in the same broad family of views about motivation; they disagree on the mechanics but are committed to the view that something mental must act on the moral content to produce motivation. (After all, if the causal powers strategy could actually be cashed out, it would meet Hume's demand for an explanation of how to move from is to ought.)

Like the causal powers strategy, this mentalist strategy also fails, but does so by violating the realist's claims about stance-independence. On the rationalist strategy, the motivational force of moral facts and properties is made to depend on a particular agent's volitional states. If an agent doesn't have a certain antecedent desire to come in and confer some special status on a particular moral belief, that belief and the moral facts and properties behind it, in turn, fail to produce any action. Volitional states almost always have intentional content—my desires, intentions, plans, etc. are all *about* something, and it need not be the case that that something exists. Because the mentalist strategy hinges the motivational features of moral properties on certain mental states of observers, those features turn out to be stance-dependent. This is a *limited* form of stance-dependence; the moral

properties still exist and have *most of* their nature independently of the intentional stances of observers, but the crucial portion of that nature—the part concerned with motivation—is made to depend on observers' mental states.

Moral realists are, of course, invited to show that a causal powers strategy can work for possible worlds or that a mentalist strategy does not induce stance-dependence, or, more generally, that causation can occur cross-world or even that the apparatus of possible worlds is unnecessary in the first place. But, at present, it is unclear how moral realists can account for moral mental states that are not directly about this world in a way that preserves both stance-independence and motivational force.

3.4. Conclusion

In this chapter, I showed that arguments for moral realism failed to establish the full range of claims about stance-independent moral facts and properties. Experience alone does not establish the existence of these facts and properties because our thought and talk, in virtue of intentional inexistence, can always be about things that do not exist. Moreover, other arguments for moral realism fail to establish the existence of these facts and properties. When we provisionally examined moral realism more closely, we saw that appeals to possible worlds via indirect moral judgments may be unable to account for moral progress and that it was not clear how motivation works on realist accounts. The causal powers strategy was unclear given the bounds of possible worlds, and mentalist strategies secured motivation only at the cost of stance-dependence.

Chapter 4

Metaethical Internalism, or Morals by Experience

> Out of the crooked timber of humanity no straight thing was ever made.
>
> —Immanuel Kant, "The Idea of Universal History"

In the previous chapter, I criticized metaethical methodological externalism for attempting to locate stance-independent moral facts and properties in the world. In particular, I argued that moral realists cannot argue from someone's having a moral experience to the existence of such facts and properties, and I cast doubt on the thesis that moral facts and properties in other possible worlds could motivate agents to act.

Opposed to this account are metaethical methodological internalists, who deny that metaethical questions can be addressed by discovering facts, properties, objects, states of affairs, etc. in the world. Instead, they argue, our investigation must begin from our own experience (however stipulated) of moral judgment, reflection, attitudes, and so on. Mandelbaum gives a forceful statement of this position as follows: "If the system which the metaphysician deduces is not consonant with the judgments of value and obligation which men actually make, no amount of argument will convince us that the system is valid and its metaphysical basis true….[T]he validity of a metaphysical ethics must be tested through an appeal to what one is willing to acknowledge to be an enlightened

moral consciousness" ([1955] 1969, 17–18). Though internalists may disagree on exactly what parts of (or whose) experience is to be taken as a starting point or what restrictions are to be imposed on that experience to yield appropriate moral conclusions, they agree with Mandelbaum's rule that "phenomenological description is a necessary propaedeutic to causal explanation" (25).

As part of this general methodology, internalists often defer to what Baier calls "the moral point of view." On his view, to ask a moral question is not simply to ask a question involving the use of certain moral words; it is to ask a question from within a certain framework: "when an answer of certain sort is wanted, an answer that can stand up to certain complicated tests; in other words, when the questioner wants the person questioned first to consider and then to answer the question *from the point of view of morality*" (1954, 105). In taking up this view, he claims, we are committed to viewing "human beings as equally engaged in the pursuit of their legitimate interests" (126). Accordingly, we must choose rules for the regulation of society that affect everyone alike.

What should already be clear from this brief introduction is that internalists take the results of their methodology to be far from subjective or personal reflections on the moral life. Quite the contrary, internalists (including Kantians) often hold normative views that are the most demanding with respect to impartiality toward others and indifference toward one's own personal projects. The task of this chapter will be to get clear on the commitments of the chief internalist account: constructivism. In particular, I address two questions: How do

constructivists account for moral mental states? and Can they achieve the objectivity to which they aspire?

My strategy is to show that in making sense of moral mental states, constructivists must abandon their objective pretensions and accept something more like the account I give in Chapter 6, which allows for a healthy dose of relativism. I begin in §4.1 by formulating a working account of constructivism with reference to moral mental states and the core tenets of constructivism: stance-dependence and objectivity. Following that, I consider how constructivists account for the world-to-mind and mind-to-world satisfaction conditions of moral mental states. In particular, I argue in §4.2 that Kantian requirements of moral motive provide a poor understanding of the world-to-mind satisfaction conditions for moral mental states. In §4.3, I explore the constructivist account of mind-to-world satisfaction conditions, concluding that constructivists are forced to abandon either stance-dependence or objectivity or both. For reasons I discuss in §4.4, it is also important and desirable to consider constructivism in connection with phenomenological views. I attempt a rehabilitation of constructivism using phenomenology throughout §§4.4.1–4.4.2 but find that this is ultimately unsuccessful in garnering objectivity and may still collapse into a kind of externalism.

4.1. Constructivism, Its Methods and Claims

Constructivists disagree with moral realists about the method of confirmation for moral claims. Whereas moral realists, like all externalists, attempt to establish a

match between moral mental states and stance-independent moral facts and properties, constructivists look toward the process of practical reasoning undertaken by moral agents. This approach, they claim, does not diminish the correctness of moral mental states. Thus, Korsgaard writes: "what makes the conception correct will be that it solves the problem, not that it describes some piece of external reality. Rather, as the term "constructivism" suggests, our use of the concept when guided by the correct conception *constructs* an essentially human reality—the just society, the Kingdom of Ends—that solves the problem from which the concept springs. The truths that result describe that constructed reality" (2003a, 117). Constructivists do not claim that the moral reality they discuss is constructed by us; rather, this "construction" is a rhetorical device for bringing out the substantive claims already entailed by one's experience from the moral point of view. To put the point a bit differently, the device of construction reveals the criteria for satisfying the mind-to-world satisfaction conditions of moral mental states. (We will attend to world-to-mind conditions in a moment.)

In a recent article, Street (2010) distinguishes between two kinds of constructivist accounts. According to "proceduralist" accounts, "normative truth [is understood] as not merely *uncovered by* or *coinciding with* the outcome of a certain procedure, but as *constituted* by emergence from that procedure" (365). This is the canonical form of constructivism found in Rawls' ([1971] 1999) original position. The parties to the bargaining game are described as free, equal, and rational, but situated behind a veil of ignorance that restricts knowledge of their accidental features, such as age, race, sex, etc. According to Rawls, the outcome of this

procedure does not merely *coincide* with the principles of justice; it *constitutes* them, and there is no truth about justice independent of the procedure itself.

As David Enoch (2006) argues, this kind of constructivism amounts to little more than a naturalistic reduction of normative facts to the responses of idealized agents. These objections do not reflect the spirit of procedural constructivism. As Rawls says, "Kantian constructivism holds that moral objectivity is to be understood in terms of a suitably constructed point of view that all can accept. Apart from the procedure of constructing the principles of justice, there are no moral facts" (1999, 307). A charitable interpretation of constructivism should preserve the internalist spirit of Rawls' account by resisting Enoch's proposal that constructivism is some kind of ideal observer theory in which, along externalist lines, subjects detect and respond in appropriate ways to moral facts and properties that exist independently of the mental acts of the subject.

Street offers a second form of constructivism, which she calls "practical standpoint" constructivism and I shall refer to as "point-of-view entailmentism." According to this, the subject is understood as taking up a certain point of view: the moral point of view. In so doing, he/she adopts a certain attitude: "The claim [of constructivism] is that we have an understanding of this attitude even if we do not yet understand what value itself is" (2010, 366). Using this viewpoint and the idea of entailment, "we need only make observations about what is constitutively involved in the attitude of valuing or normative judgment itself…" (367). Point-of-view entailment constructivism seems to reflect internalist methodology more prominently than proceduralism. For one, the outcomes seem to flow directly from

the subject's apprehension of the moral point of view as opposed to procedural constraints that are, in Rawls' case at least, foisted upon hypothetical bargainers. In addition, it makes clear that any moral facts and properties recognized by the constructivist are entailments (or logical constructs) out of the subject's point of view, not responses to independently existing moral properties.

To form a working account of constructivism, we must preserve these upshots of Street's formulation. Before giving that account, I want to take one more pass at the literature on constructivism, with reference to Ron Milo's statement of constructivism, which represents a sizable portion of all constructivist positions. Following Rawls, Milo brings out the rhetorical aspects of procedures and standpoint-taking quite clearly by giving constructivism a more contractarian flavor: "contractarian constructivism views moral truths as truths about what norms and standards hypothetical contractors would have reason to choose. It is true that lying is wrong just in case there is reason for human beings, from an idealized social point of view, to choose norms that prohibit lying" (1995, 186). Since all forms of constructivism are not committed to contractarianism,[65] we can omit the reference to social choice theory, yet retain the modal flavor of Milo's account.

---

[65] In an article comparing these two methods with reference to Rawls' and Scanlon's views, O'Neill concludes: "If Scanlon's account of practical reason is best read along these lines, his work might be viewed as more constructivist than contractualist, more Kantian than Rousseauian, and more constructivist than Rawls's constructivism. Paradoxically Rawls who characterizes his work as *constructivist* might reasonably be viewed as a *contractualist*; and Scanlon who terms himself a *contractualist* makes no basic use of the notion of agreement, and might well be called a *constructivist*" (2003, 330–31). This dizzying reversal invites suspicion that the theories are, indeed, distinguishable, and as O'Neill herself notes, "Seemingly they are not wholly different, and certainly not incompatible, since some writers have described themselves as both" (319).

Milo portrays constructivists as interested in what judgments subjects *would* make under certain idealized conditions. This point parallels our discussion in the previous chapter of moral realism in connection with possible worlds. But whereas realists were committed to acknowledging judgments about possible worlds in *some* cases, constructivists seem to avert to them in *all* cases by considering what judgments idealized subjects would, hypothetically speaking, make. The device of possible worlds is an attractive one for constructivists, for it embodies the imaginative stance of taking up the moral point of view or imagining the unfolding of a certain procedure. With this in mind, let us take constructivism to be the following thesis:

CONSTRUCTIVISM Call the actual world α. There is a nearby possible world **W** in which an ideal state of affairs obtains ("nearby" because otherwise moral progress is doomed from the start). When I make the moral judgment that murder is wrong or assert, "Murder is wrong," the predicate 'wrong' picks out the class of objects and events that bear the property *m* in **W**. To bear the property *m* just is to be an object of moral condemnation by a subject S in ***W***. And so on for 'right' and moral approval and all other moral predicates and their corresponding judgments according to S.

Most constructivists add a Kantian strain to this thesis by substituting the kingdom of ends for the ideal state of affairs in **W**.[66] As Kant told us, one subject there is as good as another—every rational subject "is *subject* to the moral law, but…at the same time *lawgiving* with respect to it" ([1785] 2002, 4:440)—so there can be no quarrel with these constructivists about *whose* moral judgments are being picked out; any S will do. Further, there is no possibility that S could be wrong about *m*, since S's judgment is *constitutive* of *m* and thus, authoritative. These restrictions on the valuing subject and the authority of his/her moral mental states should, in principle, vest constructivism with the objectivity Kant and others seek; we will examine whether it does shortly. For the moment, we should pause to consider whether this formulation faithfully reflects the core of constructivist accounts.

It might be objected that this thesis distorts the constructivist position because it mentions moral properties, from which constructivists have distanced themselves, along with moral realism. However, the crucial difference between the moral realist's properties and those referenced above is that the constructivist hangs the genesis of those properties on a special kind of mental content: the moral mental states of idealized subjects. As such, constructivist moral properties are stance-*dependent*—they are both mind-dependent (i.e., no mental states, no

---

[66] Rawls' (1980) classic article sketches the case for Kantian constructivism. Several later authors argue that Kantian constructivism is superior to other forms of constructivism, including Rawls' alleged formulation. Hill (1989) does so on the basis of the kingdom of ends offering a more robust framework than Rawls' more minimal original position. Tiffany (2006) maintains that constructivism only succeeds given Kant's moral psychology, including a presupposition of transcendental freedom. Timmons (2003), while not explicitly endorsing Kantianism, argues that Scanlon's contractualist constructivism commits him to relativism, which Timmons regards as unappealing.

properties) and not independent of the intentional states of subjects as we saw with realism in the previous chapter. Moreover, the constructivist does not claim that *m* is causally efficacious of S's judgment. Rather, *m* is a logical construct or theoretical posit out of S's judgment. To bear the property *m* just is to be an object of S's moral mental states.

Any further objections about constructivist ontology must amount to the claim that constructivists simply don't *have* any moral ontology. Constructivism, it is often said, is a theory about the confirmation of moral claims, not a theory of semantics or moral reality. As Street puts it: "the constructivist thinks the debate about mind-dependence…is where the most important philosophical action is. This isn't to say that the semantic task should be ignored. But it is to say we won't understand the nature of value until we settle the debate between realism and metaethical constructivism" (2010: 380). Even if constructivists have more stringent motives for examining epistemology rather than ontology, it does not count against the claim that constructivists *do* have an ontology, albeit an implicit one. As Quine (1948) told us, a theory is committed to whatever ontology is necessary to make its affirmations true. The constructivist acknowledges idealized subjects, their moral mental states, situations, objects, and events in good standing; stance-dependent moral facts and properties come along for the ride.

The constructivist's claims about stance-dependence and objectivity are a matter of some complexity. To many, they seem less straightforward than externalist accounts on which we simply aim to discover that which already exists and we hold objective views to the extent that they conform to this independently

existing reality. Constructivists, by contrast, have more indirect and subtle notions of what moral reality and objectivity consist in, and their claims require closer scrutiny.

4.1.1. Stance-Dependence

To get clear on the constructivist's notion of stance-dependence, we need to consider carefully the various stances involved in their theoretical framework. As stated above, constructivism involves at least two: S, the bearer of the moral mental state in **W**, and some creature in α who apprehends S's judgment in **W**. It is an interesting (and usually unaddressed) question *who* this creature in the actual world is—the constructivist? a modal logician? a philosopher trained in moral theory? an ordinary person?[67] I think we can safely set this problem aside, since the true focus of constructivism is the person in the idealized world having a moral mental state. Still, it must be noted, the cash value of constructivism is how well it enables people in the actual world to apprehend moral truths and act accordingly, and I will return to this point in §4.3.2.

Call O the correlate of S in the actual world—correlate not in the sense of being a doppelganger of S, but in being a sentient agent similar enough to S such that O finds S's judgments, attitudes, beliefs, etc. to be cognitively salient and possible grounds for action. According to constructivism, truth for any token moral

---

[67] This question reflects Baier's (1954) consideration of a person's contexts in asking the question, What shall I do? As Baier notes, that person could be a moral agent, asking questions with a view toward doing something, or a moral critic, a critic of morality, or a reformer.

judgment about a particular object or event will consist in O's recognition of the following constructivist fact:

F In **W**, S judges that object or event to be an object of approval (condemnation).

As S encounters new moral situations and has corresponding moral mental states, O will have opportunity to observe F1, F2, and so on, and the set $\{F_{i-j}\}$ will comprise the set of all moral truths. The correctness or incorrectness of O's moral mental states will depend on this set. In other words, the mind-to-world satisfaction conditions of O's moral mental states will be determined by the set, even though, in this case, the "world" is not the actual world, but an ideal one. With respect to O in α, F and $\{F_{i-j}\}$ are *not* stance-dependent. What garners condemnation or praise from subjects in the kingdom of ends does not depend on *our* attitudes in *this* world about objects or events or idealized subjects. On the contrary, from *our* perspective, the constitution of F is stance-independent. The crucial point for constructivists is that *m*, the moral properties in question, are constituted by S's judgment—that is, they are stance-dependent with respect to S in **W** and that, from S's standpoint, F and $\{F_{i-j}\}$ are stance-dependent.

Let us pause for a moment to bring home the similarities and differences between moral realism and constructivism. Both acknowledge that subjects in the actual world are observers of specific kinds of facts. For the realist, these facts are stance-independent moral facts, which have their basis in the stance-independent moral properties of actions, institutions, objects, and events in the actual world or ideal world(s). For the constructivist, these facts are stance-independent *non*-moral

facts about the moral mental states of subjects in an ideal world, out of which are constructed stance-dependent moral properties in the actions, institutions, objects, and events. According to the realist, in a world without observers, there would still be moral facts and properties (though they might not cause, say, pleasure and other mental phenomena). According to the constructivist, in a world without moral concerns—that is, without subjects and their moral mental states or without at least observers able to engage in the rhetorical device of *imagining* such concerns—there would be no moral facts and properties.

Given the constructivist's emphasis on observing the mental states of subjects in an ideal world, it is fair to say that the theory is partially externalist in the sense discussed in the last chapter. However, it must be noted that constructivists are not externalists with respect to moral properties. They do not believe they can be detected or perceived as the realist does. Instead, they are logical constructs or theoretical posits out of S's moral mental states. Since the path to apprehending moral properties, according to constructivism, proceeds from a person's experience (i.e., S's experience), the theory is rightly regarded as internalist.

Two questions loom: How do people in the actual world access the set of constructivist truths (i.e., how does O in α access F1, F2, and so on)? and What is gained by the constructivist averting to an ideal world? With respect to the first question, apprehending F or $\{F_{i-j}\}$ is no more difficult in principle than apprehending any other fact about a possible world. If you think someone can know (or try to know) what his/her life would have been like had he/she become a physicist rather than a philosopher, you should also agree that someone can know

(or try to know) what kinds of judgments an idealized subject in the kingdom of ends would make. Questions of epistemic access are tricky, but the point here is that constructivists face no special problems on those grounds.

So far, so good. But we cannot forget that the facts in question concern *moral* mental content. As we saw in the previous chapter, externalists had problems showing that properties in other worlds could cause *us* to act. All of us in the actual world are externalists with respect to F; *we* want to know what, *independently* of our attitudes, beliefs, epistemic and moral shortcomings, etc., idealized subjects judge to be right and wrong. In **W**, moral facts and properties are stance-dependent, but from α, F is stance-*independent*. Regardless of what *we* think about S or the situation with which he/she is presented, S makes the same judgment in **W** and we simply apprehend that fact. Even though F is a non-moral fact (i.e., it simply describes what judgments S makes), it does concern moral mental content (i.e., what things S extends judgments of approval and condemnation toward). At least one problem for constructivists will be explaining how apprehension of F motivates people to perform morally right actions any more than the realist's stance-independent moral facts and properties motivate right action. The motivation problem we explored in the last chapter seems to apply in full force to constructivist accounts.[68] I am willing to set aside this particular problem so that we can deal with more interesting

[68] I think a response here is possible along the lines of practical reason: roughly, that our process of practical reasoning is similar enough to the process of practical reasoning by S in *W*, so, by analogy, what S finds to be motivating is similarly motivating for us. Kant bases this similarity on the rational agency of subjects and observers, but given the variance in rational abilities found among actual humans and, indeed, uncertainty about what the concept of rationality even amounts to, it would be best to find other foundations. I explore such foundations in Chapter 6.

aspects of constructivism, including why the constructivist shifts the analysis to **W** in the first place.

One might think that subjects in **W** are somehow epistemically better situated to form the moral mental states that they do. Street is quick to caution that constructivism is not a form of ideal observer theory. It is just not the case that S is better positioned to receive causal inputs than we are or that S is somehow causally stationed in an advantageous way that enhances this process, perhaps by removing complicating factors. This reading, as Street correctly points out, distorts the constructivist position because it grants causal processes (and with them, I would add, externalism) the center stage and removes the stance-dependence thesis for which the constructivist was plumping: "Standard ideal response reductions introduce extraneous elements by making facts about what is valuable hostage to the outcome of irrelevant causal processes. According to constructivism, in contrast, normative questions aren't question about what would emerge from any *causal process* (whether real or hypothetical), but rather questions about what is *entailed* from within the standpoint of a creature who values things" (2010, 374). Moreover, it should not be thought that S is somehow just lucky—epistemically or otherwise—since this would garner little moral credit from Kantians or others moral theorists interested in the agent's intentions and moral motive. As I argued in Chapter 2, our metaethical analysis should be as neutral as possible between normative theories, so we don't want to rule out certain normative theories in our interpretation of constructivism. And anyway, we could imagine a subject in the actual world being similarly lucky and then there would be no reason for shifting

the analysis to **W**. So the luck interpretation fails to explain the special role that the ideal world plays in the constructivist account.

Rather than imagining S as somehow epistemically fit or lucky, we would do better to imagine S as *morally* better off. In taking up the process of practical reason, S is better able to derive the proper conclusions from it and apply those conclusions to constitute the constructivist fact F in **W**. This emphasis on standpoint—S's superiority in taking up the moral point of view—underscores the stance-dependence of constructivist accounts. Naturally, one might wonder how S's superiority helps *us*; what good is it to us that S can better take up the moral point of view and form appropriate moral judgments from within it? At the very least, the purchase of the constructivist machinery seems to buy us an opportunity to observe S and his/her moral mental states. In doing so, we learn more about the moral point of view and we begin to understand the rule on which S acts in forming his/her own moral mental states—that is, we start to learn what is entailed by the moral point of view and the standards of correctness for engaging in the activity of practical reasoning. By analogy, we could imagine a rookie player, through the rhetorical device of an ideal athlete, observing the actions of an ideal athlete and his/her judgments about which actions are appropriate from within the standpoint of the game, and the novice player, by degrees, learning the rules of the game. Similarly, in observing S and apprehending F, the facts about his/her moral judgments, we come to learn about stance-dependent properties and the mind-to-world satisfaction conditions for our own moral mental states.

Another note on observation and possible worlds: Where the constructivist world **W** is concerned, the ideal world must be roughly similar to the actual world, otherwise it would not have any relevance for our actions in this world. However, in stipulating other features of the world and its subjects, the constructivist, in essence, determines what follows from the moral point of view in that world—what Street referred to as "entailment." So in discussing "observation" of **W**, we are not discussing a process like scientific observation where we may be antecedently ignorant of the phenomenon that is occurring or about to occur. Instead "observation" stands in rhetorically for the devices of imagination and entailment, and serves only to bring out the hint of indirect externalism (and hopefully objectivity) that constructivists endorse.

4.1.2. Objectivity

Objectivity is normally discussed in one of two ways. In one sense, objectivity requires that one's beliefs, judgments, etc. correspond to an independent standard external to the believer, judger, etc. This notion of objectivity has obvious affinities with a correspondence theory of truth, which accords truth-values to beliefs, judgments, etc. based on whether they reflect the mind-independent world. The other notion of objectivity, often discussed in contrast to the correspondence notion, treats objectivity as the convergence of all observers in the limit of inquiry. A belief, judgment, etc. is objective just in case it is what all observers would agree upon given infinite time, resources, opportunities for observation, ideal conditions, and so on.

Constructivists are bound to reject the first accounts based on their thesis of stance-dependence. On their account, there is no independent moral reality against which the idealized subject's moral mental states may be queried for correspondence. However, they do not straightforwardly embrace the second understanding either. On the Kantian possible worlds construction we have been considering, a *single* subject's moral judgment is necessary and sufficient for determining the standards of correctness for all moral judgments of the same type, say, those concerning murder. If we abandon this Kantian line and say that *some* subjects' moral mental states are relevant, we introduce further problems about how our observer in the actual world picks out the right subject(s) to observe in the ideal world. Indeed, it would be as if I had to know who was making the correct moral judgments before I knew what correct moral judgments were. There is obvious reason to adhere to the legislative reading of the categorical imperative, and anyway, we could imagine **W** as containing only one subject in the first place. In any case, the notion of having multiple observers converge on similar judgments will not be available to the constructivist, given that the subjects are *active* in their contribution or construction of the properties in question and not mere passive perceivers of them.

What, then, does objectivity amount to for the constructivist? Rawls (1993) gives five essential elements of objectivity that will help frame our discussion. On his view, a conception of objectivity

(1) "must establish a public framework of thought sufficient for the concept of judgment to apply and for conclusions to be reached on

the basis of reasons and evidence after discussion and due reflection;" (110)

(2) "aims at being reasonable, or true, as the case may be;" (111)

(3) "must specify an order of reasons as given by its principles and criteria, and it must assign these reasons to agents, whether individual or corporate, as the reasons they are to weigh and be guided by in certain circumstances;" (111)

(4) "must distinguish the objective point of view—as given, say, by the point of view of certain appropriately defined reasonable and rational agents—from the point of view of any particular agent, individual or corporate, or of any particular group of agents, at any particular time;" (111) and

(5) have "an account of agreement in judgment among reasonable agents" (112).

In order to satisfy these constraints, the constructivist must demonstrate that moral experience contains within it some procedure that would be agreed to and could be employed by all other agents to reach a set of substantive conclusions about reasonable (correct) moral judgments. This procedure *qua* reasoning process will specify principles and criteria the agents are to observe in making their determinations.

Within the possible worlds framework we have been exploring, constructivists should say that a subject's judgment (in the actual world) is objective if it:

(1′) adheres to the constructivist framework;

(2′) aims at agreement with the judgment of the (relevant) subject(s) in an ideal world;

(3′) specifies a procedure for accessing that ideal world and the (relevant) subject's judgments within it;

(4′) embodies cognitively salient reasons or principles that could be articulated by or abstracted from the (relevant) agent; and

(5′) submits (1′)–(4′) to scrutiny of public examination.

Constructivism could fail, then, on several points. First, as I already hinted, it could fail in specifying how observers in the actual world access the moral mental states of subjects in the ideal world.[69] Presumably, the window to this ideal world is our own moral experience of taking up the moral point of view and engaging in the process of practical reason. The constructivist will need to specify this experience in detail to demonstrate that, indeed, there is such access. If such experience turned out to be extremely varied, we would have reason to doubt that agents in this world could actually access the world at which constructivist accounts aim. I take up consideration of these objections in §4.3.1.

Second, constructivists will need to give a robust account of the moral mental states of idealized subjects. This account will be important in assuring us that constructivist moral facts and properties are, indeed, stance-dependent. There are interesting and very important objections to the effect that these properties

---

[69] By 'access' here and elsewhere, I mean 'construct' or 'stipulate' in accordance with Kripke's notion of possible worlds. For further discussion, see Note 59.

are, in fact, stance-independent and that constructivism collapses completely into a variety of externalism. I develop these objections in §4.3.3.

Finally, there are deep affinities between phenomenology and constructivism. Though I do not attempt to sketch them in full here, I do offer responses to the first two worries along phenomenalist lines. These concepts, in turn, invite new problems, which I think underscore some of the deep and pervasive difficulties with constructivism. §4.4 presents these considerations. Before considering further issues involved in objectivity and mind-to-world satisfaction conditions, it will be helpful to consider the constructivist's views of world-to-mind satisfaction conditions.

4.2. Satisfying Rules and the Demands of Morality

On the constructivist account we have been examining, idealized subjects encounter objects and events, form moral mental states, and perform verbal and non-verbal behaviors that provide evidence for us observers of what stance-dependent moral facts and properties the subjects establish in the course of this exercise. Beyond this minimal account, most constructivists also endorse Kant's views of moral action, which adds a further constraint: that agents must act from a regard for the moral rule, which requires an awareness of the rule and the appropriate intentions to act on that rule.[70] To put this a bit differently,

---

[70] It might be objected that moral motive is a requirement on *us* in the actual world and not idealized subjects in the ideal world. We simply learn from them the content of morality, and other arguments must be marshaled in support of moral motive. I think this response is unattractive for two reasons. First, I don't see why *idealized* subjects, by the Kantian's own lights, won't act from moral motive. Second, I'm not sure what a constructivist formulation of the argument for moral motive might look like. Are we to imagine *two*

constructivists' idealized subjects form their moral mental states from a particular standpoint—the moral point of view—and the Kantian adds the requirement that they are aware of this process in the sense of recognizing the rules entailed by that standpoint and intending to act upon them. This requirement forms the core of Kantian morality: "*Morality* is thus the relation of actions to the autonomy of the will….A will whose maxims necessarily harmonize with the laws of autonomy is a *holy*, absolutely good will….The objective necessity of an action from obligation is called *duty*" ([1785] 2002, 4:439). While imposing awareness and specific intentions on the mental states of idealized subjects may seem innocuous enough, this is no minor addition. It places a fundamental constraint on the world-to-mind satisfaction conditions for moral mental states, and substantive normative conclusions about rewards and punishments follow from it. To examine this point more critically, we need to consider the connection between rules, mental states, and behavior in detail.

Satisfying a rule is more than just having it be the case that such and such has been brought about. After all, this state of affairs may be brought about quite accidentally. For example, a rule about watering my garden is not satisfied merely if the plants receive water, for my kindly neighbor might have watered the plants for me, or it might have just rained. For *me* to fulfill the rule—indeed, for *it* to be fulfilled—it must be the case that *I* have watered the garden, not someone or

---

possible worlds, one with idealized subjects who determine the content of morality and another of idealized agents who apprehend that possible world and consciously follow whatever morality winds up being? This proposal is decidedly un-Kantian, for it splits the faculty of perception or judgment from the faculty of action and practical reason, and Kant's most important contribution seems to be his view that the two constitute a unity (in moral knowledge and elsewhere).

something else. I must have done something to bring it about that the plants are watered. This "something" stands in need of further explanation.

As we saw before, satisfaction conditions for behavior often contain a certain amount of ambiguity. A rule may pick out acts at one level of description, but at different levels, those acts may be realized in different ways. To return to our example, fulfilling a rule about watering my garden does require that I do something to give my plants water, but it is ambiguous between different modes of watering. I can use a watering can or I can use a hose. So long as the plants receive water and I am the one who gives it to them, it doesn't seem to matter *how* I give it to them. Notice, however, that I must still give them *water*, as opposed to, say, giving them plant food. The rule here picks out a certain class of behaviors (i.e., waterings), but it does not contain any fine-grained description of what behaviors count as waterings; it allows me to fulfill the demand in some appreciably indeterminate way.

But could it really be that *any* watering will do? Suppose I trip and spill a glass of water on my plants. My spilling brings it about that the plants receive water, and it is *my* spilling, something I do. But it hardly seems that I have just watered my plants. So what distinguishes the two cases? What makes deliberate "spillings" relevantly different from accidental spillings? Remember that the rules under consideration are rules of *action.* Spilling water accidentally isn't an action *per se*, but simply a case in which a certain state of affairs (i.e., water being spilt) is brought about. My spilling involves no real volitions on my part, no intentions to use a hose or fetch a watering can. My *body* may move in a certain way, but I am

not the cause of accidentally spilling water in the same way that I am the cause of pouring water deliberately out of a spout. The former involves no intentions on my part and is compatible with my having no intentions to spill the water (perhaps I am thirsty), while the latter, *ceteris paribus*, involves the use of some concepts (e.g., HOSE, WATER) or recognitional capacities (e.g., that a hose is nearby, that doing such and such will put water near the plants). With accidental spillings, a state of affairs has been brought about, but a rule has not been satisfied.

Sellars puts this another way by noting that actions *are* events—that is, they are processes and states of affairs in the world that have a non-action core—but unlike other events, actions involve a person and are "*ceteris paribus* caused by the person's willing to do them" (1973, 195). As such, actions have performance conditions, which are satisfied, in part, by certain predictable mental states and processes. These "Volitions are *conceptual* episodes….the sort of episode which [are] manifested, *ceteris paribus* (thus in the absence of paralysis and in the presence of favorable circumstances), by a doing of A, e.g. a raising of the hand" (1967, 177, my emphasis). Accidents do not fulfill the performance conditions required for satisfying rules because they lack the corresponding conceptual episodes. When those episodes are present, one usually performs the corresponding behavior, as when, in the parallel case, one candidly thinks out loud. When those episodes are absent, one can be *present* in an event, but one fails to *perform* that event (i.e., to act or to commit an action).[71]

---

[71] That is Sellars' quick answer to how behaviors fulfill rules. Elsewhere (1954, 1974), he presents a more detailed account that avoids a paradox that arises from the quick answer. The traditional account

So far, we have sketched an account of rules and behaviors that distinguishes actions from accidents and successful behaviors from unsuccessful ones. Before leaving this discussion, we need to make one more distinction, and it involves how agents act with regard to rules. One can fulfill a rule in one of two ways: with an awareness of the rule in mind, or without such awareness in mind. This distinction may be cashed in terms of the conceptual episodes that precede deliberate behavior. Those episodes may consist, in part, in recognitional capacities and concepts, and they may contain, in addition, a thought of some rule that specifies an action to be performed. To return to our example, there are two ways to water one's garden in the relevant sense. One might water one's garden as a matter of sheer habit (e.g., at 7 o'clock on alternating evenings), or one might water one's garden quite soon after thinking that the plants ought to be watered or that I should be the one to water them or even after reasoning from the former to the

---

holds that learning to use a language is learning to obey the rules for the use of its expressions. However, it may be objected that a rule for an expression is expressed in a metalanguage which contains the object expression. Therefore, learning a language presupposes learning a metalanguage and that presupposes a metametalanguage and so on. The simple solution to this problem is to talk about *conforming* to rules rather than obeying them. Obeying a language rule *would* require knowing a metalanguage, and a metametalanguage, and so on. Conforming does not; it is simply doing A when the circumstances are C. At this step, Sellars points out a false dichotomy between *merely conforming* to rules (doing A in C where these doings "just happen" to contribute to the realization of a complex pattern) and *obeying* rules (doing A in C with the intention of fulfilling the demands of a system of rules). This distinction relies on the claim that unless the agent conceives of the system, his/her conformity is merely accidental. Sellars points out that this is true in the non-intended sense of "accidental," but there is another sense in which "accidental" means non-necessary. In contrast to this sense, there can be an unintended action that is necessarily related to a system of acts (e.g., bee dancing). This leads Sellars to a distinction between *pattern-governed* behavior and *rule-obeying* behavior. The former is to become conditioned to arrange perceptible elements into patterns and to form more complex patterns and sequences from these. The latter presupposes the former, but also contains a game and a metagame that contains the rules obeyed in playing the game as a piece of rule-obeying behavior. Pattern-governed behavior involves positions and moves that *would* be called formation and transformation rules in its metagame if it *were* rule-obeying behavior. Sellars' aim is to show that we can engage in pattern-governed behavior to play a game and still be doing it "because of a system" without having to *obey* rules and hence play a meta<u></u>language game (and enjoin the regress). "Pattern-governed" and "rule-obeying" correspond roughly to the satisficing–satisfying distinction I am about to introduce, though rule-obeying does not require an awareness of rules in the way that satisfying does.

latter. In the first case, recognitional capacities and concepts are present in the preceding conceptual episode. If they were not, I would fumble about and my finding a watering can or hose would be a matter of sheer luck, as would my directing the water in the neighborhood of the plants. The conceptual episode that precedes the second case also involves recognitional capacities and concepts. But, in addition, it contains some form of a rule (propositional, if you like) about watering one's garden that expresses that rule. The difference between the two cases lies in their conceptual episodes, the volitions that *ceteris paribus* lead to successful behavior of a certain type.

For our purposes here, let us distinguish two relations between rules and pieces of behavior. Let us say that a rule has been *satisfied* when the conceptual episode contains, in part, a thought of some rule that subsequently issues in an agent successfully performing a behavior specified by that rule. Let us say that a rule has been *satisficed* (minimally satisfied) when the conceptual episode does not contain any such thought.[72] This distinction allows us to say that habitual waterings satisfice certain rules, while waterings that stem from tokenings of water-rule thoughts satisfy those rules.

What we have to consider now is whether the constructivist framework provides resources for showing that idealized subjects actually *satisfy* moral rules (with awareness and intention, as Kantians claim) or whether they merely satisfice

---

[72] For a broader discussion of satisficing and moral theory, see Byron (1998, 2004).

them. Kant suggests at several points that behaviors originating from moral motive are not, in the long run, coextensive with other kinds of behaviors:

> From love of humankind I am willing to admit that even most of our actions are in conformity with duty; but if we look more closely at the intentions and aspirations in them we everywhere come upon the dear self, which is always turning up; and it is on this that their purpose is based, not on the strict command of duty, which would often require self-denial. One need not be an enemy of virtue but only a cool observer, who does not take the liveliest wish for the good straightaway as its reality, to become doubtful at certain moments (especially with increasing years, when experience has made one's judgment partly more shrewd and partly more acute in observation) whether any true virtue is to be found in the world. And then nothing can protect us against falling away completely from our ideas of duty and can preserve in our soul a well-grounded respect for its law… ([1785] 2002, 4:407–8)

Kant claims that any motivation *other* than respect for the moral law itself will fail to issue in moral action at some time or other, and, accordingly, only respect for the moral law can provide a firm groundwork for morality and garner the moral praise and respect of others. As the foregoing discussion makes clear, these empirical assumptions are misguided. Agents need not act with an awareness of a rule in mind in order to regularly satisfice that rule. One we admit this point, it seems that Kant is only haggling over the price—that is, *how frequently* would other motivations issue in moral action, or is it *certain* that they would do so.

To adjudicate this point, we must recall that the idealized subjects in question are stipulated by the possible worlds construction we have been employing. There is nothing, in principle, that forbids us from stipulating that idealized agents *always do* exhibit feelings of sympathy, generosity, benevolence, and the like. It is a substantive assumption that they (or we) will occasionally lack these feelings and that moral motive is the *only* guarantee of right action. If idealized subjects need not have any respect for the moral law in order to perform moral actions (even with *certain* regularity!), why hold their observers (i.e., humans in the actual world) to any stricter standard? If, *ex hypothesi*, it is not part of the *content* of the moral judgments of idealized subjects that their actions proceed from respect for the moral law, there is no reason to think that such respect or awareness forms part of the world-to-mind satisfaction conditions of moral mental states.

Moreover, we must recall that the only evidence that we have are observations of verbal and non-verbal behaviors. On our possible worlds interpretation of constructivism, idealized subjects encounter moral situations, form moral mental states, and, in the process of doing so, confer stance-dependent moral properties on objects and events. No constructivist we have seen claims that we have direct access to the mental states of these idealized subjects—indeed, how could we, or what would be the point of using the mental states of idealized subjects as opposed to our own. Instead, we appear to have access to their overt behaviors, both verbal and non-verbal, in much the same way that we have indirect access to others' thoughts in the actual world through these behaviors. Thus,

readers of the constructivist literature find the idealized subjects engaging in bargaining behavior, vetoing principles they find repugnant, making moral proclamations, expressing their moral judgments, and so on. Based on this behavioral evidence alone, it is impossible to distinguish between actions that are performed in conformity with moral law and actions that are performed out of a respect for it. So even *if* idealized subjects always acted from a respect for the moral law—even if they have as part of their mental states an awareness of particular rules—*we* could never know that they do.

I am not saying that behavioral evidence *cannot*, in principle, provide support for claims that subjects have awareness. On the contrary, I think that may be the best (or only reliable) evidence for the attribute of mental states. I am arguing that Kantians have not done enough work to show that is the case here, and that the constructivist framework does not necessarily imply moral motive as part of the world-to-mind satisfaction conditions for moral mental states. A full rehabilitation of Kantian moral motive in the terms discussed here is beyond the scope of this project. In §4.4.2, I will make a brief and limited attempt to do so using phenomenal properties, but they, too, prove inadequate. Having discussed this issue of world-to-mind satisfaction conditions, let us now turn to mind-to-world satisfaction conditions and the constructivists' attempt to secure an objective framework for moral theory.

4.3. Objectivity and the Moral Point of View

Constructivists claim that, in taking up the moral point of view, we gain epistemic

access to—or rather construct or stipulate the conditions in—the idealized world that forms the basis of their stance-dependent moral facts and properties. It is crucial, then, for us to get clear on what constructivists mean by "the moral point of view." In other words, we must consider what grounds the constructivist has for identifying one ideal world **W** out of the set of all possible (ideal) worlds and for assuring us that idealized subjects form such and such moral mental states in that world.

The next section (§4.3.1) casts doubt on the commonality of this point of view in light of empirical facts about the variety of moral experience. Given cultural variation in norms and customs, it seems very implausible that there is some common viewpoint from which all observers could proceed to converge on **W**. Absent that foundation, it is unclear that constructivists can achieve the objectivity they desire. Following that, I take up a brief consideration (§4.3.2) of what an ideal subject would look like and how that compares to our own experience. Finally, I develop a supervenience objection to constructivism (§4.3.3) that suggests constructivism actually collapses into a kind of externalism. The common thrust of these three sections is that the constructivist account of mind-to-world satisfaction conditions is inaccurate, too demanding, and ultimately reliant on stance-*independent* moral facts and properties.

4.3.1. The Varieties of Moral Experience

Herodotus' *Histories* gives a classic example of the varieties of moral experience: "If one were to order all mankind to choose the best set of rules in the world, each

group would, after due consideration, choose its own customs; each group regards its own as being by far the best" ([c. 440 B.C.E.] 1998, 185). In a famous example, he cites Darius, king of Persia, who asked a group of Greeks how much it would cost for them to eat the corpses of their fathers. Horrified, the Greeks say they would not do it for any amount of money. Darius then asks several Callatines how much money it would take for them to burn the corpses of their fathers. Horrified, the Callatines say they would not do it for any amount of money. Herodotus concludes, "I think Pindar was right to have said in his poem that custom is king of all" (186).

Faced with this example, the constructivist may respond that disagreement about funerary rights does not count against the objectivity of their moral claims because funerary rights are not a moral issue in the first place. This response is *exactly* what one would expect from anyone who held a substantive view of moral mental content. The bounds of the moral are demarcated with reference to certain kinds of mental content, and the constructivist here asserts that funerary rights fall outside of that domain. As I argued in Chapter 2, this is an unattractive view of moral content, for it rules out *by fiat* too many instances where genuine moral questions seem to lurk.

Though we cannot draw the conclusion that *all* experience is relative to culture—there are surely some norms about truth-telling, contracts, property, and so on that are common touchstones across cultures—it should be quite obvious that there *is* variation in moral experience across cultures; simply recall Williams' notion of thick moral content, according to which certain moral terms or contents

acquire their meaning from particular cultural contexts. Regardless of whether any of these viewpoints is superior, it is an empirical fact that moral experience varies from culture to culture. Given this variation, what entitles the constructivist to stipulate one ideal world over any other?

In principle, the constructivist framework should be neutral between competing normative theories. That is, it should be equally possible to be a Kantian constructivist as it would be to be a utilitarian constructivist. The difference between them lies in how they pick out or describe the ideal world. For the utilitarian, it is one in which idealized subjects form moral mental states that aim at producing the greatest amount of happiness for the greatest number. For the more common Kantian constructivist, the ideal world is one in which the kingdom of ends is instantiated and each idealized subject respects other persons as ends in themselves by following the law of the categorical imperative. Now it may turn out that some of these worlds are incoherent or ones we wouldn't wish to adopt as ideal—and I think this is precisely Kant's point about utilitarian worlds. Still, we should be able to at least *examine* such worlds using the constructivist framework in the same sense that we can, on realist accounts, entertain the hypothesis that the property of goodness is identical with pleasure or with moral motive, antecedent to and in the course of our investigation of those properties.

The point is that any constructivist owes us an explanation of *why* the ideal world he/she stipulates as part of the constructivist framework is a worthwhile one, or in keeping with the moral point of view, or one we should care about. Stipulation without this explanation will beg the question against alternative

normative theories, probably along the lines of substantive accounts that I criticized in Chapter 2. But on what grounds can the constructivist offer this explanation? If constructivism is to be regarded as a metaethical theory in full standing, it had better pursue this explanation along internalist lines; relying on another metaethical framework at this level will undermine the autonomy and independence of constructivism as a metaethical theory. The problem is precisely that internalism forces the constructivist (and everyone else, on that framework) from starting from his/her own moral experience, which, as I have been arguing, can vary wildly. The objective pretensions of constructivism seem to be in direct tension with its internalist methodology.

As a matter of course, constructivists seem to get their objectivity by stipulating certain formal constraints on moral mental states (e.g., universalizability). Defending these foundations is extremely difficult work, but it's not work that externalists or prescriptivists or interactionists have to accomplish in the same way—the difficulty is precisely the result of the constructivist's internalist leanings. Constructivists might do well to endorse Henry Sidgwick's stated goal for his own discussion of intuitionism: "by reflection on the common morality which I and my reader share, and to which appeal is so often made in moral disputes, to obtain as explicit, exact, and coherent a statement as possible of its fundamental rules" ([1907] 1981, 216). If Rawls, for example, announced that he wished only to tell us *postindustrial Westerners* what *we* already believe about justice, we might agree that he had succeeded in his task. But constructivists, especially Kantians, take themselves to be saying something objective about moral philosophy, and often

something that is incapable of being overturned by the staunchest skeptic or the most compelling evidence about the actual consequences of any of their schemes.

This hubris reflects a deficiency of all internalist views: reliance on private experience at the expense of public evidence. Internalism takes as its starting point individual experience and attempts to systematize that into an account that is true of all other subjects (i.e., objective). Faced with data about the varieties of experience, the internalist may try to beg off counterexamples by denying that they are what they purport to be—in this case, moral claims—or he/she may refine the account by relativizing it and reducing its scope in more reasonable ways. At best, the constructivist's moral point of view is more subjective or personal than he/she wants to admit; at worst, it is question begging.

4.3.2. The Idealized Subject

In the previous section, I raised concerns that the ideal world specified in the constructivist framework could not be justified on the basis of moral experience. Too much variation threatened the objective pretensions of constructivism. On a related note, I want to pause briefly to consider the idealized subject that is also part of the constructivist framework. These two criticisms are related in the sense that, in taking up the moral point of view and stipulating the ideal world in constructivists, we similarly stipulate facts about idealized subjects in that world, including the moral mental states they are bound to reach through entailment from the moral point of view.

The most common examples of constructivist accounts portray these subjects as overly rationalistic, aheadonistic, performing what some would consider supererogatory actions, and so on—problems best summarized by David Velleman's discussion of "the guise of the good:"

> The agent portrayed in much of philosophy of action is, let's face it, a square. He does nothing intentionally unless he regards it or its consequences as desirable. The reason is that he acts intentionally only when he acts out of a desire for some anticipated outcome; and in desiring that outcome, he must regard it as having some value. All of his intentional actions are therefore directed at outcomes regarded *sub specie boni*: under the guise of the good.
>
> Our moral psychology has characterized, not the generic agent, but a particular species of agent, and a particularly bland species of agent, at that. It has characterized the earnest agent while ignoring those agents who are disaffected, refractory, silly, satanic, or punk. I hope for a moral psychology that has room for the whole motley crew. (1992, 3)

What Velleman takes issue with here is the extremely narrow conception of idealized subjects on many accounts. Susan Wolf has made related criticisms in her discussion of "moral saints," those individuals who are as morally worthy as can be, even at the cost of their own personal projects (e.g., gourmet cooking) and happiness. Wolf criticizes moral sainthood as an ideal on the grounds that "when

such ideals are present, they are not ideals to which it is particularly reasonable or healthy or desirable for human beings to aspire" (1982, 433).[73]

The common thrust of these criticisms is that idealized subjects are perhaps often not idealized moral agents but idealized *super*moral agents. They are unlike us in respect of their psychology, selfless drives, and unwavering pursuit of the right and the good. One wonders, then, what wisdom the experience of such agents would hold for lesser beings, such as ourselves. Just as our own moral experience may be too varied for constructivist convergence on a single moral point of view, perhaps these idealized subjects are so foreign to our own nature that any lessons learned from them would have little purchase in the actual world. This narrow-minded conception of moral agents is not an inherent feature of constructivism—we can certainly imagine better, but still flawed versions of ourselves who are simply more skilled at practical reasoning, more patient in deliberation before forming moral mental states, and more disciplined in acting upon them. If constructivists wish to take internalism seriously, they would do well to attend closely to the nature of their moral experience and seize upon idealized agents who are not beyond humanity in the relevant moral senses, but idealized

---

[73] Interestingly, she suggests that the moral point of view is not the be-all end-all of how to live one's life: "alternative interpretations of the moral point of view do not exhaust the ways in which our actions, characters, and their consequences can be comprehensively and objectively evaluated. Let us call the point of view from which we consider what kinds of lives are good lives, and what kinds of persons it would be good for ourselves and others to be, *the point of view of individual perfection*….[I]t provides us with reasons that are independent of moral reasons for wanting ourselves and others to develop our characters and live our lives in certain ways. When we take up this point of view and ask how much it would be good for an individual to act from the moral point of view, we do not find an obvious answer." (437).

versions of the moral agents we are and find ourselves among. I leave this as further work for constructivists.

4.3.3. The Supervenience Objection

In the previous two sections, I cast doubt on the constructivist apparatus in virtue of the diversity of moral experiences found in the actual world and the (possibly) stark divergence between those experiences and the ones had by idealized subjects. Suppose we grant the constructivist his/her stipulation of ideal world in question (e.g., specify that we are looking at a possible world where the kingdom of ends is instantiated), that the idealized subject's moral mental states are indeed compelling for us, and that the stance-dependent ontology implied so far is acceptable to constructivists. What remains in question is the actual moral mental states of these idealized subjects and the referents of their judgments. What *things* do idealized subjects judge to be right and wrong? Is it indeed the case, as the constructivist claims, that these judgments are not answerable to some external reality in the ideal world? Finally, what are the mental processes by which these subjects take up the moral point of view and come to act in accordance with it?

In §4.3.1 we explored the ways in which variation in moral experience might threaten constructivism. In this section, I want to explore quite the opposite worry. Let us assume that S makes stable value judgments and that they, together with some conditions on O's observations (e.g., that O not be hallucinating), do meet the criteria for objectivity we examined in §4.1.2. Presumably, S's moral judgments will exhibit some kind of regularity or constitute a pattern, otherwise the subject's

moral mental states concerning wrongness could just as well be caused by red apples as they could be by rape. Consider the case of torture. Suppose S judges a single instance of torture to be wrong. Then, we can wait and see what happens the next time S considers torture. After seeing enough judgments that torture is wrong, we may be tempted to conclude that, for S, torture is always wrong, which is just to say that torture has the (stance-dependent) property of wrongness. Now suppose S considers a ticking bomb case in which the lives of millions could actually be saved if we tortured a person for information about the whereabouts and method for disarming the bomb and this information was unavailable anywhere else. In this case, S might judge torture to be permissible. This single instance would not severely threaten our account of S's judgments that torture is always wrong, for we could simply say that all cases of *unnecessary* torture are wrong, where all cases are unnecessary except ticking bomb cases. The point is that once we have observed enough of S's judgments—however complex—we can always construe them as exhibiting regularity or following some sort of pattern; we just adjust the grain of the description of the action or class of actions in question to account for relevant counterexamples. And so on for murder, abortion, and the like.

Now, the objection goes, once we have seen enough cases and start to learn the pattern of S's assignment of wrongness to certain situations, we can begin to look for other patterns across these situations. Not only will they belong to the set of things with the moral property *m* (the stance-dependent property that is a logical construct out of S's judgments), but they may also belong to the set of

things with the non-moral property *n* (some other property, such as causing unnecessary pain). It is hard to say, antecedently, what such properties might be, but some candidates might include being physical events, being extended in time, or, more interestingly, involving people and pain caused to them. Once we have grasped one such non-moral property *n*, we can attempt to predict S's judgments about which things are *m* based on which things are *n*. This process may be slow, taking many attempts and failed predictions about which things will be *m* and requiring many revisions in *n* until our predictive ability gains accuracy. But let us grant that after enough time, there are some properties $n_1, n_2, \ldots, n_n$ that allow us to predict what things S will assign *m* to.

At this point, it will begin to look as though S's judgments are merely responses to the presence of $n_1, n_2, \ldots, n_n$ and that S's judgment *supervenes* on these non-moral properties. This possibility amounts to a *reductio* of constructivism, for it turns what appeared to be stance-dependent properties into *responses* to stance-*independent* properties—and non-moral ones at that! S's judgment is no longer as authoritative or important as it once seemed. It is not *constitutive* of *m*; it is merely a response to $n_1, n_2, \ldots, n_n$. The only purpose that S serves here is to provide the mental vehicles of response to *n* and make an assignment of *m* to the object or event in question.

This is precisely the same role that humans play with many moral properties according to realist accounts. Milo makes this point quite clearly in discussing "weakly mind dependent" facts and truths, which "are constituted by states of affairs that include mental states as essential components. Thus, hedonistic

utilitarianism takes moral facts to be weakly mind dependent, since it makes feelings of pleasure or pain essential ingredients of all moral facts" (191). Weakly mind-dependent facts are distinct from stance dependent facts in that the latter consist "in the instantiation of some property that exists only if some thing or state of affairs is made the object of an intentional psychological state (a stance)…" (192). If constructivists agree that their idealized subjects are there merely to feel and respond to moral properties, rather than to constitute them in virtue of having moral mental states with particular intentional content, what appeared to be an internalist position will collapse into total externalism.[74]

4.4. Phenomenalism and Constructivism

In order to save the constructivist position, we might resuscitate a bit of Kant's work concerning freedom of the will. In the *Groundwork*, Kant argues that the possibility of freedom of the will is necessary for our conception of moral judgment: "If, therefore, freedom of the will is presupposed, morality together with its principle follow from it by mere analysis of its concept" (4:447). As Kant saw things, the purpose of moral judgment and giving a law to oneself was to free one, to a certain extent, from determinism in the phenomenal realm. The true upshot of freedom of the will and moral judgment was *spontaneity*—independence from causality in the phenomenal realm. In this Kantian sense, constructivists need a way to rescue S's judgments from the causal forces of non-moral properties on which

[74] One way of saving constructivism here would be to say that *all* properties in the world in question are stance-dependent. This thoroughgoing constructivism will need to come to terms with arguments for, say, scientific realism, and I will not adjudicate that debate here.

the moral properties supervene in our example. Rawls himself admits, "Constructivism holds that the objectivity of practical reason is independent of the causal theory of knowledge" (1993, 116). So constructivists must provide some explanation of the intentional content of moral mental states that, contra externalist accounts, does not appeal to causal forces in the world.

To do this, we can inject a bit of phenomenalism into the constructivist account. "Phenomenology" is a notoriously loose term for a family of views dating back to at least Kant and including, more recently, Brentano, Husserl, and Heidegger. As a method, it claims the same thing as internalism does for metaethics, but does so more generally for investigations of ontology, epistemology, and other areas of philosophy. Heidegger agrees with this basic characterization: "'phenomenology' signifies primarily a *methodological conception*. This expression does not characterize the w h a t of the objects of philosophical research as subject-matter, but rather the *how* of that research…the *science of phenomena*" ([1926] 1962, 50). "*Only as phenomenology*," he continues, "*is ontology possible*" (60).

Given the long and varied history of phenomenology, the term is doubtless loaded with much theoretical baggage. I will use 'phenomenalism' and 'phenomenal' to pick out views about the nature of mental content found in recent accounts of phenomenal intentionality. On these views, the phenomenal character of experience is, as Uriah Kriegel puts it, "not constitutively dependent upon anything outside the experiencing subject," "intrinsic," "inherently subjective," "the metaphysically most fundamental kind of intentionality," and "the source of all

other intentionality" (forthcoming, 7). When I am conscious of a red apple in my visual field, my mental state may include the concepts RED, FRUIT, and ROUND, as well phenomenal aspects of that experience. These aspects are often hard to capture in ordinary language, but we may approximate them here as *having a reddish sensory experience* or *having the appearance of something round*. Representationalists argue that these phenomenal aspects of mental states are constituted by the representational content of those states (Byrne 2001; Dretske 1995; Harman 1990; Rey 1998; Tye 2000). On this account, what it's like to see a red surface is nothing above and beyond representing a surface as red. Phenomenalists, by contrast, argue that this reduction of the phenomenal to the representational is incorrect and that the phenomenal aspects of experience require separate treatment (Horgan and Kriegel 2008; Kriegel 2002, 2007a, forthcoming; Loar 2003; Lycan 2008; Pautz 2008). To see a red surface is not simply to represent something as red; it's also, when one is conscious of one's sensory states, to have a certain kind of experience, the phenomenal character of which cannot be accounted for in conceptual terms alone. Phenomenal properties are intrinsic to the states themselves and always tied to the standpoint of the person having the moral experience.

Phenomenalism and moral theory may seem to make for strange bedfellows. Phenomenalists emphasize the role of subjective properties in giving an account of intentionality, experience, and mentality. Moral theory, on the other hand, concerns the actions of agents, which are publicly observable behaviors *par excellence*. A cornerstone of moral practice turns on observations and judgments regarding other's actions, which are essentially third personal. The dichotomous

nature of subjective and objective experiences and first-personal and third-personal data are well worn in the literature—rightly or wrongly—and moral *phenomenology* must surely seem an unholy marriage of the two. Simon Kirchin even thinks phenomenology is useless in settling metaethical debates. He argues that the data does not support one theory over another and notes that

> If 'phenomenological arguments for a metaethical position' translates as, 'we start a debate from our phenomenology and then seek to show that our position is correct or our opponent's position incorrect by appeal to argumentation (such as points about bootstrapping)' then I claim that it is misleading to call this a 'phenomenological argument' since phenomenology plays little or no part in the argumentation. (2003, 262)

Though Kirchin's point has merit, the possible links between phenomenology and morality have been an active area of inquiry in recent years,[75] and Steven Galt Crowell even claims that "the potential for mutual enrichment [between Kantianism and phenomenology] has as yet barely been tapped" (2002, 66).

In terms of constructivism, phenomenal properties would be part of the moral mental states of idealized subjects in considering objects and events. To play a crucial role, the phenomenal properties should be essential in specifying the stance-dependent moral properties that are logical constructs out of S's judgments. Thus, phenomenal properties can be added into the constructivist framework by substituting the clause

---

[75] For examples, see Drummond and Embree (2002), Bagnoli (2002), Horgan and Timmons (2005, 2008a, 2008b), and Kriegel (2007b). Findler even argues that "a phenomenological reading offers us a much richer understanding of Kantian ethics than other readings do" (1997, 167).

CONSTRUCTIVISM' …To bear the property *m* just is to be an object of the phenomenal property *p*, which constitutes the moral condemnation of a subject S in **W**…

for

CONSTRUCTIVISM …To bear the property *m* just is to be an object of moral condemnation by a subject S in **W**…

in our original formulation. The upshot of this revision is to make moral condemnation, approval, and the like irreducible to causal forces and to insulate them from the supervenience objection by locating them in the phenomenal properties of S's experience.

This strategy is, to my mind, deeply unattractive. Qualia, as Daniel Dennett says, "are not even 'something about which nothing can be said'; 'qualia' is a philosophers' term which fosters nothing but confusion, and refers in the end to no properties at all" (1988, 49). In a previous passage, Dennett signals (and derides) the exact use to which qualia are being put here: "Qualia seem to many people to be the last ditch defense of the inwardness and elusiveness of our minds, a bulwark against creeping mechanism. They are sure there must be *some* sound path from the homely cases to the redoubtable category of the philosophers, since otherwise their last bastion of specialness will be stormed by science" (48). Though I, following Dennett, think qualia are creatures of darkness, I am willing to set aside general concerns, entertain them for the moment, and consider only the special problems that qualia pose for the constructivist in this context—specifically concerning the constructivist's attempts to achieve objectivity and the Kantian

constructivist's penchant for moral motive. The first section will explore the general problem; the second section will explore the problem specific to Kantianism.

4.4.1. Moral Aberrants

To appreciate the difficulties introduced with phenomenalism, we can consider three cases adapted from the phenomenalist tradition in philosophy of mind. The first is *moral zombies*. If phenomenal properties cannot be reduced or are not constituted by, say, physical properties, then it's possible to imagine organisms that have exactly the same physical constitution as we do, yet lack the phenomenal properties that are associated, within us, with those physical properties.[76] This case of absent qualia extends to moral mental states as well. If moral mental states have phenomenal properties independent of the brain states of those who make those judgments, then it must be possible to imagine observers who would have all the same physical properties as we do, yet whose mental states lack the same phenomenal properties as ours. These moral zombies say things such as, "Torture is wrong," and, "Charity is good," and they respond to cases of ignited cats with horror and disgust (or so it would seem to us), but they lack the phenomenal qualities that we experience in making those statements and responding in that way.

---

[76] Cf. Kriegel's point that "the strongly internalist view which does not complement phenomenal intentionality with externalistic intentionality, need not deny that there are externalistic connections between phenomenally intentional states and items in the environment. Instead, it need only deny that these externalistic connections qualify as a form of intentionality" (forthcoming, 40)

If qualia can be absent, they can also be inverted, and this is the second case we need to consider. If qualia are truly independent of physical states, it would be possible for someone to have the same physical states as you, yet completely different qualia. Your doppelganger, having been raised in a physically identical environment as you were, has been trained by ostension to call the same things "red" and produce the same behavioral responses to, say, red roses. But what it's like for your doppelganger to see red is completely opposite to what it's like for you to see red. And similarly in the case of moral judgment. Your doppelganger is a *moral satan* who makes all the same moral judgments and (ideally speaking) performs all the same moral behaviors as you do, but loathes every minute of it. What your doppelganger calls 'right', he experiences as repugnant, vile, and shameful; what he calls 'wrong' is accompanied by the most sublime feelings of righteousness, justification, and pride.

If you find this case inconceivable, then you're on to something about the connection between qualitative experience and behavior. In many cases, difference in behavior is explained by difference in the qualitative character. The fact that something tastes extremely bitter to me explains why I avoid it. Since I take it that almost everyone dislikes this taste, I can pretty well infer that if you continue to drink strong tea, it must not taste extremely bitter to you. Thus, your behavior tells me something about what qualitative experiences you're having (within certain limits that usually attend induction and personal difference). But behavior and all its public physicality has no truck with the phenomenalist. Phenomenal properties are supposed to provide the bedrock for intentionality. Our moral satan should be

able to tell on the basis of his *direct* access to the phenomenal properties of his moral judgments that he loathes that which he judges to be right; he shouldn't need to infer anything about his qualitative states, as we do, by observing his behavior—and anyway, his behavior is the same as ours, *ex hypothesi*.[77]

Moral satans must be distinguished from a third class of aberrants: those for whom all the qualia normally associated with moral judgments are intact but who nonetheless exhibit different behaviors than we do. Call these patients *wantons*, and their existence remains possible on phenomenalism because qualia, as independent, non-physical properties of mental states, could turn out to be epiphenomenal. It could be the case that the phenomenal aspects of moral judgment play no causal role in downstream behavior. Thus, a wanton could make the same moral judgments I do and have exactly the same experience in making them but nevertheless fail to act in the same way because the usual connections between experience and behavior have been severed.

Lest you think that these cases are strawmen foisted upon the moral phenomenologist, consider a few perennial questions in the history of ethics:

(1) *Is moral motive necessary for right action?* Kant's entire discussion of moral motive seems aimed at excluding moral zombies from judgments of moral praise and adoration. It is not enough, he thinks, that one could have a desire to do the right thing or a mere inclination towards benevolence. These "springs of action" are

---

[77] As Dennett (1988) points out, there is no way, in principle, of telling whose qualia are *right* here. Dennett gives as an example a "brainstorm" machine, which allows me to see the world through your eyes, roughly speaking. If a technician were to pull the plug, rotate it 180˚, and reinsert it, we would have no way of knowing which orientation was authoritative. As he puts it, "no intersubjective comparison of qualia is possible, even with perfect technology" (50).

insufficient to garner true esteem; only the will to do what is right because it is right deserves moral praise. This supersensible will cannot be observed by anyone since it belongs to the noumenal realm to which we lack direct epistemic access.

(2) *Could wrong actions be right?* It's harder to find a true moral satan in the literature. This is usually because some link is presupposed between judging that an action is right and being motivated to perform that action and, more controversially, judging that an action is wrong and being disposed to refrain from doing it. (More on this in the next example.) To imagine a moral satan who hates the right and the good but acts on them faithfully is quite a feat indeed. More often, we encounter the *immoral* satan who loves the wrong and the bad and, in his love, acts upon them, and we are asked to consider whether his normative system may be generalized to everyone else. But we are more like the immoral satan than we realize at first glance, for the connections between his experience, say, of pleasure and his behavior is just like ours. His qualia aren't inverted; his judgments about what produce those qualia are, or more precisely, his causal situation is such that evil things bring *him* pleasure, whereas good things bring *us* pleasure.

The more interesting case—the truly inverted one—is the moral satan who hates his right actions. The most obvious place we find him in the literature is Bernard Mandeville's ([1732] 1988) infamous injunction, "private vices are public benefit." Mandeville claims that extreme rapaciousness alone drives the twin wheels of invention and commerce that lead to luxurious living. The libertine may

be vicious, yet his actions employ the servants, tailors, and tramps that fund society's nobler butchers, carpenters, and breadmakers. As Mandeville puts it,

> Virtue is made Friends with Vice, when industrious good People, who maintain their Families and bring up their Children handsomely, pay Taxes, and are several ways useful Members of the Society, get a Livelihood by something that chiefly depends on, or is very much influenc'd by the Vices of others, without being themselves guilty of, or accessary to them, any otherwise than by way of Trade, as a Druggist may be to Poisoning, or a Sword-Cutler to Blood-shed. (80)

But if Mandeville is correct that such wrongness has good effects, in what sense would it still merit being called "wrong?" The puzzling aspect of Mandeville's claim is not that wrong can cause good; it's that it would still be *wrong*! This claim *would* be plausible, however, if we were all moral satans. We would find that which produced good to be accompanied by experiences of repugnance and disapprobation. We would hate ourselves for doing what our better judgment demanded of us, and what is this but the inverted qualia case of doing the "right" thing while feeling wrong in doing it.

(3) *Is weakness of the will possible?* What is the akrasiac but a wanton whose qualia are exactly the same as ours but whose behavior has become unhinged from them. The Socratic response denies the possibility of akrasia ([380 B.C.E.] 1997, 358c–d). It treats the epiphenomenalism of moral qualia as a repugnant conclusion and maintains that anyone who has the experience of knowing virtue must *eo ipso* act on what he/she knows to be right. The Davidsonian, response, by contrast, tries

to drive a wedge between qualia and action by discussing the agent's rationality. While Davidson concedes that akrasiacs are possible, he thinks they nonetheless violate a principle of countenance, which requires them to "perform the action judged best on the basis of all available relevant reasons" (1970, 41). Effectively, this strategy interposes a step of practical reasoning between one's moral judgment and one's action and attributes cases of akrasia to failures of reasoning, not failures of efficacy of moral experience. But if the qualitative aspects of moral judgment can fail to motivate behavior, why think the qualitative aspects of rational judgment fare any better? The problem with both the Socratic and the Davidsonian responses is that once you've bought in to the notion of phenomenal properties as independent of physical properties, it's always possible that you'll be paid back in epiphenomenalism. Now it might be the case with most humans that the relevant qualia–behavior links are intact. But if we come across a true wanton, there is no denying that he/she counts as an akrasiac in his/her moral judgments.

Moral zombies, moral satans, and wantons all raise skeptical doubts about the common core of moral experience that's needed to get objectivity off the ground. The constructivist wants to claim that his/her favored individual's moral experience—his/her own, if privacy is to have its day—is proper fodder for the account. But how can phenomenalist constructivists, by their own lights, show that this experience is normal (for lack of a better term) and not one of the aberrant experiences of these derelicts? Couldn't the subject, in fact, be a moral zombie or a moral satan, devising a normative account that would seem completely spurious to true moral saints? Moreover, if anyone were to challenge the constructivist's

account of the moral point of view, how could he/she appeal to the subject's experiences when that subject could, in principle, be a moral zombie or—what's worse—a moral satan? And if the subject is a wanton, he/she will make the same judgments as we do and have the same moral experiences, yet act in a completely different way. What recourse shall we have there?

The possibility of any of these cases makes it seem as if the starting point for moral phenomenologists is purely subjective, for they have no way of justifying their judgments to others or resolving conflicts when they arise or criticizing others' behaviors for being inconsistent with those individual's own moral experience. There may, of course, be agreement in many cases, and there may be certain facts about human beings that constrain the possible correct moral mental states we can have. (I make this argument in Chapter 6.) The problem I am pointing to here is that the internalist, *qua* internalist, lacks access to those resources and is forced to proceed from some individual's standpoint. It's not clear how constructivists can account for the judgments of causally-bound creatures such as we are in a way that gets them the objectivity they wish. Diversity of experience threatens, regularities invite suspicion of supervenience and externalism, and even granting phenomenalism and qualia doesn't achieve what the constructivist wanted—zombies, satans, and wontons remain in the mix.

Failing direct appeal to phenomenal properties, the constructivist might avert to non-phenomenal aspects of moral mental states to construct an account of reference. But how would this strategy differ from the externalist one we saw in the last chapter? If the phenomenalist averts to representational content caused by

moral facts and properties, he/she may be able to explain how certain properties of the objects of S's judgment cause S to have the experiences S does. These had better not be or supervene on non-moral properties, or the phenomenalist will have taken on unwanted stance-independent baggage. But what else could they be? If causation is not going to be irregular in **W**, the non-moral properties that cause S's judgments in one case had better cause the same judgment in other cases, *ceteris peribus*. And if S's judgments supervene on these properties, so do the moral properties that are logical constructs out of those judgments. So much for stance-dependence. And as we saw in the last chapter, there will always be a simpler explanation of moral judgment available that avoids reference to these moral facts and properties. You thought that buying in to constructivism freed you from the ontological baggage of realism. But the problem of moral explanation applies to you as well on the phenomenalist strategy.

4.4.2. The Modal Objection to Moral Motive

In section 4.2, I claimed that Kantians failed to adequately defend their requirement that agents must act on an awareness of a rule in order to satisfy that rule. I now want to reinterpret that requirement in terms of phenomenalism to see whether it can be sustained. To do this, I am going to weaken the requirement just a bit. Rather than requiring that the agent have an awareness of the rule itself before his/her mind, let us allow that the subject is aware that he/she is taking up the moral point of view, but not that he/she is aware of every rule entailed by it or that awareness of such rules figures into his/her actions. The subject's moral mental

state will thus have certain phenomenal properties about it that involve this awareness, but we will make no further requirements about the nature of that awareness. The correlative requirement would be that agents must act for a respect for the moral law *in general* rather than the particular moral law that applies to their situation. If this fails, I see no hope for the stronger claim. So can the weaker requirement be sustained?

We agreed earlier that the constructivist is allowed to stipulate the idealized world in question (e.g., specify that we are looking at a possible world where the kingdom of ends is instantiated). The addition of phenomenalism here adds an interesting problem. If phenomenalism is going to remain an interesting theory, there must be worlds in which there are zombies, inverted spectra, absent qualia, and the like.[78] These possibilities are ineliminable because the phenomenalist thesis is precisely that phenomenal qualities are independent of physical properties and all of these cases are ones in which the two come apart in ways that they normally don't in the course of *our* experience. Call the possible worlds in which these cases occur "**Z**-worlds." There are also, of course, worlds in which cognizers make moral judgments, worlds where agents act morally, and even worlds where the kingdom of ends is realized. Call these latter worlds **K**-worlds. It is logically possible that there is some world **W*** in the intersection of **Z**- and **K**-worlds in which, for example, moral zombies live in the kingdom of ends. We can clearly conceive of both the kingdom of ends and of moral zombies and there is

[78] The phenomenalist can abandon these fictions, but only on pain of abandoning his theory, for if these cases are not possible, it's no longer clear that phenomenal properties just aren't garden-variety physical properties.

nothing conceptually incoherent in believing that there could be a world extensionally equivalent to Kant's kingdom of ends, but one in which no consciousness, awareness of the moral law, or intentions to follow it precede agents' behaviors.

Kant would have denied this possibility. On his view, to live in the kingdom ends would be to act out of a regard for duty, or—what amounts to the same thing—to treat everyone as an end. While it's possible to interpret this kind of action as a virtue or trait of habit, which need not be conscious, it's not clear that Kant intended *that* to be the foundation of morality as he described it in the *Groundwork*. Kant's view of virtue may just be what he thought humans, as best, were capable of satisfying. But the foundation of moral judgment is another thing altogether. As we have already seen, Kant's notion of moral motive seems to require that we *consciously* act from a regard for duty—that we act with the proper motive in mind, not merely because we, say, have an inclination toward right actions, such as benevolence. To act unconsciously from a regard for duty, then, would be impossible.

But is this true? Can't duty just become a habit for me? Perhaps it's true that I need to consciously reflect on my motives in order to cultivate the proper habits. But that isn't a necessary truth about virtue; it's a contingent truth about the way humans are. In other **K**-worlds, virtue could come about by magic or irregular causality or some other mechanism. There is no reason to think that virtue is predicated on some kind of moral consciousness in all possible worlds in which it exists. So not only is it logically possible that there are worlds that are both **K**-

worlds and **Z**-worlds; it is *physically* possible. In those worlds, cognizers make moral judgments without aid of the phenomenal aspects of their experience, which they lack, *ex hypothesi*. If phenomenalism is true, there must be **Z**-worlds. If constructivism and phenomenalism are consistent, there must be worlds belonging to **K** ∪ **Z**, and **W*** is among them.

The point of demonstrating the possibility of **W*** is that the constructivist has no way of showing that the ideal world he/she wishes to stipulate is not **W***. Because the evidence of the idealized subject's moral mental states is purely behavioral or extensional, the constructivist's observer has no way of knowing that S is not in **W*** as opposed to **W**. The real problem with *Kantian* constructivism is not that S is part of some causal order and that S's moral mental states can be rescued by separating them from that order. S's judgments can have as many phenomenal properties as you like. But observers will never be able to see them, and behavioral evidence will not serve to distinguish **Z**-worlds from non-**Z**-worlds because, *ex hypothesi*, agents behave in the same way in these two sets of worlds even though their phenomenal properties differ. The real problem is that internalist methodology commits Kantian constructivists to using S's moral mental states to determine the entailments of the moral point of view. But moral mental states proceeding from an awareness of the moral point of view and moral mental states that conform to the moral law (but without awareness) are extensionally equivalent, and the constructivist framework simply *cannot* establish that an idealized subject *must* form moral mental states with an awareness of (and act in accordance with) the moral law.

Without phenomenalism, it's not clear how Kantian constructivists are going to save their account of moral motive. They can't appeal to the causal efficacy of moral facts and properties in this world (if there were any), and one doubts that moral facts and properties in *other*, *non-existent* worlds are any more efficacious. Absent the representational or the phenomenal account of intentional content, it's up to Kantians to present evidence for the fact that **W** is not **W*** and that idealized subjects necessarily have an awareness of the moral point of view (or moral law) in having moral mental states.

4.5. Conclusion

In keeping with their internalist methodology, constructivists attempt to develop an objective framework of stance-dependent moral facts and properties for assessing normative claims. As I have argued, there is reason to doubt the commonality of moral experience sufficient for achieving this objectivity. Moreover, internalism provides impoverished explanatory resources for establishing this objectivity or the Kantian requirement of acting from awareness of moral law. When we examined claims about the judgments of idealized subjects, we saw that regularities in those judgments suggested a theoretical collapse into externalism, and the addition of phenomenalism failed to help save the constructivist account.

Chapter 5

Metaethical Performativism, or Morals by Utterance

> The philosophers have only interpreted the world, in various ways; the point is to change it.
>
> —Karl Marx, "Theses on Feuerbach"

In the two previous chapters, we examined metaethical theories on which moral language is understood as asserting views about the nature of moral reality. Externalists attempt to evaluate these assertions by observing reality and seeing whether there are any moral facts and properties that would make those assertions true. Moral realists believe there are such facts and properties and purport to observe them directly in the actual world and indirectly in other possible worlds. Internalists attempt to reveal the truth of moral statements by examining their own experience from the moral point of view and, in the case of constructivism, considering what that experience shows about the moral thoughts, judgments, and attitudes of ideal moral subjects. Both of these approaches endorse descriptivism about moral language and, correspondingly, cognitivity about moral mental states.

In contrast to externalism and internalism, a family of views that I shall call "performativism" attempts a different route. Where others treat moral language as descriptive, performativists treat it as primarily non-descriptive. Simon Blackburn,

for example, says that his account attempts "to place ethics on the directive side rather than the representational side" of things (2006, 161). Decades earlier, C. L. Stevenson alleged that traditional theories were wrong in their analysis of ethical judgments: "Their major use is not to indicate facts, but to *create an influence*. Instead of merely describing people's interests, they *change* or *intensify* them. They *recommend* an interest in an object, rather than state that the interest already exists" (1937, 18–19). On both of these views, there is something extremely wrong about treating moral discourse as passively stating what is the case. Moral discourse *is* about the world—the real world we interact with and the one we share with other agents. But it's not about the world in the same way that factualist forms of discourse are. Moral language does not function to describe the world or to give speakers and hearers a way to represent that world. It is, at its most fundamental level, a way of *changing* the world, including other people's attitudes toward it. Accordingly, performativists agree that in making a moral remark, I am performing a certain kind of speech act that both expresses my own attitude and also *does* something to change the world, and not just in factive ways with respect to other peoples' beliefs.

Part of performativists' views—and they have been varied—has always been a reaction against their opponents. If anything is common to performativists' approach aside from their view of speech acts I just described, it is an enduring skepticism of the metaphysical posits of realism and the idealized subjects often thought to garner constructivism its objectivity. When G. E. Moore ([1903] 1993) posited non-natural ethical properties, A. J. Ayer ([1946] 1952) retorted with

emotivism. When critics derided the apparent subjectivism of emotivism, R. M. Hare (1949, 1952, 1963) countered with universalized prescriptions. When the naturalistic forms of realism reached their peak in the 1980s, Simon Blackburn (1984, 1988, 1993, 1998) and Allan Gibbard (1990) offset these gains with powerful expressivist accounts. My discussion here will treat these authors and theories as engaged in the same broad approach to moral language and as addressing, in general, the question, How do we make sense of moral language within a naturalistic view of the world?

It is extremely hard to give any programmatic criticisms of performativism. As we will see, contemporary forms of performativism, such as quasi-realism, have become so complex that they would likely offend emotivists' more pristine sensibilities about moral discourse. Still, I think there is a common agenda that runs throughout the performativists' views and a set of questions they must address regarding moral mental states. I begin in §5.1 by giving an overview of the performativist position and describing their common views of moral utterances, moral attitudes, and the interplay of facts and values. This overview will encompass both emotivism and more recent forms of performativism, such as expressivism. In §5.1.3, I develop a problem involving mind-to-world satisfaction conditions for moral mental states according to performativism. Following that, §§5.2, 5.3, and 5.4 explore possible responses. Notably, §5.4 examines quasi-realism and its attempt to incorporate these conditions. I argue that that quasi-realism succeeds in making sense of mind-to-world satisfaction conditions only at the cost of becoming an

externalist account and abandoning performativism as a substantive metaethical approach.

5.1. Varieties of Performativism

Performativists view moral language as fundamentally *active*—that is, in the process of making a moral expression, one is also *doing* something. Strictly speaking, all speech acts *do* something; speech acts are, after all, a species of behavior. Even statements of fact such as, "It's raining," or "This chocolate tastes sweet," make assertions about states of affairs in the world, and in making such assertions, one gives, as Austin put it, a sort of guarantee "that we are sure of it or know it…" ([1961] 1979, 77). In a loose sense of 'active', moral realists and constructivists can agree with performativists that moral language is active, for they see moral language as making assertions about moral facts and properties. By saying, "Murder is wrong," a speaker is, on a realist reading, giving a guarantee that murder has the property of wrongness in the actual world or ideal world(s). On a constructivist reading, that speaker is giving a guarantee that ideal subjects would morally disapprove of murder. These two theories differ, of course, in what kinds of moral facts and properties they countenance, but the basic view of language as descriptive and of moral expressions as assertions remains the same.

The difference between these theories and varieties of performativism is that performativists think moral expressions bring into existence a certain force in the world—a force based in persuasion or perhaps recommendation of a plan of action. In this sense, moral language is fundamentally similar to the language of

promises, vows, commands, indictments, and so forth. In the very course of making such speech acts, speakers create an influence on others—and not merely concerning their beliefs about how the world actually is (as happens in the case of assertions). This influence, if it is to be moral, must be more than just an ordinary recommendation of something. Saying, "Torture is wrong," does something more (or at least different) to recommend against torture than saying, "You should take the five o'clock train," does to recommend a certain mode of transportation. If I don't take the five o'clock train, you may call me foolhardy, inefficient, or unwise. If I do so in the moral case, you may do that as well, but you may also do much more: you may call me immoral, may attempt to get others to shame me for my actions, may bring me up on charges of human rights violations and so on—all with some degree of reasonableness, warrant, or justification. (You might do the same in the non-moral case, but your efforts would not appear similarly reasonable, warranted, or justified.) Thus, performativists need to tell us more about distinctively *moral* attitudes, a subject we will consider in greater detail in §5.1.2. For the present, we can formulate a working account of performativism as follows:

PERFORMATIVISM When a speaker S makes a moral utterance U about an object or event, U expresses S's moral attitude toward that object or event, creates a moral influence on others' attitudes and their actions, and (optionally) expresses S's other, non-moral attitude(s) toward that object or event.

On this account, moral sentences are non-descriptive in respect of their moral content, for that content does not assert anything, morally speaking, about the objects and events in question. Instead, moral language is *sui generis* in function. Ayer, for example, argues that moral discourse is radically different from non-moral discourse and that attempts by subjectivists and utilitarians to reduce that discourse to psychological states are completely misguided:

> we are not, of course, denying that it is possible to invent a language in which all ethical symbols are definable in non-ethical terms, or even that it is desirable to invent such a language and adopt it in place of our own; what we are denying is that the suggested reduction of ethical to non-ethical statements is consistent with the conventions of our actual language. (105)

Moral language does not play the same role that any other kind of language plays; it expresses a distinctive mental attitude and imposes a special moral force on listeners and their subsequent actions. Many critics have misinterpreted performativists in taking them to say that speakers are talking *about* their states of mind, rather than acting, verbally, to create this special influence. Before considering this problem in further detail, it will help to attend to two features of the formulation above to appreciate the full nature of the performativist's claims.

First, performativists analyze moral language at the level of utterances, not sentences or even thoughts. As Stevenson points out, "It frequently happens that the same sentence may have a dynamic use on one occasion and not another, and that it may have different dynamic uses in different occasions" (1937, 21). Stevenson

gives as an example a man who says, "I am loaded down with work." In some situations, he may be simply be conveying how his life is going. This instance would be a descriptive or assertoric speech act. On other occasions, however, he may be dropping a hint, such as asking for help, or even trying to elicit sympathy from his neighbor. These are cases of different varieties of dynamic use. Notice also that if the man says, "Ugh," we can pretty well take him to be saying the same thing even though he uses different words. It may be less clear *exactly* what he means, but it may be clear from the context that he is generally placing a certain demand on us and talking *about*, say, his present situation. And if we imagine him saying, "I am loaded down with work and it's wrong for you to ask me to help you," we imagine yet another kind of dynamic use that has a distinctively moral force that the other interpretations lack. Utterances are the proper level of analysis for performativists because speech acts are always tied to specific to contexts and trade on general social conventions about what one is doing (or trying to do) in making a certain kind of remark. (I discuss these points further in §5.1.3.)

Second, performativists do not deny that moral utterances often have descriptive components to them. In fact, Ayer and Stevenson seem to regard these elements as crucial for resolving some forms of moral disagreement. "[W]e find," Ayer says, "if we consider the matter closely, that the dispute is not really about a question of value, but about a question of fact" (110). Stevenson is a bit more precise here, noting that moral disagreements may arise from factual agreements, but that the latter do not completely determine the former: "people who disagree in interest would often cease to do so if they knew the precise nature and

consequences of the object of their interest. To this extent disagreement in interest may be resolved by securing agreement in belief, which in turn may be secured empirically….But note that empirical facts are not inductive grounds from which the ethical judgment problematically follows" (1937, 28). The point is that the *moral* component of an utterance creates a distinctive kind of influence that cannot be reduced to or understood merely in terms of the descriptive portion of that utterance. For this reason, performativists say that the propositional content of moral statements and their moral force are unrelated, semantically speaking. That is, the meaning or significance of the moral component and the truth of the non-descriptive component have no logical relation to each other. One is aimed at creating an influence; the other is often aimed at reflecting the world, as it were.

Let us grant the performativist that the moral portion of any utterance is not truth-conditional, but the non-moral portion may be ('may' because the utterance could express a command or other non-veridical attitude just as well as it could express a belief). Presumably, whether a certain piece of verbal behavior is truth-conditional or not depends on its intentional content and the mental attitude of the mental state it expresses. Thus, assertions are truth-conditional because they express beliefs with certain intentional contents. Similarly, in saying that the moral component of any utterance is not truth-conditional, the performativist must be committed to the view that it expresses a mental state, the

attitude of which does not have a mind-to-world direction of fit.[79] For this reason, it is absolutely critical for performativists that moral utterances *do* create a special kind of influence, otherwise they will be analyzable in terms of some other class of speech act, such as commands, wishes, or even assertions. For example, consider the difference between someone's saying, "Nuclear disarmament would be nice," and "We have an obligation to eliminate nuclear weapons." The former seems to express a mere wish—I could take it or leave it, as it were. The latter, however, creates a certain demand on me (and everyone else) that I cannot simply ignore. Stevenson describes this as *magnetism*: any analysis of moral terms must account for the fact that "A person who recognizes X to be 'good' must ipso facto acquire a stronger tendency to act in its favor than he otherwise would have had" (1937, 16). While Stevenson's description may be incomplete in fully describing this attitude, it conveys the sense in which performativists see moral attitudes and their *expression* as unique in nature and not in the business of veridicality.

Before addressing these larger issues of mind-to-world satisfaction conditions, I want to discuss three preliminary issues involved with performativist accounts. One is the historical error critics have made in treating performativists as collapsing the reporting–expressing distinction. This discussion in §5.1.1 will give way to another consideration: what exactly is the moral attitude performativists have in mind on their accounts? I take up this consideration in §5.1.2, followed by a

---

[79] Rather curiously, the early performativists are silent on the intentional content of moral mental states. Rather, they see moral attitudes as not being truth-conditional and their corresponding expressions as creating a certain force, but not having any truth conditions.

discussion in §5.1.3 of performance conditions for moral speech acts and their bearing on the satisfaction conditions of moral mental states.

5.1.1. Reporting and Expressing Mental States

Many critics take performativism, especially emotivism, to be the view that moral expressions say what a speaker likes (or doesn't like).[80] Accordingly,

U1 "Murder is wrong"

means something like

U2 "Murder happens and I don't like it."

This misinterpretation may be an accident of the stress that performativists place on the independence of the moral and non-moral components of a given utterance. Consider this famous passage by Ayer:

> [I]f I say to someone, 'You acted wrongly in stealing that money,' I am not stating anything more than if I had simply said, 'You stole the money.' In adding that this action is wrong I am not making any further statement about it. I am simply evincing my moral disapproval of it….If now I generalize my previous statement and say, 'Stealing money is wrong,' I produce a sentence which has no factual meaning—that is, expresses no proposition which can either be truth or false. It is as if I had written 'Stealing money!!'—where the shape and thickness of the exclamation marks show, by a suitable convention, that a special sort of moral disapproval is the feeling which is being expressed." (107)

---

[80] For examples of this mistake, see Aiken (1952), Glassen (1959, 1963), and Alston (1968).

Ayer seems to regard moral terms as adding nothing to the content of the uttered sentence itself. Thus, he might be read as saying that in making the statement,

U3 "You acted wrongly in stealing that money,"

I might as well have said two things in sequence:

U4 "You stole the money"

and

U5 "I have a feeling of moral disapproval about it."

The problem with this formulation is that it treats all moral utterances as non-moral statements of fact about an object or event, combined with reports of the speakers' mental state. A different way of saying the same thing as (U4) and (U5) together would be

U6 "I disapprove of your stealing the money."

This interpretation makes it clear why critics saw emotivists as putting forward a simple form of subjectivism. On this analysis, if I say something is wrong, I am merely saying that I don't like it. If you say,

U7 "I think it's okay you stole the wallet,"

you are saying that it's not the case that you don't like it (i.e., that you think stealing is permissible). There is no disagreement between us because we are merely saying things about our own likes and dislikes. To put this a bit differently, the truth conditions of (U5) and (U6) are determined by the mental states of those who utter them—namely, they are true just in case the speaker actually has a feeling of moral disapproval about the theft. Both (U6) and (U7) can be true of

different speakers (or the same speaker at different times). After all, there is nothing wrong with us having different preferences (or changing our preferences). In such a case, there is no disagreement because we are simply talking about different things (i.e., our own states of mind).

This is not the way to be a performativist. Ayer himself is clear that, in making a moral utterance, "I am not making any factual statement, not even a statement about my own state of mind" (107). Stevenson, too, is eager to reject "interest theories" that treat moral language as statements about what people actually desire (1937, 15). On his view, Hobbes and Hume were both wrong for claiming that "'good' means *desired by me* (Hobbes); and 'good' means *approved by most people* (Hume, in effect)" (1937, 15). If such analyses were correct, they would turn moral expressions into assertions or descriptions of a certain state of affairs, namely the mental state of the speaker or community. However, moral utterances are not, for performativists, fundamentally in the business of describing anything but rather function to create a certain influence.

Accordingly, we must interpret the performativist as saying that moral utterances express, not report, the moral mental states of their speakers.[81] Reporting and expressing differ in respect of their satisfaction conditions. Reports are assertions *about* one's own mental states and are therefore true or false depending on whether one is right about being in a certain state. Expressions serve to communicate what one has a mental state *about* and could be fulfilled or

---

[81] For more on the reporting–expressing distinction, see Rosenthal (2005, 49–55).

unfulfilled in a variety of ways. (As we have seen, they may be true/false, fulfilled/unfulfilled, obeyed/disobeyed, faithful/unfaithful, etc.). Having made it clear that performativists view moral utterances as *expressing* moral attitudes to create an influence, we can now consider both the nature of that attitude and the mechanism by which such influence is created.

5.1.2. The Moral Attitude

On the account of performativism I have been developing, moral attitudes play a very important role. Their expression constitutes a fundamentally different type of speech act than the expression of any other mental state, such as beliefs or assertions. It distinguishes moral utterances from descriptive ones and creates a distinctively *moral* influence on listeners. What, then, is this primitive attitude that does so much work for the performativist?

Ayer is largely silent on this moral attitude, characterizing it only in terms of "feeling" and, in the specific examples he gives, "disagreement" (107–12). Stevenson describes it as having *magnetism*, as if it were a fundamental force in the world. This metaphorical vagary is no less common among other authors, including Philippa Foot, who have cast around for the "moral must"—that special kind of force that morally binds us and, through our expressions, others to certain courses of actions.[82] Simon Blackburn, who offers perhaps the most advanced views of performativism, rejects the very need for any account of this moral attitude: "Analytical philosophers demand definitions, but I do not think it is profitable to

---

[82] As I discussed in §2.1.1 and Note 21, Foot (1972) is not enthralled with this concept and thinks a weaker attitude might be sufficient.

seek a strict 'definition' of the moral attitude here. Practical life comes in many flavours, and there is no one place on the staircase that identifies a precise point, before which we are not in the sphere of the ethical, and after which we are" (1998, 13–14). While Blackburn's position may ultimately be correct, it is worth considering what the performativist *could* give as an account of moral attitude that would make clear how he/she wishes to account for the world-to-mind *and* mind-to-world satisfaction conditions of moral mental states.

As we saw in Chapter 2, different satisfaction conditions attend to different kinds of mental attitudes. Beliefs have mind-to-world satisfaction conditions, for they are concerned with veridical representations of the world. Intentions, hopes, and wishes all have world-to-mind satisfaction conditions, for they attempt to change the world in ways that bring it closer to a person's vision for it. We agreed there that moral mental states must have *both* kinds of satisfaction conditions. They are responsible for being better or worse visions of how the world should be changed. As a historical point, the early emotivists came under intense criticism in the post-war years, not so much because their views of moral language were flawed, but rather because they were branded as moral monsters who refused to offer *any* advice on how to live one's life. As Mary Warnock then put the point,

> moral philosophers in England…have for the most part fallen in happily with the positivist distinction between moral philosophers, who analyse the logic of moral discourse, and moralists, who practise it. It follows that they are inclined to believe, in theory at least, absolutely anything could count as a moral opinion, or a moral principle, provided it was framed in

> the way laid down for such principles, and used, as they are used, to guide conduct. It would be generally agreed that some opinions might be outrageous, and some principles harmful, but where we get our principles and opinions from, how we should decide between them, and what would be an example of a good one—these things they will not tell us, for to do so would be actually to express a moral opinion. (1960, 144–45)

Though contemporary performativists, including Blackburn, still resist this demand, others have tried to meet it by giving a clear account of the moral attitude and the judgments in which it issues.

Allan Gibbard gives what is perhaps the most advanced performativist account in discussing the moral attitude in connection with rationality. According to his account, to have a moral attitude about something—or, more generally, to have a normative attitude—is just to accept a rule or imperative as part of one's own process of practical reasoning: "There are…special *concepts* that figure in planning and decision—such as the concept of *being the thing to do*" (1990, 7). Thus, moral attitudes reflect the end of practical reasoning: one's conclusion about how to act. Gibbard also claims that acceptance or endorsement of a norm prohibiting a certain action *requires* that one would feel guilty upon performing that action or angry at someone else for performing it (1990, 297–300). For reasons I discussed in Chapter 2, this account is unsatisfying. Guilt and anger often accompany non-moral judgments, so their presence does not determine that a particular state is a moral mental state. Moreover, although Gibbard tries hard to show that guilt is a

cultural universal, there is certainly no necessary connection between specific culturally conditioned emotions and the nature of moral judgment in general. Guilt is expressed in various ways in different cultures, and there is no reason to think that an observer is talking about the same phenomena of moral mental states across various purported cases of guilt.

In more recent work, Gibbard (2003) amends this account by saying that we plan for different hypothetical considerations; we form what he calls "fact–plans"—given certain facts, planners prepare themselves to act in certain ways. On this account, correct moral judgments are whatever hyperplanners, who plan for all (relevant) contingencies, would decide. While this formulation avoids problems associated with specific emotions, it raises new questions about how hyperplanners go about their planning, what moral attitudes are like for them, and why different hyperplanners are bound to reach the same conclusions about what is the thing to do. We can set aside these questions, however, for this account fails to answer the basic question we asked of the performativist: What is the moral attitude? *We* are not hyperplanners, nor do we possess the cognitive abilities they allegedly exhibit. Thus, hyperplanners do not describe the moral attitude as it actually occurs in human beings. The performativist owes us an account of *our* moral mental states, not idealized ones, and the same worries I raised about idealized subjects in §4.3.2 apply to the Gibbard's hyperplanners.

One reason that performativists may be shy about defining the moral attitude and its satisfaction conditions is that they have, historically, taken issue with realists' accounts of what it means to be correct in holding a certain attitude.

To put this simply, assume that in holding a moral attitude, we feel that we are bound to do something or other. What can we do to query whether the bindingness should be there or not? Performativists seem unsure that we can ever escape this normative circle and ask, from an external standpoint, whether such attitudes are *right*. If we ask, But should this plan really be binding? we are thrown back onto the question of the bindingness of bindingness, and so on. This variant of Moore's open question argument should come as no surprise. One of the starting points of performativism was that moral language is *sui generis* and irreducible to other kinds of discourse. To suggest that we can step outside of that framework and ask about its appropriateness runs counter for this very suggestion. That is why performativists will even treat metaethical realism as a species of normative question: Should I plan to do something, come what may, or not? Gibbard puts this point as follows:

> "Normative facts are out there, subsiding independently of us" might just be a fancy way of putting an aspect of a plan for living. Return to a specific instance of the claim, "It's a normative fact, out there independent of us, that one ought not to kick dogs for fun." Accepting this might amount to planning to avoid kicking dogs for fun, planning this even for a contingency of being someone who approves of such fun, and who is surrounded by people who approve. The claim of independence, then, turns out to be internal to normative thinking—though arrayed in sumptuous rhetoric. (2003, 186)

What appeared to be a substantive metaethical claim about the existence of certain stance-independent facts turns out to be a normative position after all—"normative" in the sense that stance-independent facts turn out to be directly action guiding and that, against background assumptions of rationality, ignoring them will turn out to be wrong, misguided, careless, etc. There is, in short, no "Archimedean point" (254) from which to survey normative questions—we must do it from within the very standpoint of moral discourse itself.[83]

No performativist wants to concede that all plans are equally good or that all attitudes are equally appropriate. The struggle for them is how to assess those attitudes from within the standpoint of the very same discourse one wants to critique. Blackburn seems to think it is sufficient that he can show that we have a way of reflecting on the moral views we hold: "a theory based upon attitudes can encompass those elements in moral thought that make us reflect on how difficult it is to know what is right and wrong, good or bad" (1993, 128). What is wanted, intuitively, is not just the *possibility* of criticism; it is a *standard* for it. But this is precisely what the performativist abandons in holding his/her view of moral discourse. As we will see in §5.4, quasi-realists have attempted to give these standards, but only at the cost of abandoning the core of performativism.

5.1.3. Performance Conditions and Satisfaction Conditions

There is, undoubtedly, something distinctively *personal* about having a moral attitude—after all, the attitude is mine, not yours. Gibbard (1990) sometimes

---

[83] For other examples of this position, see Dworkin (1996) and Blackburn (1996).

describes this as being "in the grip" of a norm (though he says it's enough if one merely accepts norms rather than, say, *feels* that they are binding). Stevenson, too, recognizes this personal nature of moral attitudes in describing magnetism: "this requirement excludes any attempt to define 'good' in terms of the interest of people *other* than the speaker" (1937, 16). Given this emphasis on one's own feelings, interests, and so on, one might wonder how these attitudes transform in the course expression to create an influence on others. To paraphrase Mackie, there is something very queer about having a moral attitude and expressing it, and it lies in the symmetry of the felt bindingness of the norm. When I have a moral attitude, I feel that the plan that it recommends is binding upon me: If I think the norm is universal, it is still something that I feel binds to *me*. If I think the norm is something that applies to you, then, in the course of having the moral attitude, I feel that *if* I were in your position, *I* would feel such and such a plan to be binding. If performativists are committed to the notion that moral attitudes *feel* binding, that bindingness is only felt by specific agents for as long as those agent are in those states of mind.

But notice what happens when those agents express their moral attitudes. Suddenly, the bindingness shifts from the speaker who has the moral attitude to his/her listeners; I bring it about that *you* feel bound. This shift lies at the heart of performativist accounts of the moral speech acts, though it is not completely unique to moral language. Korsgaard gives a nice example of the normativity of everyday language in *The Sources of Normativity*:

> If I call out your name, I make you stop in your tracks. (If you love me, I make you come running.) Now you cannot proceed just as you did before. Oh, you can proceed, all right, but not just as you did before. For now if you walk on, you will be ignoring me and slighting me. It will probably be difficult for you, and you will have to muster a certain active resistance, a sense of rebellion. But why should you have to rebel against me? It is because I am a law to you. By calling out your name, I have obligated you. I have given you a reason to stop. (2003b, 140)

Korsgaard's example is inextricably social, but so are verbal behaviors such as commanding, instructing, requesting, promising, and perhaps even asserting.[84] The difference between these other performances and moral language seems to lie in the nature and degree of the felt bindingness. Expressing a moral attitude and performing a moral speech act does something very special: it shifts the attitude *itself* on to the listener, *ceterus paribus*. No other mental attitude or corresponding expression that I can think of does this (to the same degree[85]):

---

[84] Following Kant, Korsgaard rejects this stress on sociality and explains normative force as arising only when a free will "choose[s] a maxim it can regard as a law" Korsgaard (2003b, 98). But we might regard this, too, as "social" in the same way we discussed it above. As long as there is a lawgiver and a citizen, so to speak, and the former is capable of imposing oughts upon the latter (and punishing and rewarding the latter in light of the oughts' fulfillment), it doesn't matter whether the lawgiver and the citizen are individuals, groups, or even the same person. Two *roles* must be played, and 'social' merely captures this requirement.

[85] Austin and Grice thought that *all* speech acts involve *some* kind of shift onto members of the audience. However, the difference between moral expressions and other speech acts seems to be that both the mental attitude and content expressed by the speech act are imposed upon listeners, at least temporarily.

- Beliefs don't, since I can express the most implausible beliefs and you can go on just fine without believing or even *wondering* about those beliefs yourself.
- Intentions, vows, and promises don't, either, for they depend on the consent of both parties in order to create some kind of binding force.
- Even my wishes, hopes, and pleas don't, for you can just ignore them, especially if there is no relationship between us. Even if you do agree with or embrace them yourself, they are still *my* wishes, not yours. You simply hope to see my hopes fulfilled.
- Commands, even when they are obeyed, fail to impart the *same* attitude on listeners. Speakers want it to be the case that listeners do such and such, while obedient listeners just want to bring it about that such and such.

In most of these cases, there is an asymmetry in attitude between speakers and listeners. The only exceptions are intentions, vows, and promises, but those require that, antecedently, the speaker and the hearer embrace the same set of attitudes.

The moral attitude seems unique because, once expressed, it imparts a symmetrical attitude on its listeners. Even the craziest person of the lowest rank who learns to use the word 'right' has learned how, by convention, to impose his/her attitudes on others, at least temporarily. This is why I think performativists put so much emphasis on the *act* of performance itself. In using a moral term, I *intend* (explicitly or implicitly) to bring it about that anther person feels bound, morally speaking. The very utterance of the moral term transfers my attitude

toward someone else. If you tell me, "What you are doing is wrong," I may rage against the feeling you create in me or I may decide to reject it, but the whole point of moral language is that I cannot help but *feel* it—and in just the same way as you do. That is the very point of being able to use moral terms at all. If my listeners don't feel this special force, there is something wrong with *them*, no less than if someone utters "red" and a listener thinks of blue. It is built in to the *use* of moral terms that they create this effect; otherwise, we have no moral language at all.

One way of stating the criticism made in the last section—that performativists have ignored questions of satisfaction conditions—is by saying that they give no criteria for the *correct* use of moral terms. To return to the example I just gave, if you and I are both observing a blue object and I say, "Red," it is clear that I have misused the term. This is especially clear if I use the term 'blue' to describe red objects—then you can simply say I don't know the difference between 'red' and 'blue'. What corresponding story can the performativist give about the use of moral terms?

Stevenson is quite explicit that the *meaning* of moral terms *just is* their use: "The emotive meaning of a word is a tendency of a word, arising through the history of its usage, to produce (result from) *affective* responses in people" (1937, 23). But use in this context is making a performance of a certain kind; no further criteria seem to apply. So far as I can tell, the only performance conditions that performativists acknowledge is that the speaker actually have the moral attitude that he/she expresses (a sincerity condition) and, perhaps, an intention to bring

about a change in others' attitudes (an intention that is, essentially, an awareness of the special force of moral terms).[86]

One wonders what it would take to get things wrong on this account—being mistaken about one's own attitude? But, as we have seen, performativists are not claiming that we *report* our mental states, merely that we express them. If I merely have a wish that something be the case, but I use moral language to describe it, it is not clear that I have done anything "wrong" according to performativism. I have not expressed a moral attitude, but if performativists are right about the role of moral terms, I have still brought about, socially, a situation in which you *prima facie* feel challenged to do something. So there is a substantive concern that *there are no* performance conditions for using terms such as 'good', 'right', and so on. Using them correctly is just using them at all!

This is troubling indeed. What seems crucial is that speech acts have performance conditions that involve a speaker's having a certain mental state and that those corresponding mental states have some kind of satisfaction conditions that, in turn, determine the semantic possibilities for those speech acts. Accordingly, assertions express beliefs, which have mind-to-world satisfaction conditions, and assertions may be true or false. Commands express desires or intentions, which have world-to-mind satisfaction conditions, and commands may be fulfilled or unfulfilled. And so on for other speech acts and their corresponding mental states,

---

[86] Friends of Grice (1989) will take kindly to this point. Language is a rule-governed activity, and to say that I must have intentions to perform such and such a speech act is just to say that I must be aware of the rules for that speech act. For reasons discussed in the last chapter (see Note 71), I see no reason why this cannot become a habit and occur without any awareness of the rules in mind.

and the satisfaction conditions of those states and semantic assessment. Performativists, in saying that moral expressions cannot be true or false, seem to deny that moral attitudes have any mind-to-world satisfaction conditions.

This alone might not be a *reductio* of performativism—after all, we could be deeply mistaken about the nature of moral discourse.[87] However, there is also a sense in which, according to performativists, the world-to-mind satisfaction conditions of moral mental states are fulfilled once the states are expressed because the expression, *ex hypothesi*, creates a binding force on listeners that serves to bring about the content of the attitude. To put this more directly, suppose three conditions hold: (a) in having a certain moral mental state that *p*, I am responsible for bringing it about that *p*, (b) by expressing that state, I, *ex hypothesi*, bring into existence a force that instructs others to achieve *p*, and (c) *p* is actually achieved. Assuming (a)–(c), it seems that my mere *expression* has achieved the satisfaction conditions of the state it expresses. If the only performance conditions for moral terms is that one uses them, then the failure of this force lies in listeners who don't understand moral terms in the first place, not in speakers for not bringing it about that *p*.

Now, it may turn out that the world-to-mind satisfaction conditions for moral mental states are not *completely* satisfied by satisfying the performance of a certain speech act. It might matter not just that, say, certain listeners *feel* bound, but

---

[87] Notice that this thesis is a bit different and more thoroughgoing than error theory. Error theorists believe that moral mental states and their corresponding expressions *do* have truth-conditions. In fact, their account depends on it, for they then go on to say that nothing exists that could possibly make them true, so they must all be false. This non-cognitivist position says that moral mental states and their corresponding expressions aren't even in the business of truth in the first place.

that they, in fact, *do* what the speaker's moral mental state envisages them doing or that, for some reason, the speaker himself/herself is responsible for bringing about the state of affairs. However, the mere fact that a speech *act*, in part, fulfills the satisfaction conditions of moral mental states may be responsible for critics' sense that performativism is mere subjectivism or, to put the point differently, that moral attitudes are self-fulfilling.

These reflections suggest that performativists have serious work to do in terms of the satisfaction conditions of moral mental states—and not just in terms of mind-to-world conditions, either. In the next three sections, I examine different strategies by which performativists have tried to account for the mind-to-world satisfaction conditions of moral mental states. One way, which I examine in §5.2, is to claim that *two* states are actually expressed in most moral utterances: a belief and a moral attitude. The former, having a mind-to-world direction of fit will have truth conditions in the standard sense. The latter, having a world-to-mind direction of fit, will not. This strategy accommodates the insights of Chapter 2 by splitting every moral mental state in two. Another strategy, which I consider in §5.3, tries to account for those same insights about moral mental states by arguing that moral mental states are indeed truth-conditional. As I will argue, truth is not where the action is, and this strategy ultimately fails because it does not say anything about what really matters: the mind-to-world satisfaction conditions for moral mental states. Historically, this position has also given way to quasi-realist accounts, which do genuinely address this question, but only at the cost of abandoning the central commitments of performativism. I consider quasi-realism in §5.4.

5.2. The Two-State Solution

The first attempt we shall examine to accommodate mind-to-world satisfaction conditions for moral mental states does so by splitting every moral mental state into two separate states, one with mind-to-world satisfaction conditions and another with world-to-mind satisfaction conditions. It takes as its basis the sense in which all performativists agree that mind-to-world satisfaction conditions are important with respect to moral attitudes: World-to-mind states succeed or fail in virtue of how well they bring about certain states of affairs in the world. If these states of affairs already exist, then one will automatically fail in meeting the satisfaction conditions of world-to-mind intentional states that aim to bring them about. This isn't a necessary condition for someone actually *having* some world-to-mind intentional state—I can very well intend to create the world. Instead, the state of affairs' not existing is a necessary condition for one's fulfillment of the world-to-mind mental state—if the world already exists, even *God* can't succeed in intending to bring it about (for the first time)! One may intend to create the world and indeed it may turn out that the world exists, but there is still a problem in the fact that the world's existence did not come about in the proper way, namely the way specified by the content of the intention. So there is a very minimal sense in which success in achieving world-to-mind intentional states presupposes success in mind-to-world intentional states: my intention to do something will necessarily fail unless I (correctly) believe that it is not already done.

Consider a case in which I intend to buy groceries, but, as it turns out, I have already purchased them. It will be impossible for me to fulfill the world-to-mind

satisfaction conditions for my intention because it is impossible for me to *now* buy those very same groceries (i.e., the very ones I need). I can, of course, buy *other* groceries (i.e., duplicates or additional ones), but not the ones I intend to buy under the description we have in mind. To put this a bit more formally, let

¬ßp stand for ⌜I don't believe that p⌝,

ß¬p stand for ⌜I believe that ¬p⌝, and

ιp stand for ⌜I intend that p⌝.

The counterfactual claim ¬(¬ßp) ⊃ ¬ιp holds, but ¬(ß¬p) ⊃ ¬ιp does not. I might believe that I don't have groceries yet still not intend to buy them, but I won't intend to buy groceries if I believe I have them (i.e., if it is not the case that I don't believe that I have them). If I do, something has gone wrong, not in the sense of logical falsehood, but in a more practical sense of planning and rationality.

All of this may be a more sophisticated way of demonstrating what Stevenson told us: factual beliefs shed a lot of light on non-factual moral attitudes. Thus, a performativist may adapt this solution by arguing that moral utterances actually express *two* states: one involving a belief about the world, the other involving the moral attitude. The belief and its satisfaction conditions (and the semantic value of the part of the expression corresponding to it) will be distinct from the moral mental attitude and its satisfaction conditions; the former will be mind-to-world, as in the case of all beliefs (and assertions), while the latter will be world-to-mind, as in the case of intentions and commands. On this account, the two states will be related in the following sense: error in the former may lead to

some kind of failure in the latter. If I say, "Your torturing that person is morally wrong!" and my statement relies on the belief that you are torturing someone but, as it turns out, you are only tickling the person, then the error of my belief will have led to some failure in my use of moral terms. We can examine this claim a bit more closely by adding to our previous notational scheme

ωp, which stands for 「I morally disapprove of p」.

According to the two-state strategy, $\neg ß p \supset \neg \omega p$—that is, no moral disapproval without some antecedent belief in what's going on. And if I come to learn that $\neg p$ (i.e., if I acquire the belief $ß \neg p$), then $ß \neg p \supset \neg \omega p$ (i.e., if I believe it's not torture, I won't morally disapprove of it as being torture).

While this strategy makes some inroads to locating the mind-to-world satisfaction conditions for moral mental states, it is inadequate on the whole. First, it assumes that the only way we can say someone should change his/her moral views is if we can point to a falsehood in his/her beliefs. Stevenson may be correct, descriptively, that this is the case. But surely we sometimes want to say that someone is just *wrong* or *should* change his/her views, independently of whatever he/she believes. Suppose, for example, someone understands that torture is being performed, understands all the other facts associated with it, and yet he/she *still* morally approves of it. We want room to say that factual beliefs underdetermine moral views such that *we* can say of him/her that his/her moral views are wrong, facts not withstanding.

Moreover, the two-state view seems rather implausible in light of our more general views about moral mental states and their expressions. If I say, "The earth is

solid and round," we normally take this to be the expression of a single mental state: *that* the earth is solid and round. We don't treat it as two separate mental states: *that* the earth is solid and that the earth is round. Why not treat, "Your stealing his wallet is wrong" in the same way as non-moral mental states: *that* you stole his wallet and were wrong to do so. There doesn't seem to be any *independent* linguistic evidence for carving up speech acts and mental states in the way the two-state solution does. And even if the solution works, it doesn't show that moral mental states (or the moral portions of them) have mind-to-world satisfaction conditions, only that in some ways, those states are intimately connected with mental states that have mind-to-world satisfaction conditions. Failing powerful arguments to the contrary, we have no reason to accept the two-state solution.

5.3. The Two-Logics Solution

Perhaps in response to worries that performativism leaves something out—and also in response to the Frege-Geach problem (§3.2.3)—contemporary performativists have found it attractive to develop formal models for regimenting the moral attitudes. Blackburn's (1988) "logic of the attitudes" is the most prominent of these attempts, but there are others.[88] The common thrust of these attempts is to show that moral utterances do have truth conditions, or rather that the moral components of those utterances have truth conditions in respect of formal constraints on which pairs of attitudes one can hold. For example, if one believes that stealing is wrong *in general*, one cannot, in consistency, believe that a

[88] See Horgan and Timmons (2006) and Gibbard (2003) .

particular case of stealing would be okay. Similarly, if one is committed to the view that all human beings have rights and the view that fetuses are human beings, then one cannot believe that abortion is permissible.

At first glance, this is a major deviation from the performativist accounts of Ayer and Stevenson, who denied steadfastly that moral utterances had any truth conditions (beyond their non-moral components). On the other hand, this line of thinking is consistent with performativism in the following respect: the truth conditions of the moral components are given in terms of a logic of moral attitudes, which differs from, say, the logic of beliefs. Moral language remains *sui generis* and irreducible to non-moral language; a separate logic and a separate class of truth are established for the attitudes. Whereas the previous solution we examined attempted to account for truth-conditional aspects of moral judgments by positing two states behind every moral utterance, this solution employs two logics to achieve the same result.

I will not belabor the discussion here by examining Blackburn's logic of the attitudes in depth. I already presented the basic operators in §1.1.2, and the notion that there is a clash between attitudes of moral approval and moral disapproval is intuitive enough, as are the notions of entailment and that of approval logically entraining approval and so on. What seems most interesting is Blackburn's attempt to widen logic by discussing *all* attitudes—or at least attitudes of belief—in terms of commitments, just as he would moral attitudes: "To avow something of the form 'If p then q' is to commit oneself to the combination 'Either not-p or q'" (1998, 72). Thus, even assertoric statements express one's commitment to something, and this

notion of commitment, can in turn, make the notion of moral *beliefs* more palatable for the performativist: "Subtlety with the concept of belief, or with the concept of truth or of fact, may enable the expressivist to soften this opposition [to moral beliefs]. Theory may enable us to understand how a commitment with its center in the expression of subjective determinations of the mind can also function as expressing belief, or be capable of sustaining the truth predicate—properly called 'true' or 'false'" (1993, 185).

These insights are interesting in respect of their capacity to solve the hybrid semantic problems I pointed to in §3.2.3. There, I argued that the problem of reconciling descriptive and non-descriptive discourse did not fall in the realists' favor, for there are a number of speech acts (e.g., commands) that are non-descriptive and any viable semantics should be able to handle mixed statements. Performativists, particularly Blackburn, seem sensitive to this problem, and analyzing all commitments—beliefs, imperatives, and anything else—in more general terms may help to achieve a unified semantics from the start. This would amount to a "one-logic" solution to the performativists' problem and provide a real advantage against realists who seem incapable of making sense of non-descriptive discourse on their externalist accounts.

Where the two-logic or even the one-logic approach seems to break down is that it is a trivial solution to the problem we have been examining—"trivial" not because it is meaningless or even obvious but rather because it fails to address the more substantive question at hand. In raising the issue of satisfaction conditions for moral mental states on performativist accounts, we were not demanding a

formal analysis or a logic of the attitudes. We were not asking whether committing oneself to something in light of one's other commitments was valid or appropriate or even true. Instead, we were asking whether the performativist could make sense of mind-to-world satisfaction conditions. Merely demonstrating that there is an inner logic to the moral attitudes will not suffice to answer this question because, to put it simply, it says nothing about the world. To hold inconsistent attitudes and be frustrated in practical matters is, as Crispin Wright alleges, a "moral failing" (1998, 33).[89] It is not necessarily to get anything wrong with respect to the way the world really is; it is not necessarily to hold incorrect beliefs, only mutually unsupporting ones. A revisionary account of 'belief' or 'truth' may be desirable, but the question will *still* arise: What is the relation between the mind and the world with respect to moral attitudes? Until the performativist comes to terms with *this* question, his/her account remains incomplete.

In the next section, I survey Blackburn's own attempt to come to terms with this relation: quasi-realism. I think it diverges significantly from performativist methodology, even to the point that it collapses into realism proper. I do not deny that an alternative account could be developed that would retain the key points of performativism. But I think it will become clear by the end of the next section that performativism is better regarded as an adjunct to some other metaethical methodology than a robust framework in itself.

---

[89] Wright's objection is actually that Blackburn has failed to demonstrate that there is any logical inconsistency in a "clash of attitudes." Wright was then objecting to Blackburn's 1984 account given in *Spreading the Word*. By 1993, Blackburn had revised is account along the lines in which I have presented it here. Still, I think Wright could object that a mistake on that formulation amounted to a moral failing, not a failing to make one's mind like the world.

5.4. Quasi-Realism

Suppose we grant that some general mental faculty for planning and evaluation could have evolved in humans[90] and that social conventions for expressing those moral mental states could have developed such that using moral terms would create a binding force on listeners. These two assumptions are crucial for performativist accounts, for they establish the possibility of moral attitudes and the performative effects of moral expressions. The performativist still owes us an account of *why* moral attitudes and their corresponding expressions occur when they do—not why we have moral mental states at all, but why we form the particular moral mental states that we do when we do. This question lies at the heart of the connection between the mind and the world, and I think a supervenience objection along the lines I sketched in the previous chapter is in the offing for performativists.

Suppose people routinely have moral mental state M when presented with certain situations (e.g., torture, rape, or unnecessary violence). They react to these situations with feelings of moral disapproval and express their moral disapproval by saying things such as "That's wrong," "How immoral," or "What a bad thing," and thereby create an influence on others not to engage in such behaviors. Once this connection between certain situations and certain responses becomes robust enough, it will be plausible to speculate that M supervenes on certain aspects of the situations in question. These aspects could be stance-independent or stance-

---

[90] For an extended discussion of how evolutionary facts bear on our knowledge of moral truths, see Street (2006).

dependent—the important point is that a pattern or regularity will start to form across these various moral situations and their corresponding attitudinal responses in observers.

To some extent, this is exactly as it should be. It would be implausible (perhaps immoral) to suggest that observers' moral responses should exhibit wild variation across different cases. Quite the contrary, we could only hope that their reactions should be so consistent as to generate this problem! Nonetheless, once we are able to talk about certain moral facts or properties as eliciting responses in observers, the performativist is faced with a new issue: deviation from this baseline will start to appear incorrect, wrong, or somehow misguided, which is just to say that we will start to regard moral attitudes as having mind-to-world satisfaction conditions. Again, this is exactly as it should be, for the main difficulty we have seen with performativism is that it has little to offer theoretically with respect to these conditions. The problem, however, is that once these moral facts and properties are on the scene, performativism starts to collapse into realism or some other kind of metaethical account. To appreciate this point more fully, let us pause for a moment and survey a few possible ways that this could come about.

Suppose that there are no moral facts or properties antecedent to the first occurrence of someone's having a moral attitude. Once that person expresses the moral attitude, he/she creates an influence on others, and suppose they, too, have the same moral attitude with respect to the same situation and express it and create the same influence on others, and so on. This "contagion," as Stevenson (1963, 29) calls it, could explain how awareness of stance-*dependent* moral facts and

properties could propagate throughout a linguistic community. There may be some deviant individuals in whom different moral attitudes persist, but the same thing could be said of false beliefs, and, at least in the moral case, it is even more likely that their attitudes will eventually be swamped out, given the special force that attends speech acts involving moral terms.[91] The problem with this approach seems to be that there is no constraint on what a community could approve or disapprove of, morally speaking. The mere effect of contagion might ensure that *any* moral attitude that finds itself in the initial position gets propagated throughout a community. Blackburn also argues that this stance-dependent view of moral facts and properties is wrongheaded:

> Values are the children of our sentiments in the sense that the full explanation of what we do when we moralize cites only the natural properties of things and natural reactions to them. But they are not the children of our sentiments in the sense that were our sentiments to vanish, the moral truths would alter as well. The way in which we guild or stain the world with the colours borrowed from internal sentiment gives our creation its own life, and its own dependence on facts. So we should not say or think that were our sentiments to alter or disappear, moral facts would do so as well. (1984, 219, note 21)

On my view, this line of thinking actually leads to a stance-dependent view of moral facts and properties. If I am right, then some kind of interactionist or

---

[91] This contagion explanation is also consistent with the stance-independence of moral facts and properties, assuming the moral attitudes line up correctly with the facts and properties.

constructivist account is appropriate, and performativism serves merely to explain how moral language propagates moral attitudes across members of a linguistic community. However, we can at least *consider* the suggestion that moral facts and properties are stance-independent, as quasi-realists claim.

On this approach, we could say that the members of the community in our example were becoming sensitive to stance-independent properties in the right way or learning how to respond correctly to certain moral situations. This route finds touchstones in the literature on response-dependent or secondary properties.[92] Just as observers can train themselves to become more sensitive to small variation in hues of color or subtle differences in the taste of wine, so too could *moral* observers be said to become more sensitive to moral facts and properties.

Alternatively, one might approach stance-independent moral facts and properties by saying that we are constrained by certain aspects of human nature in the kinds of evaluative stances we can take towards situations. This is Blackburn's own quasi-realist solution to the problem of mind-to-world satisfaction conditions: "Just as the senses constrain what we can believe about the empirical world, so our natures and desires, needs and pleasures, constrain what we can admire and commend, tolerate and work for. There are not so many livable, unfragmented, developed, consistent, and coherent systems of attitude" (1984: 197). We cannot find just *anything* valuable or abhorable—our moral attitudes must be

---

[92] For the classic exposition of moral properties as secondary properties, see McDowell (1984).

constrained by facts about our natures, interests, and potentials. Let us amend our formulation of performativism to account for this quasi-realist insight:

PERFORMATIVISM' There is a set of stance-independent facts $\{F_{i-j}\}$, which constrain the possible moral attitudes one can consistently and coherently adopt. When a speaker S makes a moral utterance U about an object or event, U expresses S's moral attitude toward that object or event, creates a moral influence on others' attitudes and their actions, and (optionally) expresses S's other, non-moral attitude(s) toward that object or event. S's attitude is evaluable with respect to its consistency and coherence with $\{F_{i-j}\}$.

Though this proposal adds some stance-independent facts (and their attendant properties) into the mix, it is not clear that it makes *moral* facts and properties stance-independent. Consider the metaethical non-naturalist who believes that certain facts about the physical world constrain how non-natural moral properties are instantiated in the actual. One can imagine Moore or Shafer-Landau saying that it's a contingent fact that goodness and pleasure always co-occur in the actual world, but that does not show that goodness inherits any of the *qualities* of the natural facts that constrain it. Similarly, just because the facts that constrain moral attitudes are stance-independent, it does not follow that the *content* of moral attitudes references stance-independent moral facts and properties. After all, the very process of having moral attitudes will be constrained by certain natural,

stance-independent facts about the brain, but that does nothing to show that the content of the moral attitudes *themselves* countenance stance-independent or natural moral facts and properties. Blackburn's quasi-realism, as formulated above, seems to confuse implementation-level questions about moral mental states with content-level questions about the nature of moral facts and properties.

Still, there is a route to stance-independent facts and properties on this suggestion, and to see that route, let us resume our supervenience objection to performativism. Along either the stance-dependent or stance-independent routes discussed above, members of a linguistic community could be trained *when* to use moral terms and, correspondingly, *when* to have certain moral attitudes. Now, suppose we want to investigate whether someone is correct in holding a certain moral attitude. We will begin by querying the mind-to-world satisfaction conditions for that attitude, and, depending on whether the facts and properties in question are stance-independent or stance-dependent, we will adopt an externalist or an internalist or interactionist methodology for examining moral mental states. What, then, is the special purchase of performativism?

As long as there are no such things as moral facts and properties—or, to put the point more precisely, as long as the performativist is correct that our moral attitudes have no mind-to-world satisfaction conditions and metaethical questions are merely normative advice about what facts to take into our plans and preferences—then performativism seems to be a distinctive metaethical contribution. But once we leave behind our performativist beginnings and begin to talk about moral attitudes as responses to moral facts and properties, the theory

does not seem to compete with other metaethical frameworks; it is surpassed by them.

In Chapter 4, I marshaled this supervenience objection to question whether moral facts and properties really were stance-dependent for constructivists. Here, I am using the same broad argument to allege that performativism isn't, at the end of the day, a competing metaethical framework. As soon as it countenances—as it should—the regularity of moral attitudes across different situations and different subjects, it gives way to *other* metaethical approaches. I think that Blackburn hints at this much in his first attempt to define quasi-realism: "I call the enterprise of showing that there is none [no mistake in the expressive and truth-conditional origins of our evaluative practices]—that even on anti-realist grounds there is nothing improper, nothing "diseased" in projected predicates—the enterprise of *quasi-realism*. The point is that it tries to earn, on the slender basis, the features of moral language…which tempt people to realism" (1981, 171). On one interpretation of this passage, Blackburn is conceding that a realist view of moral language is accurate but needs justification or clarification by means of the performativist framework. The performativists' moral attitude (what he calls "projectivism") provides little more than the equivalent of a proof in constructive logic, in which no statement is true until there is a *proof* that it is true and no statement is false until there is a *proof* that it is false. Blackburn himself notes this analogy in discussing Wittgenstein's approach to mathematics:

> It is clear that what he wants to do is place mathematical practice, not as a representation of the mathematical realm, but as "a different kind of

> instrument," commitment to which is not like central cases of belief but much more like other kinds of stance….The proposition expresses a norm that arises in the course of human activities, but does not describe those activities, and it has no use in which the correctness of the norm (the truth of the proposition) depends upon the existence of form of those activities. *That* question cannot be posed; it treats what is not a dependent state of affairs belonging at all to the natural world as if it were. (1993, 175)

For Blackburn, it seems there are no "earned" moral facts and properties until a speaker has a moral attitude and expresses it using the conventions for moral speech acts. Once this framework is in place and different speakers can counterspeak, as it were, and engage in normative debate, *then* we are entitled to talk about the truth and falsity of moral mental states with respect to mind-independent moral facts and properties. What performativism seems to do on this account is nothing more than clear the ground for moral realism—important groundclearing, to be sure, but groundclearing all the same.

This collapse of quasi-realist performativism into mere realism has been noted by several of Blackburn's critics. Crispin Wright goes so far as to pose a general dilemma for the quasi-realist:

> The goal of the quasi-realist is to explain how *all* the features of some problematic region of discourse that might inspire a realist construal of it can be harmonized with projectivism. But if this program succeeds, and provides *inter alia*—as Blackburn himself anticipates—an account of

> what appears to be ascriptions of truth and falsity to statements in the region, then we shall wind up—running the connection between truth and assertion in the opposite direction—with a rehabilitation of the notion that such statements rank as assertions, with truth conditions, after all. Blackburn's quasi-realist thus confronts a rather obvious dilemma. Either his program fails—in which case he does not, after all, explain how the projectivism that inspires it can satisfactorily account for the linguistic practices in question—or it succeeds, in which case it makes good all the things the projectivist started out wanting to deny: that the discourse in question is genuinely assertoric, aimed at truth, and so on (1988, 35)

Though the battle over metaethics shouldn't be won by what kind of *–ism* a position is made out to be, Wright seems correct to bring out the sense in which quasi-realism, if it succeeds in establishing realism, winds up simply *being* realism (not unlike what we saw with constructivism and the supervenience objection in the previous chapter).

As I argued in §5.3, cognitivity or truth conditions are not the interesting issue in metaethics. There is a trivial sense in which truth is relative to any set of ontological commitments. Ultimately, one's metaethical account is going to bottom out in terms of how the world is (or isn't). And then we will want to know about what qualities of the world occasion our moral mental states and our all-important moral utterances, and we will want to know whether those qualities are there, independently of us, or whether they are products of our own intentional stances.

Either moral facts and properties are stance-independent or they are not. If they are—and this seems to be contemporary performativists' preferred solution—what's the upshot of performativism? Blackburn says we've earned the right to our realism. If so, performativism provides the basis for realism, but I think that project will ultimately run afoul of the troubles I raised in Chapter 3. If moral facts and properties are not stance-independent, what's the upshot of performativism? The existence of a convention that would allow us to construct these properties is interesting, but it will always come down to individuals' attitudes and their expressions of those attitudes—and the attendant satisfaction conditions of their content.

Unless the performativist has some way of telling us whether someone's utterance is called for or not (the correlate of 'true' here), I don't think we learn much about the nature of stance-dependence. It will turn out that either constructivism (some kind of idealist analysis) or interactionism (some kind of empirical analysis) is correct. How do reflections on performativism aid us in this question? Once performativists establish moral facts and properties, they seem to adopt other metaethical approaches. But without those approaches, the performativist has no robust way of making sense of mind-to-world satisfaction conditions.

5.5. Conclusion

As I have argued in this chapter, performativists have not done enough to show that their account can accommodate the mind-to-world satisfaction conditions of

moral mental states. This task would be made easier if performativists gave a robust account of what they take the moral attitude(s) to consist in and what satisfaction conditions hold for that attitude(s). In giving this account, they will need to show that the mind-to-world satisfaction conditions for the use of moral terms (or for having thoughts with certain moral content) is not just the performance conditions for the use of those terms. The attempt by quasi-realists to add stance-independent moral facts and properties into a larger performativist framework is an important move in this direction, but if performativism is to remain a competing metaethical framework, performativists must show that appeal to stance-independent facts and properties does not commit them to garden-variety externalism.

Chapter 6

Metaethical Interactionism, or Morals by Experiment

> Constant and effective interaction of knowledge and practice is something quite different from an exaltation of activity for its own sake. Action, when directed by knowledge, is method and means, not an end. The aim and end is the securer, freer and more widely shared embodiment of values in experience by means of that active control of objects which knowledge alone makes possible. From this point of view, the problem of philosophy concerns the interaction of our judgments about ends to be sought.
>
> —John Dewey, *The Quest for Certainty*

In Chapter 2, I outlined three existing methods in metaethics—externalism, internalism, and performativism—each of which I discussed in greater detail in the previous chapters, with special reference to its ability to meet the criteria of adequacy for moral mental states we established in Chapter 2. Each of these methodologies aims to answer fundamental questions about the relation between the mind and the world with respect to moral mental states. In the broadest sense, we can see each approach providing a different answer to the question, Should we begin metaethical investigations with the mind or the world?

- Those who answer, "World," are externalists, attempting to discover stance-independent moral facts and properties.
- Those who answer, "Mind," are internalists, examining the nature of moral experience for confirmation of normative principles.

- Those who answer, "Neither, attend to the nature of moral language itself" are performativists, rejecting other attempts as unearned.

A possible approach that is not reflected here is the answer "Both," on which morality arises as a function of the interplay between minds and the world and metaethical questions can only be addressed by taking into account that relation. I call this approach "metaethical methodological interactionism" ("interactionism" for short) for the special emphasis it places on the interaction between moral mental states and conditions in the world that occasion them. In this chapter, I want to sketch one possible interactionist account and argue that, in the absence of clear solutions to the objections raised in the previous chapters against the other methodologies, we should provisionally endorse it as our working metaethical framework.

It is doubly hard to defend the adequacy and autonomy of any hybrid theory. Such a theory, it will be alleged, inherits all the central problems of its constituent theories. Worse than that, where it succeeds, it will be said to do so on the basis of only one of its constituents, making the need for a hybrid appear altogether spurious in the first place. If it finds touchstones in the literature, it will be cast as the same old theory under new guise; where it diverges, it is likely to be called irrelevant.

These are tough demands for interactionism to meet—perhaps impossible. But I do not put it forward on its own accord, as it were. This chapter presents only a statement of the position and preliminary considerations in its favor (e.g., that it

is not self-defeating, that it can provide a framework for normative theory, that it meets certain constraints on metaethical theory that most find intuitive). The full argument for interactionism also turns on the *inadequacy* of existing metaethical approaches, including moral realism, constructivism, and expressivism—an argument that I have developed at length over the preceding chapters. The failure of these theories to present a compelling account of moral mental states urges us to consider alternatives, among them interactionism. A full defense of interactionism is beyond the scope of this project. My intent here is only to establish the broad framework for an interactionist theory and to present one such account. Some interactionists may follow my own emphasis on the methods of science and empirical investigation; others may follow alternative routes. Just as there are many moral realisms, so too could there be many interactionisms. At the end of the day, each metaethical account must be assessed on its own merits, particularly on its treatment of moral mental states, if my overall stance here is correct.

6.1. A Sketch of Interactionism

When Descartes and his colleagues banished alchemy, magic, and superstition from the scientific realm, they did not intend to exile morality as well. But ever since the rise of scientific materialism, philosophers have been hard-pressed to find a home for value in the world of material substance. Sellars (1962) once captured this tension in describing two different images of humankind: the "manifest image," which focuses on persons as beings who conceive of themselves as sentient

perceivers, cognitive knowers, and deliberative agents; and the "scientific image," which presents some sophisticated account of "atoms in the void." While the former is home to agents, meanings, and values, the latter reflects the dream of nineteenth-century French mathematician Pierre Laplace—the world reduced to fundamental physical particles, computed by a sufficiently powerful machine. The apparent difficulty in reconciling these two images has led many to conclude that the fundamental problem in the foundation of ethics—the fundamental problem of *metaethical* inquiry—is how to make sense of the so-called world of value within a general scientific framework.

If we are not to be externalists—beginning with investigations of the world—or internalists—beginning with investigations of the mind—where *shall* we begin? In the rest of this chapter, I argue that a scientific approach to questions of value is possible, adequate, and even desirable, so long as that approach takes seriously the role of the mind in determining the nature of value. As one might imagine, the challenge here will be avoiding some kind of subjectivist voluntarism according to which each individual determines what is good, right, and so on, and value on the whole (if it even makes sense to talk that way) is wildly disparate and unsystematizable. I begin by giving an account of the sense in which the mind makes a crucial contribution to value and by formulating the interactionist account I wish to defend. Following some exegetical remarks on this formulation, I consider in §6.2 the extent to which this account is scientific—that is, how the science of morals parallels the science of the natural world. In §6.3, I present three empirical theses intended to show the general unity of individual and social interests. If

these are correct, we have no reason, on the whole, to fear wild subjectivism about value on my account. As empirical theses, these need empirical confirmation, so my account here can only sketch the hypotheses and their relevance.

To begin formulating this interactionist account, we need to consider the contribution of mind to the construction of value. Samuel Pufendorf puts forward a useful image for capturing this contribution: "Now as the original way of producing physical entities is creation, so the way in which moral entities are produced can scarcely be better expressed than by the word *imposition*. For they do not arise out of the intrinsic nature of the physical properties of things, but they are superadded, at the will of intelligent entities, to things already existent and physically complete, and to their natural effects" ([1703] 1990, 171). On this metaphor of imposition or projection, *we* are responsible for the creation of moral value by adding something to material substance. Hume, too, adopts this same general stance, saying that the "productive faculty" of taste is responsible for "gilding or staining all natural objects with the colours, borrowed from internal sentiment" ([1751] 1998, Appendix I, 21).[93] While this general view of the origin of moral properties may be correct, it offers little guidance as to *how* we impose value on the world or, more importantly, how we might do so in better or worse ways. Put differently, it does nothing to address the issue of mind-to-world satisfaction conditions for moral mental states. One could impose value on the world willy-nilly, it seems, with moral consciousness amounting to little more than artistic

[93] Blackburn claims that his quasi-realist account is based in projectivism. As I argued in Chapter 5, his ambitions for securing stance-*independent* moral facts and properties are at odds with this projectivist foundation.

temperament. This line of thinking, *by itself*, seems to lead to the worst kind of subjectivism about value. More problematically, it suggests that the method of science and the method of ethics are fundamentally opposed, for they concern two completely different types of activity. Science aims to discover the world as it is; morality adds an independent world of value on top of that.

John Dewey alleges that this radical split between science and value is responsible for a number of errors in the theory of value:

> The neglect of sciences that deal specifically with facts of the natural and social environment leads to a side-tracking of moral forces into an unreal privacy of an unreal self. It is impossible to say how much of the remediable suffering of the world is due to the fact that physical science is looked upon as merely physical. It is impossible to say how much of the unnecessary slavery of the world is due to the conception that moral issues can be settled within conscience or human sentiment apart from consistent study of facts and application of specific knowledge in industry, law, and politics. (1922, 10–11)

But if we cannot locate value in the material world or in mere mental powers of projection alone, wither morality? One suggestion is to study the imposition of value using the methods of science: to generate hypotheses about the imposition of value and to confirm them by observational evidence. Such an exercise would be little more than an anthropological or perhaps sociological study of the ways in which humans project value on to the world—*unless* we could construct a framework for studying not how people actually *do* assign value, but how they

ought to. This suggestion smacks of circularity, for it suggests that the proper method for studying the *criteria* for valuing something is by studying not the actual assignments of value, but the *criteria* for valuing.

To improve this view, we might begin by considering cases in which people, by their own lights, admit to making mistakes about value. If we find similarities in this process of *correction* of values, then we might apply those same procedures antecedently to *valuing*, as it were, to avoid the mistakes in the first place. On this approach, imposing values onto the world will be a kind of skill. Before understanding this skill completely, we may be unsure what rules govern the practice. But, if we can identify cases in which it seems we have improved our practice of the skill, we can retrofit general rules that will guide our practice on future occasions. Nothing in this account presupposes that there are such rules writ into the fabric of the universe; to adopt that perspective would be to adopt a very different methodological approach. Instead, I am suggesting that by attempting cases of valuation and, through incremental revisions, attempting to improve that practice, we can begin to fashion rules governing that practice that allow us to say when a person has done better or worse in his/her performance of it. What we aim for is a rapprochement between that practice and our motives for engaging in it, noting cases of failure (and success) along the way and using that information to gradually improve our abilities.

This very method for determining the satisfaction conditions of moral mental states parallels an important feature of *normative* reasoning: the interplay of facts and values. Very often, in pursuing a certain course of action, we pursue it

under guise of achieving a certain end or bringing about a certain state of affairs. Another way of putting this point is that, in acting, we attempt to fulfill some desire, preference, or interest. These ends or motives may appear rather base, for they are often accompanied by pleasure, and some moralists will confuse this point with the claim that we always seek pleasure or, more generally, always seek to do what is in our own interest. These claims are patently false; people sometimes act without regard for pleasure or against their own interest.[94] On a more charitable interpretation of my point, we act to achieve certain ends, which may be things that benefit us or may be quite noble-minded ideals, such as respecting the rights of others. Thus, in saying that we act to fulfill certain ends, let us remain neutral and allow those ends to range from the completely selfish to the completely selfless.

The importance of bringing out the guises under which we act is that the most expedient way to change someone's views about how he/she should act (or even what ideals to hold) is by showing him/her that the specific actions he/she has in mind will not bring about the ends he/she seeks. Call this the *direct revision method for action–ends*. You thought that by performing some action you would be able to achieve a certain end, but I am able to show you that that action does not achieve that end but only frustrates it. There are also cases of *indirect revision* wherein the end at which we aim is brought about, but in experiencing that end, we come to realize it is not an end we actually want, desire, prefer, etc. You thought

---

[94] For refutations of psychological egoism, see Butler ([1726] 1827) and Feinberg ([1958] 2008).

that a certain end should be achieved, but, upon achieving it, you realize it shouldn't have been achieved. These are hard-learned lessons about the connection between what we think should be brought about and the actuality of its being brought about. There may be unintended or unforeseen consequences of our actions or miscalculations about ourselves or the world that lead to some poor fit between the way we act and the way we later judge we *should* have acted. With the long years of experience, we may come to learn that honesty really is the best policy most of the time, or perhaps that a life of material luxury is not ultimately satisfying.

The point of discussing these processes of plan-revision is to draw out the differences between *ex-ante* and *post-facto* information about the effects of certain actions. The epistemic asymmetry of these situations is that, in the former, we have only estimations of what ends will attend a particular action, while, in the latter, we have the advantage of knowing the outcome and, in hindsight, assessing whether the action was something that should have been undertaken. This point should not be confused with normative consequentialist views. On the framework I am presenting, the outcome of a particular action could be the violation of someone's rights (in the Kantian sense) just as much as it could be the production of some additional utility (in the Benthamite sense).[95] The upshot of appealing to outcomes is to give greater informational resources to the agent planning certain decisions, thereby giving him/her greater control over the things to which he/she attributes

---

[95] For further discussion of this point, see Dreier (1993) and Louise (2004).

value and pursues as such. In other words, such revisions improve the fit between the agent's desires, preferences, and interests and the achievement of them.

Taking this insight on board, let us formulate a working view of interactionism as:

INTERACTIONISM When a speaker S makes a moral utterance U about an object or event, S asserts the hypothesis that, in the nearby possible world **W** where that object or event is brought about, S' would approve (disapprove) of it, and tries to influence others to approve (disapprove) of it.[96]

Roughly, the account says that when a person claims a certain action is right, he/she is asserting that, were that action undertaken and information about its outcomes available, he/she would *still* approve of the action; this assertion amounts to a hypothesis about the speaker in the nearby possible world where the action is performed. This formulation can be permuted to generate similar definitions for 'permissible', 'obligatory', 'wrong', and so on (e.g., when a person claim a certain action is *permissible*, he/she is asserting that, were the action

---

[96] The account presented here is indebted to Welchman's reconstruction of Dewey's ethical theory, especially the material presented in the *Ethics* (1908). Welchman gives a vivid example of an adult woman faced with the option to persuade her aging mother to move out of a physically challenging home into one where she would be more safe and convenient. Welchman concludes: "If the actual consequences of Jane's decision confirm her predictions, Jane may treat that as empirical confirmation of her original judgment and may consider a judgment to continue as she has begun both provisionally and experimentally warranted. If, however, the consequences do not confirm her predictions, her original decision, though justifiable given the circumstances, must now be judged unreliable and a decision to continue as she has begun unjustified. Jane's deliberation must be resumed from scratch" (1995, 180).

undertaken and information about its outcomes available, he/she would still *tolerate* the action).

In addition to these points about the reference of moral terms, the account claims that the speaker's utterance has a certain prescriptive force that recommends that course of action to others. For reasons I discussed in Chapter 5, this performative analysis of moral speech acts is persuasive and worth incorporating into any metaethical account. But where performativists were ill equipped to make sense of the mind-to-world success conditions of moral mental states, this interactionist account addresses that deficiency by adding that moral utterances are also assertions about hypothetical valuations (under conditions of improved information), and therein lie standards for assessing the correctness of the moral mental states expressed in those utterances.

Having clarified these initial matters with our interactionist account, we can now consider two further elements specific to this account. First, we must consider what role *prediction* has on this interactionist theory (i.e., why we should regard speakers as putting forward hypotheses about what they would value under conditions of improved information). A second issue to consider is the relation between the speaker S on this theory and his/her counterpart S* in **W**, who has the benefit of additional information about the outcome of the objects or states of affairs recommended by S (i.e., what is gained by shifting the analysis to S*). Put differently, this account is fundamentally based in possible world semantics, and we must consider what constraints there are on stipulating facts about our counterparts. What special contribution do our more informed counterparts make

to our overall analysis of the imposition of value on an antecedently valueless world?

6.1.1. Predicting Hypothetical Valuations

On the version of interactionism presented here, a speaker's use of moral terms reflects his/her predictions about what he/she would value, given improved knowledge of the outcome of those actions. Prediction here plays the role that practical reasoning or the application of a certain decision procedure plays in normative theories. The speaker must take into account certain information about the action under consideration, constraints imposed by conditions in the world, human psychological tendencies (e.g., status quo bias, endowment effect, confirmation bias), and his/her own desires, preferences, and interests. Where this process seems to diverge most from standard normative accounts is in its notion of consistency. To take two examples, Kantianism and utilitarianism both require that the agent apply a consistent set of principles or a consistent decision procedure in the process of deciding what to do. To some extent, this seems to require that S gives a guarantee, as it were, that S* would act exactly the same way because S and S* must be taken to apply the same set of criteria in their imposition of values. The reason this standard requirement of consistency will not work here is that the whole point of averting to this hypothetical valuation is to allow the valuations of S* to diverge from those of S such that S can learn how to make *better* valuations. To put the point more directly, interactionists don't want S* to make the same mistakes as S. The only way interactionism can provide normative guidance is if the

valuations of S* are allowed to diverge from S and, by that process, allowing for improvement in the imposition of value.

To assess the scope of this consistency problem, we need to get clear on the demands placed on interactionism in virtue of its being a metaethical theory. First, interactionism should give us some guidance about which normative theory to adopt. I was careful to argue in Chapter 2 that this guidance should not amount to question-begging assumptions about which normative theory is correct. Instead, it should establish a framework within which normative theories may be developed and assessed. To take a parallel example, nothing in a moral realist's account pre-judges the question of whether, say, Kantianism or utilitarianism is correct. Moral realism gives us a framework for discovering moral facts and properties, and normative theories will succeed or fail according to whether they employ a correct notion of these facts and properties. Some normative theories will, presumably, be ruled out (if moral realism is true) once we discover some moral facts and properties. Others may have to await adjudication until we have a richer notion of moral facts and properties.

Similarly, interactionism as a metaethical theory should not be assessed in the same way that a normative theorist would assess a moral agent according to whether he/she consistently applies some favored moral principles or procedure. Interactionism is an incremental theory that approximates moral truth in the limit of hypothetical valuations. A single case of an agent's predicting what he/she would hypothetically value (under conditions of improved information) does not constitute the whole moral truth about some object or event. We will need

repeated attempts at valuation and prediction—probably by other agents and probably even accumulated over several generations (I discuss these in §6.2)—to get a sufficiently reliable notion of truth for a given object or event. Just as the moral realist may not discover *all* moral facts and properties in a single day, so too should the interactionist not be expected to generate robust criteria for valuation in a single valuation. Our expectations for interactionism should be the same as our expectations for standard science: we *treat* as true whatever hypotheses are most well confirmed in the present, and we reserve the final judgment for whatever hypotheses are most well confirmed at the end of scientific inquiry. On an interactionist account, we model our working criteria on whatever valuations are stable in the face of improved information, and we reserve final judgment about the criteria for whatever valuations are left at the end of all hypothetical valuation. It may turn out that at present or in fifty years, we can regard utilitarianism or Kantianism as meeting the criteria for correct valuational assignments. More likely (and as is the case, historically), we can expect incremental revisions to those theories as we learn more about what their outcomes would entail. So fluctuation in hypothetical valuations is not just permissible in the process of assessing normative theories; according to interactionism, it is required for moral progress.

6.1.2. Improved Moral Agents

In shifting the context of valuation from an agent in the actual world to his/her counterpart in the nearby possible world where some object or event is brought

about, interactionism locates the determinant of correct valuations in the attitudes of hypothetical agents. This shift has two important implications.

First, it means that moral facts and properties are stance-dependent. We can treat moral properties as logical constructs out of the attitudes of S*. Whatever S* approves of has the property of goodness or rightness; correlatively, whatever S* disapproves of has the property of badness or wrongness. These properties depend on the attitudes of S* and are therefore stance-dependent. What is gained by shifting the stance, as it were, from S in the actual world to S* in **W** is an improved epistemic situation with respect to the fit between the object or event under consideration and S's desires, preferences, and interests. S* is in a better position to know whether the object or event will actually further those desires, preferences, or interests because S* has occasion to observe the outcome of that object or event. The response that S* makes, if it diverges from S, may be of one of two types: changing the object or event to achieve a better fit, or changing the desires, preferences, and interests themselves. Thus, in saying that S* can better observe the outcome of the object or event, I am saying that S* has potential access *both* to the *fit* between it and what S desires, prefers, or has interest in and also, in the case where fit is achieved, to the actual *experience* of having S's original desires, preferences, or interests fulfilled. The former case of change I called *direct revision*; the latter case I termed *indirect revision*.

Second, in shifting the analysis to possible worlds, we gain an important tool: stipulation. Not only do we stipulate the fact that the object or event is actually brought about in that world, we might also stipulate changes in S* that

might place S* in a better position to approve or disapprove of the object or event. Where such stipulation will run afoul is if we stipulate S* to have different normative views than S *antecedent to S*'s looking at the outcomes*—that will beg the question against S's original valuation—or where we load in assumptions about idealized agents. In §4.3.2, I criticized constructivism for its assumptions about idealized subjects. A key difference between interactionism and Kantian constructivism is what additional stipulations are made about **W** and S'. Kantian constructivists take **W** to be the kingdom of ends in which idealized agents both legislate and are governed by rules of action. In addition, idealized agents are stipulated to only be concerned with moral motive and acting from a regard for the moral law, never with the outcome of their actions.

The hypothetical agents of interactionism are *improved*, not idealized moral agents. In principle, a normative theorist could build in more and more improvement to the point that they are idealized moral agents. But, more plausibly, S* is just like S, only situated in a better epistemic position *vis-à-vis* his/her access to information about the outcome of the object or event under consideration. An interesting puzzle here concerns the point at which we consider S's counterpart. Is the possible world in which S* makes his/her valuation supposed to "occur" immediately after the object or event comes to be? Or some time later? Or perhaps even at the end of inquiry, at which point all outcomes could be known? Consider, for example, a teenager who believes that smoking is in his/her interest in the near future—or even at age 40—even though smoking regularly may be against his/her interest by age 70. We surely want to say that S* in this case is not S in the nearer

future, for that fails to account for possibly relevant outcomes further down the road. On the other hand, we do not want to project S* billions of years into the future to consider the effects, in the fullness of time, of an individual electing to smoke. To do that, it seems to me, would to commit us to some kind of idealized subject—idealized in the sense that he/she has access to *vastly* more information than the agent himself/herself has. A reasonable cashing out of the *ceteris paribus* clause would likely involve no more than average human capacities ranging over average human lifespan, with both averages indexed to the socially salient categories (e.g., sex, race, class) into which an individual is born. A full analysis of this problem, however, requires further inquiry.[97]

For the moment, I wish to examine one further advantage of stipulation in this context: it may be desirable to stipulate that S* has additional resources, computational ability, or time to complete his/her valuation compared to S. One could, for example, recognize that his/her situation requires making a valuation that is likely suboptimal in these respects and allow that S* has advantage of these enhancements. Moreover, one might want to stipulate that S* is without some recalcitrant trait or attitude that plagues his/her own valuations. To the extent that the person can plausibly eliminate this trait or attitude and, realistically, come

---

[97] Genetic and developmental effects may have significant impact on the prospects and plans of any one individual. Interpreting the *ceteris paribus* clause in this individualistic way would greatly hinder, if not prevent, the entire counterfactual analysis from going through. Instead, I suggest that comparisons involve average life expectancies and capacities, recognizing that any deviations would significantly alter the results. These averages will vary from time period to time period (e.g., fourteenth century vs. twenty-first century), place to place (e.g., developed vs. developing country), and even social groups (e.g., poor African-American female vs. affluent white male). The term 'socially salient' is used here to prevent the construction of arbitrary or non-standard groups (e.g., all people with a genetic predisposition to lung cancer) that would draw the analysis back to an individualistic approach. It is important to note, however, that what is a salient grouping will itself shift over time.

around to having the same attitudes as S*, this stipulated improvement is unproblematic. The *general* constraint on stipulation is that the more the situation of S* ceases to resemble conditions in the actual world, the less likely it is that actual agents will be able to correctly assign values (i.e., to meet certain demands of morality) in accordance with their counterparts. A normative account within interactionism has the burden of showing that its stipulations have purchase in the real world by allowing *us* some possibility of moral improvement. At minimum, though, interactionism requires that S* have epistemic benefit of additional information about the outcome of the objects and events in question, both their fit with his/her desires, preferences, and interests and, if the fit is right, the experience of actually having those fulfilled.[98]

6.2. A Science of Morals: Hypothesis and Confirmation

In the previous section, I suggested that one way of bridging the apparent gap between scientific materialism and the world of value was by studying the process of valuation scientifically—that is, by applying the methods of science to the

[98] Kahneman and Sugden's (2005) work on "experienced utility" is useful in this regard. Rather than employing rational decision theory or preference maximization, they argue that an understanding of people's actual attitudes towards having certain experiences are crucial in making good economic and policy decisions: "Even when participants in contingent valuation surveys report hypothetical choices rather than attitudes, their responses may take the form of choices that cannot be rationalised in terms of consistent preferences, and hence cannot be translated into decision utility" (167). In particular, they note two errors in affective forecasting: use of "transition heuristics" that consider hypothetical states as arising out of present ones and fail to take account of adaptation effects, and the focusing illusion, which causes participants to put undue stress on the issue presented. They cite an example from an earlier study by Kahneman and Snell (1990) in which students were asked to predict their happiness over eight consecutive days of eating their favorite ice cream flavor while listening to rock music: "The daily repetitions of the experience produced some substantial changes in ratings—mostly, but not always, in a negative direction—but, overall, the correlation between actual and predicted changes in liking was close to zero" (2005, 170). Like my interactionist account, this account is rooted in actual experiences rather than these forecastings. As I explain in §6.2, my account is more tolerant of such forecasting when errors can be expected to be minimal (e.g., similar subjects, time-tested maxims that hold for large numbers of people, theoretical knowledge).

process of criticizing and improving our valuations. In discussing a "science of morals," as it were, we need to be careful to distinguish the *objects* of (natural) science from the general *method* by which scientists pursue their investigations. Dewey puts this point rather clearly in discussing scientific treatments of morality: "the term "science" is likely to suggest those bodies of knowledge which are most familiar to us in physical matters; and thus to give the impression that what is sought is reduction of matters of conduct to similarly physical or even quasi-mathematical form. It is, however, analogy with the method of inquiry, not with the final products, which is intended" (1903, 116). According to Dewey, the method of scientific inquiry requires that judgments about phenomena follow from an accepted set of descriptions and hypotheses about a given subject matter and that they be confirmable empirically. Accordingly, the sense in which interactionism strives to be scientific is that judgments about the criteria for valuation (i.e., the mind-to-world satisfaction conditions for moral mental states) should follow from certain definitions (i.e., of 'object', 'event', etc.) and hypotheses about valuations and be subject to empirical confirmation. Issues about hypotheses and the process of predicting the attitudes of counterparts were discussed in the previous sections. In what follows, I want to consider what empirical confirmation amounts to according to interactionism.

First, it should be noted that the hypotheses in question concern nearby possible worlds. Accordingly, "observation" will play a slightly different role in our moral science that it does in natural science, in which observations are made of the actual world. In the case of natural science, observation usually occurs in

laboratories with specialized instruments. Here, observations will be more commonsense observations about counterfactual situations in which one's counterpart has access to additional information about the outcome of certain plans. While we do not observe possible worlds with powerful telescopes, there is no problem, in principle, with being able to say what would happen in this counterfactual situation any more than there is with other counterfactual situations, or past or future events. If you think you can know (or at least reasonably predict) what your life would have been like had you slept in this morning or what happened in the distant past or, most difficultly, what will happen tomorrow, then you should have no problem, *in principle*, with the idea of the relevant counterfactual for interactionism.

As a matter of course, S *will* predict that S* will have an attitude that corresponds to S's own valuation with respect to some object or event. You will, as a matter of fact, hypothesize that your counterpart will approve of the things you believe to be valuable, disapprove of the things you find deleterious to value, and so on. If this were not true, we would doubt that you actually believed in or imposed that value in the first place. The crucial contribution of adding the hypothetical valuation is that it allows for S to be *wrong* about his/her valuations with respect to certain actions or certain desires, preferences, and interests. By saying that moral utterances are, in part, assertions about hypothetical valuations, interactionism allows that, although one gives a guarantee in the course of making the assertion that one's counterpart will value something in a certain way—namely the same way as one does oneself—one could be *wrong* about this. Interactionism

then goes on to employ these cases of error in the service of improved valuations and improved desires, preferences, and interests.

The most straightforward way in which one tests this hypothesis is to implement the course of action recommended and to see the results—in effect, to make S into S*. This gives one firsthand opportunities to observe the outcome of the recommendation and to experience the fulfillment of the desire, preference, or interest that one decided to act upon. It is worth stressing that the object or event in question may resolve into a series of objects or events, each of which presents an opportunity for new information and possible changes in one's overall valuation. Suppose, for example, a person decides that a particular career serving the cause of social justice is his/her best option and the right thing to do. Presumably, he/she decides to major in it in some educational setting, specifically, let's say, by registering for a course in the selected field. Even at this early step, one might learn that there is some poor fit between the object (i.e., the career) one valued and one's own desires, preferences, and interests, or one might decide to change those desires, preferences, and interests in light of this new information and commit oneself even further to the original plan (e.g., by getting a doctorate in the field rather than just a bachelor's degree). The more complex the object or event initially valued, the more opportunities there are for changes in one's plans. So it's not the case that one needs to take up costly commitments to get the information one needs—opportunities for observation may present themselves from the very first stages.

Beyond direct trial, one can also defer to the trials of others, either single trials of single individuals or the aggregated experiences of many agents. Let us consider each of these cases of deferment in turn. Suppose you know someone who is similar to you in the sense that they share similar life prospects and similar desires, preferences, or interests. They may attach value to some object or event and, upon having it brought about, make changes to their valuations. Given the similarity between this agent and yourself, it makes good sense to defer to him/her in relevant cases and to treat his/her empirical confirmations on par with your own. This process of deferment is, of course, used by natural scientists quite regularly. It is simply not possible to run every possible experiment one wants to run; there are constraints of resources, time, and even one's own abilities. Instead, one defers to other members of the scientific community who are known to be reliable sources. The standards of reliability in the case of natural science (most likely peer review[99]) are bound to be different than the standards used in moral science (most likely similarities between agents). But the basic strategy of deferment remains the same.

Another kind of deferment occurs on a much larger scale. Instead of taking the results of other individuals or even their single trials, one might take aggregated results from a variety of individuals, including even those who are

---

[99] One disanalogy with scientific confirmation is that, in principle, *any* scientist is properly situated to verify the results of another scientist, while in the moral case, not every individual is properly situated to verify the valuations of every other individual (or so I have been arguing). Strictly speaking, it should be noted that not *all* scientists are well positioned to evaluate the work of *all other* scientists; disciplinary boundaries and intense specialization may restrict the pool. Moreover, it is often the case that experimental measurements vary ever so slightly, even when the same apparatus is employed. The scientific community is willing to accept a certain margin of "error" ("difference" might be more appropriate) and I see no reason we cannot say the same thing of the moral community, recognizing that the scale of the difference may be larger.

dissimilar from oneself and, importantly, from each other in the sense discussed above. Presumably, the operating *assumption* here is that if a robust pattern emerges out of dissimilar individuals, that pattern can be trusted to apply to oneself as well. This strategy mirrors generalizability in the case of scientific experiments, particularly social scientific experiments. Provided that the sample pool does not exhibit some selection bias (i.e., a relevant difference from the cases to which we want to apply it), we can safely assume that the results apply to other cases within some acceptable margin of error.

This assumption may be particularly useful for incorporating historical information from other individuals or regarding large-scale policies or experiments. Given enough historical data, one may be able to conclude (provisionally) that the practice of slavery is unstable for a society over any great length of time or that democratic processes lead to the most long-lasting and stable governments. With centuries of human existence at our disposal, we have, in principle, a finite but very large set of observational data to contribute to our criteria for valuation. We must be cautious, here, in avoiding the presumption that just because people continue to value something (e.g., slavery) that that itself counts as the relevant observational evidence. The best cases for assimilation are ones in which agents have direct experience with the object or event in question (e.g., slaveholders or slaves). Mere guessing at what the outcomes might be like is no better than guessing what the results of a laboratory experiment might be. Moreover, we should be cautious in looking for patterns of robust disagreement between groups (e.g., slaves and slaveholders) that can be constructed out of features otherwise

irrelevant to the object or event in question (e.g., race, class, gender, sexual orientation, disability, nationality). Such cases might point to structural inequality or injustice in human experience. Instead, we should attend to cases of wide agreement among subjects who are dissimilar from each other.

Beyond trial and deferment, we can also incorporate theoretical knowledge provided by the sciences and public policy. Data on evolutionary adaptations, biological needs, social universals, psychological dispositions, economic habits, and so forth can all help to predict how our counterparts would respond to the outcome of certain objects and events. These data are empirical, through and through, and probably more reliable sources than others' testimony or even our own anecdotal experience. However, given the strict standards of confirmation in these fields, it is quite possible that they do not always offer results on the complicated and muddled options we face on a daily basis. Thus, theoretical knowledge alone cannot be the basis for predicting what our counterparts would actually find valuable. There may be particular individual differences that make the case at hand unique—an outlier to the otherwise observed tendency reported by these disciplines. Thus, as a practical matter, direct trial and indirect trial via testimony will be needed to fill out the full range of information we need to predict what our counterparts would value.

6.3. Social Individuals

So far, I have presented an overview of interactionism and discussed in greater detail the extent to which it mimics natural science. We have grounds for saying

that individuals impose value according to the judgments of their better-informed counterparts (who have the benefit of hindsight), and by iterating this critique on cases, we have a notion of refinement and moral progress in the assignment of value. The question I now want to take up is one crucial to the task of developing a normative account: the extent to which interactionism is pure egoism. In developing interactionism so far, I claimed that agents pursue various ends relative to their own desires, preferences, and interests. From this formulation, it does not follow that, even in cases of improvement, agents are necessarily doing anything more than prudential reasoning about how to best achieve their own aims and ends. This selfish or *potentially* egoistic picture may, after all, be the best we can do, morally speaking. But I disagree, and I suspect that many ethicists would be tempted to reject this metaethical framework if it did not entail that correctness involved more self-sacrificing and altruistic behavior, at least on some occasions. What I want to consider in this section, then, is the degree to which individual desires, preferences, and interests can take into account the welfare of others on an interactionist account.

To defend the notion that correct moral judgments (according to interactionism) will sometimes require self-sacrifice and advantage-reducing policies,[100] I advance forward three empirical hypotheses about human nature—or rather, the conditions in which human agents find themselves—that show other-regarding concerns are actually built in, as it were, to one's own desires,

---

[100] See Wilson (2004).

preferences, and interests, at least frequently enough to meet the minimal requirements for normative theory to get off the ground. The first hypothesis, discussed in §6.3.1, concerns prosocial tendencies that appear to have developed during the early adaptive environment. The other two figure prominently in the history of ethics, perhaps most notably in the work of Adam Smith. One is the notion of linked prosperity (§6.3.2)—the sense in which agents' well being is bound together—and the other involves social reputation (§6.3.3) and the role that helping others plays in one's own social standing. These hypotheses, as empirical hypotheses, require confirmation on empirical grounds, probably by evolutionary psychology, sociology, and social psychology, respectively. In each case, I try to note the limitations of each of these empirical phenomena—that is, the limit to which it shows that other-regarding concerns are indeed part of one's own desires, preferences, and interests. In the case of the last two hypotheses (linked prosperity and social reputation), I also argue that cases of failure or deviation are not sufficiently threatening in frequency to undermine the robust tendency of these phenomena to aid interactionism in meeting the minimum requirement. For what it's worth, the truth of any one of these hypotheses might be enough to achieve the minimal requirement.[101] If two or more of them are true (or if there are others I have not anticipated), all the better.

---

[101] I will not adjudicate the question of individual or joint sufficiency here—better to let the empirical work carry on. Once we have a more developed notion of each of these, we can decide whether it alone or in combination with another or others is sufficient.

6.3.1. Evolved Prosocial Tendencies

In 1871, Charles Darwin speculated that "any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man" (83). Contemporary theorists have often continued this line of inquiry, discussing altruism, sharing, cooperation, conflict resolution, and egalitarianism in higher primates.[102] Recent inquiries in evolutionary biology, anthropology, and evolutionary psychology all converge on the notion that limited prosocial tendencies have evolved in human beings. The results of this body of literature are best reflected in Christopher Boehm's account, according to which human nature contains "a very large dose of egoism, a hefty dose of nepotism, but at least a modest and socially significant dose of altruism" (2000, 214).

These evolutionary reflections give some philosophers hope for human morality. Steven Pinker thus exclaims: "The new sciences of human nature can help lead the way to a realistic, biologically informed humanism….They promise a naturalness in human relationships, encouraging us to treat people in terms of how they do feel rather than how some theory says they ought to feel. They offer a touchstone by which we can identify suffering and oppression wherever they occur….And they enhance the insights of artists and philosophers who have reflected on the human condition for millennia" (2002, xi). Pinker's enthusiasm may

---

[102] See Sober and Wilson (1998), Boehm (2000), and Flack and deWaal (2000).

be a bit premature. Evolved prosocial tendencies are limited at best, and there are two problems with their relevance for contemporary moral issues. First, it is not clear that these tendencies offer much guidance in situations that have only come about in the past several decades or even centuries. For millennia, our distant ancestors engaged in behaviors of resource use and reproduction, so it comes as little surprise that resource allocation, sexual strategies, and childrearing may all exhibit evolutionary pressures. By contrast, behaviors of political participation, developing one's talents, and equal opportunity employment are newcomers on the evolutionary scene and usually tied to particular social and cultural structures. There has simply not been enough *time* for evolutionary pressures to exert much—if any—force on judgments about them.

Second, it is not clear that the tendencies are specific enough to guide actual decisions, even in cases where they do range over the situation, broadly speaking. Suppose a tendency toward egalitarianism disposes one to share resources equally among members of one's community and results in a judgment that each member should receive an equal share of every resource. What would happen if one possessed a tendency to produce different behaviors that fulfilled the negation of that judgment—or, worse yet, what if the same tendency produced such a contradiction? Suppose that, in addition to egalitarian tendencies, the organisms in mind have retributive tendencies that dispose them to punish wrongdoers for harmful and non-prosocial behaviors. These tendencies would dispose them to withhold resources from a wrongdoer, while their egalitarian tendencies would dispose them to distribute the resources to the wrongdoer as a

member of the community. In this case, one tendency would explain a certain moral behavior, while the other would explain its inhibition. The second tendency need not even be connected to other moral judgments, as the retribution mechanism probably is. Even a mechanism to invest more resources in one's own offspring—only sometimes regarded as altruism[103]—would be sufficient to generate a disposition against egalitarianism in certain cases and to get the objection off the ground.

The joint force of these two objections is the claim that evolved prosocial tendencies, even where they can be established empirically, do not *guarantee* that individuals will, as a matter of course, project values in ways that benefit others. Interactionism can accept this limitation in stride. The purpose of the three empirical hypotheses in this section is to make plausible the claim that *all* of an individual's valuations will not be purely self-regarding. But just as they are not all self-regarding, we need not show that they are all *other*-regarding. Clearly there will be a mix of the two kinds of valuations that depends, in part, on just *how* prosocial the individual in question is (i.e., how much his/her evolved prosocial tendencies are actualized). This will be determined by many factors including how developmental forces have impacted the expression of prosocial tendencies (including the frequency of their expression), the individual's genetic predisposition to various prosocial tendencies (as a result of mutation and the interplay of genes), the resources available to the agent, and cultural factors, such

---

[103] The meaning of 'altruism' is highly debated among evolutionary theorists. For a very brief overview, see Dawkins ([1982] 1999, 57).

as the degree of expression of different tendencies in different societies. The evolutionary hypothesis about inherited prosocial tendencies serves to disarm opponents who want to push the hard line that interactionism is a radical form of egoism. If the hypothesis is correct, we can naturally expect people to make some significant amount of valuations that benefit others, providing empirical evidence against the worry of egoism.

6.3.2. Linked Prosperity

Against this general backdrop of evolved prosocial tendencies, there are two additional empirical hypotheses that cast doubt on the fact that our valuation would be purely self-regarding. Though both hypotheses are present in the literature to some extent from antiquity onwards, they both find clear exposition in the work of Adam Smith. The first notion, linked prosperity, is a structural claim about societies and the mutual effects of certain aims.[104] The second notion, to be discussed in the next section, is a psychological claim about social relations and our motivation for action.

Linked prosperity occurs when the individual welfare of mutually connected members of some organizational unit hangs or falls together. The literature on

---

[104] The idea of linked prosperity predates Smith—one finds it in Plato's *Republic*, for example. But it is interesting that Smith accords it such a central place in his ethical theory. I think the prominence of linked prosperity there and his own interest in capitalism are no coincidence and cannot be understressed. Capitalism is one of the first practices that links people on a global scale. Travel, commerce, and war all do so on a smaller scale before the early modern period, but capitalism is the first large-scale experiment in global citizenry, as it were. Approaching the present, we find increasing examples of global connectedness (e.g., communist revolutions, world and cold wars, greater ease and frequency of travel, internet and telecommunications) and, correspondingly, increasing discussions of human rights, just war theory, environmental ethics, and so on. As the former increase the magnitude of linked prosperity, the latter, unsurprisingly, are developed to advocate policies that enhance that prosperity.

business ethics often appeals to linked prosperity in claiming that consumers and producers are engaged in reciprocal, cooperative practices in the course of business.[105] But the application of linked prosperity extends much further. Smith cites it at the level of entire societies: "No society can surely be flourishing and happy, of which the far greater part of the members are poor and miserable. It is but equity, besides, that they who feed, clothe, and lodge the whole body of people, should have such a share of the produce of their own labour as to be themselves tolerably well fed, clothed, and lodged" ([1776] 1986, 203). The end of this passage presents a normative principle regarding equity, a principle perhaps based in some unstated view about the rights of labor and property ownership. The first sentence, however, expresses the relevant empirical claim that the state of flourishing of the organization as a whole depends on the flourishing of all of its parts.

For Smith, it is not enough that aggregate flourishing be achieved by the greater flourishing of certain parts, perhaps at the expense of others. Smith denies that society *as a whole* can be happy if some portion of it suffers. One might point out that Smith imagines "the far greater part of" society to be suffering—that is, one might suspect that society in this example is not flourishing because of the greater *numbers* of those that are suffering, not because of the amount of their individual suffering. If that were true, Smith should tolerate a situation in which the

---

[105]Freeman (1984) and Donaldson and Preson (1995) give the classic formulation of stakeholder theory according to which managers have responsibilities to all stakeholders (i.e., all parties that can affect or be affected by the practice of business). Luetge (2005) presents a fresh account of stakeholder theory in discussing the "mutual advantage" of ending child labor.

greater happiness of a small subset of society outweighs the lesser misery of the masses. But his own normative prescription contradicts this reading, so we can interpret Smith as rejecting this crude form of consequentialism. Smith's insight may be recast as, "No society can be flourishing and happy, of which *any significant part* of the members are poor and miserable." The assumption behind this claim must be that the outcomes of actions for some are also the outcomes of those actions for others. Some may be doing well—perhaps well enough to make the average well being of the society quite high—but if their well-being comes at the expense of others or if a significant number of others are simply suffering (for whatever reason), then the society as a whole is not doing well. This again reflects linked prosperity; when I recommend a certain course of action for its intended aims, that action may also impact others. In a reverse of Kant's dictum, we might say, "He who wills the means necessarily wills the ends."

So what, it might be objected, an individual can simply discount the scope of the outcomes when it comes to others and continue to assign values strictly in terms of the outcomes for himself/herself. Strictly speaking, this is correct, but one *can* do a lot of things: one can not engage in the process of imposing values at all,[106] or one can impose values that have outcomes that are completely *other*-regarding and pay no heed to the individual's own well being. It is not the burden

---

[106] I doubt this is possible, any more than suspension of beliefs is possible. In life, as Hume says, "When he [the skeptic] awakes from his dream, he will be the first to join in the laugh against himself, and to confess, that all his objections are mere amusement, and can have no other tendency than to show the whimsical condition of mankind, who must act and reason and believe; though they are not able, by their most diligent enquiry, to satisfy themselves concerning the foundation of these operations, or to remove the objections, which may be raised against them" ([1748] 1999, XII.2).

of a metaethical theory to show that there is a reason why individuals *must* be moral (in a rational or normative sense of 'must'). If the hypotheses of the previous and following sections are correct, we naturally *do* have motives to take others into our process of valuation. The most we should expect from a metaethical framework is a confirmation procedure for normative proposals—an account of moral mental states that allows us to assess how successfully a normative theory or normative judgment meets the mind-to-world and world-to-mind satisfaction conditions for those states.

Linked prosperity guarantees that, in a significant number of cases, one's own actions have effects on others, particularly those to whom one is bound as a family or community member, citizen, human, or perhaps even fellow sentient being—and vice versa. Once one recognizes the mutual influences that people have on each other, one will naturally revise certain valuations because they have detrimental effects on certain subjects one antecedently cares about. Moreover, one cannot, in sincerity, consistently recommend *to* others that they not take into account these unintended side-effects of actions, provided one also expects that others take the effects of *their* actions on oneself into account.[107] Both of these entailments show that there are many cases in which we would correct our valuations because we *do*, as a matter of fact, care about their effects on others, if we simply had knowledge of linked prosperity. Thus, interactionism is not *hopelessly* wed to egoism. One is not bound to take others into account, either, but one never

---

[107] This minimal requirement for moral prescriptions is defended forcefully in Baier (1958).

could be in the first place according to interactionism unless *that* principle was somehow confirmed through the interactionist methodology.

6.3.3. Social Reputation

Beyond prosocial tendencies and linked prosperity, there is another, perhaps less admirable,[108] reason for thinking that valuations would take into account outcomes for others and not just oneself. Smith discusses the special role of social reputation in guiding moral judgments as follows:

> Man naturally desires, not only to be loved, but to be lovely; or to be that thing which is the natural and proper object of love. He naturally dreads, not only to be hated, but to be hateful; or to be that thing which is the natural and proper object of hatred. He desires, not only praise, but praise-worthiness; or to be that thing which, though it should be praised by nobody, is, however the natural and proper object of praise. He dreads, not only blame, but blame-worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the nature and proper object of blame….But in order to attain this satisfaction, we must become the impartial spectators of our own character and conduct. We must endeavour to view them with the eye of other people, or as other people are likely to view them. ([1759] 1986, 103)

Leaving behind Smith's own notion of the impartial spectator, we see here the suggestion that people act out of a sense of other's judgments about them—that

---

[108] This mechanism will not sit well with Kantians who leverage moral praise on agents' motivations. I offer critical remarks on this approach in §4.2.

is, they act out of a regard for their social reputation. This hypothesis was appreciated no less by Hobbes, who noted that "every man looketh that his companion should value him at the same rate he sets upon himself" ([1651] 1994, 75–6). Unfortunately, Hobbes drew the negative conclusion that reputation facilitates the state of war by causing even those with adequate material resources to quarrel with each other to obtain superabundant resources or otherwise enhance their social standing. But Hobbes' conclusion follows only if we regard fear or awe as the content of others' judgments of one's own reputation. If we allow, as Smith does, that observers make judgments about whether one is lovely or praiseworthy, hateful or blameworthy, then we avoid the conclusion that social reputation only enters into moral theory as a prop for discord. Even if people make *all* of these judgments about one's character and conduct—even if Hobbes and Smith are both right about the content of judgments of reputation—the negative conclusion does not follow. Given the ways in which fear and awe will prevent us from being seen as lovely or praiseworthy, we will need to decide, on balance, what kind of reputation we want to have.

The notion of social reputation follows as a corollary of linked prosperity. The objects and events we recommend have outcomes for ourselves and others. Even if *we* don't care how those outcomes affect others, *they* will still make valuations of our character and conduct, and the empirical hypothesis of social reputation is the claim that we do, in fact, care about these judgments. As with linked prosperity, it is, of course, possible that someone could shrug off the valuations of others and dismiss their points of view. However, our linked nature in

the practice of society will entail that they may frustrate our aims, outright prevent us from achieving certain ends, or even punish us directly for our actions. This is why Epicurus (1993) was correct in arguing that mere pleasure and bodily pains were not the whole of happiness; social goods (such as friendship of the "fullest intimacy" discussed in §40), wisdom and justice (§6), and freedom from fear (§10) matter as well. The latter hinge on maintaining some good social reputation, and agents will not discount this in their valuations, or such is the hypothesis under examination.

This hypothesis faces two problems, each of which is represented by a recurring character in the moral literature. The first is the ascetic hermit who desires nothing from society, and, it will be said, could impose values that took no account of others. The second is Hobbes' "foole" who believes it is in his/her interest to act wrongly whenever he/she believes his/her misdeeds will go undiscovered. Let us address each of these characters in turn.

The ascetic hermit is of no concern to the social reputation hypothesis. If he has indeed separated himself from society and his own reputation has no effect on him, *ex hypothesi*, then it seems his actions have absolutely no outcomes for others. If, on the other hand, he only *believes* he has separated himself, then we are no longer dealing with a true hermit, and others will make judgments about his reputation, which will therefore expose him to the effects we have been considering. Let us consider the example of a self-purported hermit who produces a large amount of trash that he tosses down the hill outside of his window. If he has indeed achieved separation from others—if indeed he has unlinked himself, as

it were—then his trash has no effect on others and they are unlikely to make any serious judgments about his reputation. He is safe from social derision and its attendant backlash, and society is safe from his valuations and actions. If, on the other hand, his trash winds up in someone else's yard or a city's water supply, then he has not, in fact, untethered himself from society and he will be subject to scorn and its consequences regardless of his mere *belief* that he is a hermit or his *indifference* to the welfare or rights of others. So either the hermit is irrelevant or he is simply a foole.

The foole (sometimes referred to as the "free-rider") is a more interesting case. She violates social rules (or takes resources without contributing back to the general stock) because she believes her actions will go undetected. Either she is genuinely *foolish* (because she incorrectly believes she will go undetected) or she poses a significant problem (because her valuations and actions may have serious effects on others). As a matter of fact, fooles probably fall into the former category more often than the latter. Epicurus rather cleverly points out that the foole can never be at peace: "It is impossible for the man who secretly violates any article of the social compact to feel confident that he will remain undiscovered…" (§35). People who break their contracts will gain little honor and possibly great scorn for their actions. On Epicurus' view, it is in the individual's own larger interest to fulfill agreements even when those agreements benefit another more than himself/herself. Moreover, humans have developed increasingly sophisticated systems of surveillance to detect such cases and punish them accordingly. Gossip

plays a major role in disseminating information about others' reputations from the early adaptive environment right up to contemporary online social networks.[109]

Modern societies have implemented advanced and computerized systems for tracking people's identifying characteristics, incomes, tax payments, movements, and so forth. Even as early as the nineteenth century, the French anarchist Pierre-Joseph Proudhon derided,

> To be GOVERNED is to be watched, inspected, spied upon, directed, law-driven, numbered, regulated, enrolled, indoctrinated, preached at, controlled, checked, estimated, valued, censured, commanded, by creatures who have neither the right nor the wisdom nor the virtue to do so. To be GOVERNED is to be at every operation, at every transaction noted, registered, counted, taxed, stamped, measured, numbered, assessed, licensed, authorized, admonished, prevented, forbidden, reformed, corrected, punished. It is, under pretext of public utility, and in the name of the general interest, to be place under contribution, drilled, fleeced, exploited, monopolized, extorted from, squeezed, hoaxed, robbed; then, at the slightest resistance, the first word of complaint, to be repressed, fined, vilified, harassed, hunted down, abused, clubbed, disarmed, bound, choked, imprisoned, judged, condemned, shot, deported, sacrificed, sold, betrayed; and to crown all, mocked, ridiculed,

---

[109] See Note 34.

> derided, outraged, dishonored. That is government; that is its justice; that is its morality. ([1851] 1923, 293–4)

Proudhon's worries predate a computerized age in which records and their retrieval have aided in massive political actions, for better or worse: passports, national identification cards, international financial institutions, entitlement programs, national health systems, mass exterminations, anti-terrorism, genetic sequencing, etc. All of these activities and their prerequisite surveillance—indeed, we are tracked more than Hobbes' leviathan could have dared to imagine—make it extremely unlikely that a foole actually *will* go undetected.

The frequency with which individuals can avoid damaging their social reputations is as much an empirical matter as whether and to what extent social reputation is a robust phenomenon that guides people's valuations. If it and the other hypotheses are correct, then the case for taking others into account in one's own valuations is quite strong. We will have general, evolved tendencies that guide us in limited ways toward prosocial behaviors. Failing those, linked prosperity will remind us that the objects and events we value have effects on others, some of whom we antecedently care about. In any case, we will know that others valuate our own valuations in ways that potentially limit our own ability to pursue our desires, preferences, and interests. Indeed, those "self-regarding" entities are social, through and through, and at least some of them will turn out to be other-regarding.

6.4. Conclusion

In this chapter, I developed a new metaethical framework, interactionism, which stresses the continuity between the methods of ethics and the methods of science. On this approach, our moral mental states are understood as hypotheses about what moral mental states we would form were we to possess additional knowledge about the outcomes of the plans we recommend or forbid. Such hypotheses may be informed by our own cases of trial and error, by deferring to others (either individually or aggregated experience), or by appeal to theoretical knowledge. I argued that interactionism does not commit us to egoism, and I advanced three empirical hypotheses that claim we do, as a matter of fact, pursue plans of action that are other-regarding. These hypotheses need empirical confirmation, to be sure, but so do our own valuations on the interactionist account. Experiments, both scientific and valuational, must continue.

Chapter 7

Conclusion

> I do not come with timeless truths.
> My consciousness is not illuminated with ultimate radiances.
> Nevertheless, in complete composure, I think it would be good if certain things were said.
> These things I am going to say, not shout. For it is a long time since shouting has gone out of my life.
> So very long…
>
> —Frantz Fanon, *Black Skin, White Masks*

My reader may recall the 1990s television series, *The X-Files*, in which Special Agent Fox Mulder continuously searches for proof positive of the paranormal. A UFO poster in Mulder's office reads, "I want to believe."

I've always been struck by the parallel between Mulder's poster and the weird ontologies of many philosophers. They want to believe in things such as universals, the essential properties of continuant objects, *stance-independent moral facts and properties*—things that exist beyond the reach of our experience. As Special Agent Dana Scully once pointed out, "common sense alone will tell you that these legends, these unverified rumors are ridiculous." Mulder retorts, "But nonetheless, unverifiable, and therefore true in the sense that they're believed to be true."[110]

---

[110] *The X-Files*. Episode no. 103, first broadcast 7 December 1997 by FOX. Directed by Peter Markle and written by Vince Gilligan, John Shiban, and Frank Spotnitz.

This dissertation is an attempt to say exactly what that belief amounts to. I began by pointing out that investigations into the nature of moral discourse—the historical centerpiece of metaethics—are ultimately investigations into the nature of moral mental states: those beliefs, judgments, thoughts, etc. in virtue of which our moral discourse has its meaning. I argued in Chapter 2 that our working account of these states should remain neutral with respect to moral content and normative prescriptions, and I developed a functional account of moral mental states to meet those criteria. On this account, moral mental states are intentional-volitional hybrid states with both mind-to-world and world-to-mind satisfaction conditions. As intentional states, they exhibit the property of intentional inexistence, which allows them to be about things that do not exist. As volitional states, they are responsible for bringing about certain behaviors in the world (if sometimes only verbal ones) and that, in contrast to emotions, they must aim to be consistent with one's other volitional states, including one's preferences, desires, goals, etc. With this framework in place, I examined three methods of metaethics present in the literature: externalism, internalism, and prescriptivism.

In Chapter 3, I considered the case for externalism with reference to moral realism, understood as the view that the content of moral mental states and moral expressions are truth-conditional and that, in at least one case, that content is true with respect to stance-independent moral facts and properties in the actual world. I examined three arguments for moral realism, concluding that arguments from the surface grammar of moral discourse and embedded contexts and truth-preservation ultimately fail to establish the full case for moral realism. The most

common and compelling argument, which is based on the transparency of moral experience, fails to establish the existence of stance-independent moral facts and properties because intentional inexistence bars all arguments from the mere presence of a mental state that is about something to the existence of that thing. Setting these worries aside, I considered what moral realism would look like if it *could* be established, and I noted that a realist must appeal to moral facts and properties in other possible worlds to make sense of indirect moral judgments, which are not about occurrent states of affairs. In light of this apparatus, I cast doubt on the notion that moral realism could provide illuminating insights into normative claims and on the idea that stance-independent moral facts and properties in other possible worlds could motivate behavior in the actual world.

In Chapter 4, I turned attention to internalist accounts by considering constructivist views on which the content of moral mental states are true or false in virtue of the moral mental states of idealized observers in nearby possible ideal worlds. Constructivists claim that our own experience of the moral point of view reveals the moral mental states these idealized subjects would hold, but I argued that this impoverished view of evidence fails to secure claims about Kantian moral motive, that diversity of moral experience in the actual world casts doubt upon the reliability of this procedure, and that idealized subjects are too exotic to provide a useful account of moral behavior for humans. I also developed a supervenience objection to constructivism according to which regularity in the judgments of idealized subjects is explained by their response to certain features of the world that occasion those judgments. Accordingly, there is no case for saying that these

subjects *construct* moral facts and properties in the course of forming moral mental states rather than merely *detect* stance-independent moral facts and properties already there. I attempted a resuscitation of constructivism using phenomenological views but concluded that, even with phenomenal properties in play, constructivism might still lapse into externalism and that claims about moral motive could not be substantiated.

Finally, in Chapter 5, I considered a class of performativist views that includes emotivism, expressivism, and quasi-realism. I noted that most performativists have failed to give any robust account of moral attitudes sufficient for establishing mind-to-world satisfaction conditions for moral mental states. The only exception here was Gibbard's account, which suffers from the same problems as the constructivists' appeal to idealized subjects. I then examined three possible solutions on behalf of the performativist. According to the first, every moral mental state is understood to reflect two separate mental states: a factive mental state about states of affairs in the world, and a related moral mental state with prescriptive force. This view defers the problem of making sense of mind-to-world satisfaction conditions for moral mental states. The second solution involves developing a logic for purportedly non-truth-conditional moral attitudes, but I noted that truth in this context is a red herring; the real issue is what relation moral mental states bear to the world, and the logic of the attitudes simply postpones this question. Finally, I examined Blackburn's quasi-realist account and concluded that his appeal to stance-independent moral facts undermines the central project of performativism and ultimately reflects a species of moral realism proper.

Given the failure of these approaches, I developed a fourth alternative in Chapter 6. On my account, the practice of imposing values on the world is an attempt to bring about certain outcomes that we seek. We can learn how to do this better (or worse) by attending to cases in which we revise our valuations in light of additional evidence of the outcome of certain actions or the experience of having the ends we seek actually brought about. Accordingly, interactionism approximates moral truth in the limit of hypothetical valuations—what our better-informed counterparts *would* choose, given the opportunity. I argued that we are not limited in this process to mere trial and error; deferment to others (whether in individual cases or the aggregate experience of many) and theoretical knowledge both have an important part to play. I also argued that interactionism does not entail that self-serving actions will always be correct. Quite the contrary, I advanced three empirical hypotheses—evolved prosocial tendencies, linked prosperity, and social reputation—to show that we *will*, as a matter of course, often undertake actions that benefit others.

While interactionism may rule out any universal moral prescriptions or stance-independent moral facts and properties, it incorporates what I see as the strongest elements of competing frameworks: a broadly scientific approach (realism), stance-dependence (constructivism), and thorough attention to the active role of moral discourse and practice (performativism). There are those of us who don't want to believe in stance-independent moral facts and properties, who think that they should be banished from the desert of the real along with Quine's creatures of darkness. There are also those of us who want to take individual

preferences and differences seriously and who think that Kantianism, with its idealized agents and impartial duties, prevent us from doing so. Interactionism does justice to moral mental states without taking on this unwanted baggage.

I conclude by noting that R. M. Hare, one of the biggest critics of the moral realism of his day, was also one of the strongest defenders of individual freedom. Hare noted that if we could hold a person's preferences constant, we could say with analytic strength that certain prescriptions were morally right and others, morally wrong. Unfortunately (or happily, as the case may be), it *is* possible for people to change their preferences and thereby endorse prescriptions that would otherwise seem horrible. That, Hare thought, is the price we pay for our freedom.

Hare did not despair at the state of moral theory any more than Philippa Foot did after arguing that moral laws were mere hypothetical imperatives—albeit well enforced ones. These imperatives, lacking the binding strength of their categorical cousins, depended simply on people's desires, inclinations, and interests.

Hare rested comfortably with the fact that "men and the world being what they are, we can be very sure that hardly anybody is going to take it [a horrible prescription] with his eyes open" (1963, 111) Foot recalled the citizens of Leningrad, who were not paralyzed with fear during the German siege because their defenders had a *merely contingent* loyalty to their city. She took heart in a vision of "volunteers banded together to fight for liberty and justice and against inhumanity and oppression" (1972, 315). These two champions of freedom did not need stance-independent moral facts and properties or idealized subjects to ground their moral

theories; they needed only humans as they are and as we would like to see them—as we would like to see ourselves.

I want to believe in people's preferences, and I want to believe in theories that avoid positing mysterious moral entities. We can do both by attending to the nature of the imposition of value and the methods of science and by pursuing a method of metaethics that does not help itself to stance-independence or idealized subjects. To the extent that this account meets our criteria for a theory of moral mental states, it reflects a viable alternative to realism, constructivism, and expressivism. To the extent those others fail, interactionism is our *best* account and our *surest* hope for attaining success in the normative realm. There is no guarantee that such an end exists, much less that we will achieve it. We may try nonetheless, and our moral practice may help us to believe.

Bibliography

Aiken, Henry David. 1952. "The Authority of Moral Judgments." *Philosophy and Phenomenological Research* 12 (4): 513–25.

Alston, William P. 1968. "Moral Attitudes and Moral Judgments." *Noûs* 2 (1): 1–23.

Anscombe, G. E. M. 1958. "Modern Moral Philosophy." *Philosophy* 33: 1–19.

———. 1958. "On Brute Facts." *Analysis* 18 (3): 69–72.

———. [1959] 2000. *Intention*. Cambridge, Mass.: Harvard University Press.

Audi, Robert. 2005. *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton, N.J.: Princeton University Press.

Austin, J. L. [1961] 1979. *Philosophical Papers*. 3rd ed. New York: Oxford University Press.

———. [1962] 1974. *How to Do Things with Words*. 2nd ed. Cambridge, Mass.: Harvard University Press.

Ayer, A. J. [1946] 1952. *Language, Truth, and Logic*. 2nd ed. New York: Dover Publication.

Bagnoli, Carla. 2002. "Moral Constructivism: A Phenomenological Argument." *Topoi* 21: 125–38.

Baier, Kurt. 1954. "The Point of View of Morality." *Australasian Journal of Philosophy* 32 (2): 104–35.

———. 1958. *The Moral Point of View: A Rational Basis of Ethics*. Ithica, N.Y.: Cornell University Press.

Baldwin, Thomas. 2002. "Three Phases of Intuitionism." In *Ethical Intuitionism: Re-Evaluations*, edited by P. Stratton-Lake, 92–112. New York: Oxford University Press.

Blackburn, Simon. 1981. "Rule-Following and Moral Realism." In *Wittgenstein: To Follow a Rule*, edited by S. Holtzman and C. Leach, 163–87. London: Routledge & Kegan Paul.

———. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. New York: Oxford University Press.

———. 1988. "Attitudes and Contents." *Ethics* 98 (3): 501–17.

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

———. 1996. "Blackburn Reviews Dworkin." *Brown Electronic Article Review Service*, http://www.brown.edu/Departments/Philosophy/bears/9611blac.html.

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

———. 2006. "Antirealist Expressivism and Quasi-Realism." In *The Oxford Handbook of Ethical Theory*, edited by D. Copp, 146–62. New York: Oxford University Press.

Blair, Robert James. 1995. "A Cognitivite Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57: 1–29.

Boehm, Christopher. 2000. "Conflict and the Evoution of Social Contract." In *Evolutionary Origins of Morality*, edited by L. D. Katz, 79–101. Bowling Green, Ohio: Imprint Academic.

Boehm, Christopher. 2000. "Group Selection in the Upper Paleolithic." *Journal of Consciousness Studies* 7: 211–15.

Boyd, Richard N. 1988. "How to Be a Moral Realist." In *Essays on Moral Realism*, edited by G. Sayre-McCord, 181–228. Ithica, N.Y.: Cornell University Press.

Brentano, Franz. [1874] 2009. *Psychology from an Empirical Standpoint*. New York: Routledge.

———. [1889] 1902. *On the Origin of Our Knowledge of Right and Wrong*. Westminster: Archibale Constable & Co.

Brink, David O. 1997. "Moral Motivation." *Ethics* 108 (1): 4–32.

Butler, Joseph. [1726] 1827. *Fifteen Sermons Preached at Rolls Chapel*. Boston: Hilliard, Gray, Little, and Wilkins. http://anglicanhistory.org/butler/rolls/index.html.

Byrne, Alex. 2001. "Intentionalism Defended." *Philosophical Review* 110: 49–90.

Byron, Michael. 1998. "Satisficing and Optimality." *Ethics* 109: 67–93.

———, ed. 2004. *Satisficing and Optimizing: Moral Theorists on Practical Reason*. New York: Cambridge University Press.

Caruso, Eugene. 2010. "When the Future Feels Worse than the Past: A Temporal Inconsistency in Moral Judgment." *Journal of Experimental Psychology: General* 139 (4): 610–24.

Casebeer, William D., and Patricia S. Churchland. 2003. "The Neural Mechanisms of Moral Cognition: A Multiple-Aspect Approach to Moral Judgment and Decision-Making." *Biology and Philosophy* 18: 169–94.

Chisholm, Roderick M., and Wilfrid Sellars. 1957. "Intentionality and the Mental: Chisholm-Sellars Correspondence on Intentionality." In *Minnesota Studies in the Philosophy of Science*, edited by M. S. H. Feigl, and G. Maxwell, 521–539. Minneapolis: University of Minnesota Press.

Chrisman, Matthew. 2008. "Expressivism, Inferentialism, and Saving the Debate." *Philosophy and Phenomenological Research* 77 (2): 334–58.

Crowell, Steven Galt. 2002. "Kantianism and Phenomenology." In *Phenomenological Approaches to Moral Philosophy*, edited by John Drummond and Lester Embree, 47–67. Dordrecht: Kluwer.

Dancy, Jonathan. 1986. "Two Conceptions of Moral Realism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 60: 167–87.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex*. Penguin.

Davidson, Donald. 1970. "How Is Weakness of the Will Possible?" In *Moral Concepts*, edited by J. Feinberg, 21–42. New York: Oxford University Press.

———. 1977. "The Method of Truth in Metaphysics." In *Midwest Studies in Philosophy 2: Studies in the Philosophy of Language*, edited by P. A. French, T. E. Uehling, Jr. and H. K. Wettstein, 244–54. Morris, Minn.: University of Minnesota Press.

Dawkins, Richard. [1982] 1999. *The Extended Phenotype: The Long Reach of the Gene*. New York: Oxford University Press.

Dennett, Daniel C.. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery, Vt.: Bradford Books.

———. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

———. 1988. "Quining Qualia." In *Consciousness in Contemporary Science*, edited by A. J. Marcel and E. Bisiach, 42–75. New York: Oxford University Press.

Devitt, Michael. [1984] 1991. *Realism and Truth*. 2nd ed. Princeton, N.J.: Princeton University Press.

Dewey, John. 1903. "Logical Conditions of a Scientific Treatment of Morality." *Decennial Publications of the University of Chicago* 3: 115–39.

———. 1922. *Human Nature and Conduct*. New York: Henry Holt and Company.

———. 1929. *The Quest for Certainty*. New York: Minton, Blach & Company.

Dewey, John, and James H. Tufts. 1908. *Ethics*. New York: Henry Holt and Company.

Donaldson, Thomas, and Lee E. Preson. 1995. "The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications." *The Academy of Management Review* 20 (1): 65–91.

Dreier, James. 1993. "The Structure of Normative Theories." *Monist* 76: 22–40.

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

———. 1995. *Naturalizing the Mind*. Cambridge, Mass.: MIT Press.

Drummond, John J., and Lester Embree. 2002. *Phenomenological Approaches to Moral Philosophy: A Handbook*. Boston: Kluwer Academic Publishers.

Dunbar, Robert. 1996. *Grooming, Gossip, and the Evolution of Language*. Cambridge, Mass.: Harvard University Press.

Dworkin, Ronald. 1996. "Objectivity and Truth: You'd Better Believe It." *Philosophy and Public Affairs* 25 (2): 87–139.

Enoch, David. 2006. "Agency, Schmagency: Why Normativity Won't Come from What is Constitutive of Agency." *Philosophical Review* 15: 169–98.

Epicurus. 1993. *The Essential Epicurus: Letters, Principal Doctrines, Vatican Sayings, and Fragments*. Amherst, N.Y.: Prometheus Books.

Fanon, Frantz. [1952] 2008. *Black Skin, White Masks*. New York: Grove Press.

Feinberg, Joel. [1958] 2008. "Psychological Egoism." In *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, edited by J. Feinberg and R. Shafer-Landau, 520–32. Belmont, Calif.: Thomson Wadsworth.

Field, Hartry. 1978. “Mental Representation.” *Erkenntnis* 13: 9–61.

———. 2000. “A Prioricity as an Evaluative Notion.” In *New Essays on the A Priori*, edited by Paul Boghossian and Christopher Peacocke, 117–49. New York: Oxford University Press.

Findler, Richard S. 1997. “Kant's Phenomenological Ethics.” *Research in Phenomenology* 2 (1): 167–88.

Flack, Jessica C., and Frans B. M. deWaal. 2000. “‘Any Animal Whatever’: Darwinian Building Blocks of Morality in Monkeys and Apes.” In *Evolutionary Origins of Morality*, edited by L. D. Katz, 1–31. Bowling Green, Ohio: Imprint Academic.

Fodor, Jerry A. 1983. *The Modularity of Mind*. Cambridge, Mass.: MIT Press.

———. 1987. *Psychosemantics*. Cambridge, Mass.: MIT Press.

Foot, Philippa. 1958. “Moral Arguments.” *Mind* 67: 502–13.

———. 1959. “Moral Beliefs.” *Proceedings of the Aristotelian Society* 59: 83–104.

———. 1972. “Morality as a System of Hypothetical Imperatives.” *Philosophical Review* 81: 305–16.

———. 1974. “‘Is Morality a System of Hypothetical Imperatives?’ A Reply to Mr. Holmes.” *Analysis* 35: 53–56.

———. 1975. “A Reply to Professor Frankena.” *Philosophy* 50: 455–59.

Frankena, William K. 1966. “The Concept of Morality.” *The Journal of Philosophy* 63: 688–96.

———. 1974. “The Philosopher's Attack on Morality.” *Philosophy* 49: 345–56.

Freeman, R. Edward. 1984. *Strategic Management: A Stakeholder Approach*. New York: Cambridge University Press.

Geach, P. T. 1965. "Assertion." *The Philosophical Review* 74 (7): 449–65.

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

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

———. 2006. "Normative Properties." In *Metaethics after Moore*, edited by T. Horgan and M. Timmons, 319–337. New York: Oxford University Press.

Glassen, Peter. 1959. "The Cognitivity of Moral Judgments." *Mind* 68: 57–72.

———. 1963. "The Cognitivity of Moral Judgments: A Rejoinder to Miss Schuster." *Mind* 72: 137–40.

Gordon, Robert M. 1973. "Judgmental Emotions." *Analysis* 34 (2): 40–48.

———. 1987. *The Structure of Emotions: Investigations in Cognitive Philosophy*. New York: Cambridge University Press.

Grice, Paul. 1989. *Studies in the Way of Words*. Cambridge, Mass.: Harvard University Press.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–34.

Hare, R. M. 1949. "Imperative Sentences." *Mind* 58: 21–39.

———. 1952. *The Language of Morals*. New York: Oxford University Press.

———. 1963. *Freedom and Reason*. New York: Oxford University Press.

Harman, Gilbert. 1977. *The Nature of Morality*. New York: Oxford University Press.

———. 1990. "The Intrinsic Quality of Experience." In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, edited by J. E. Tomberlin, 31–52. Atascadero, Calif.: Ridgeview Publishing Company.

Haugeland, John 1981. "Semantic Engines: an Introduction to Mind Design." In *Mind Design, Philosophy, Psychology, Artificial Intelligence*, edited by J. Haugeland, 1–28. Cambridge, Mass.: MIT Press.

Heidegger, Martin. [1926] 1962. *Being and Time*. New York: Harper & Row.

Herodotus. [c. 440 B.C.E.] 1998. *The Histories*. New York: Oxford University Press.

Hess, Nicole H., and Edward H. Hagen. 2006. "Psychological Adaptations for Assessing Gossip Veracity." *Human Nature* 17 (3): 337–54.

Hill, Thomas E., Jr. 1989. "Kantian Constructivism in Ethics." *Ethics* 99: 752–70.

Hobbes, Thomas, ed. [1651] 1994. *Leviathan*. Edited by E. Curley. Indianapolis, Ind.: Hackett.

Holmes, Robert L. 1974. "Is Morality a System of Hypothetical Imperatives?" *Analysis* 34: 96–100.

———. 1976. "Philippa Foot on Hypothetical Imperatives." *Analysis* 36: 199–200.

Hooker, Brad, and Margaret Olivia Little. 2000. *Moral Particularism*. New York: Oxford University Press.

Horgan, Terry, and Mark Timmons. 2005. "Moral Phenomenology and Moral Theory." *Philosophical Issues* 15: 56–77.

———. 2006. "Cognitivist Expressivism." In *Metaethics After Moore*, edited by T. Horgan and M. Timmons, 255–98. New York: Oxford University Press.

———. 2008a. "Prolegomena to a Future Phenomenology of Morals." *Phenomenology and the Cognitive Sciences* 7 (1): 115–31.

———. 2008b. "What Does Moral Phenomenology Tell Us About Moral Objectivity?" *Social Philosophy & Policy* 25 (1): 267–300.

Horgan, Terry, and Uriah Kriegel. 2008. "Phenomenal Intentionality Meets the Extended Mind." *Monist* 91: 347–73.

Huemer, Michael. 2005. *Ethical Intuitionism*. New York: Palgrave Macmillan.

Hume, David. [1751] 1998. *An Enquiry concerning the Principles of Morals*. Edited by T. L. Beauchamp. New York: Oxford University Press.

———. [1748] 1999. *An Enquiry concerning Human Understanding*. Edited by T. L. Beauchamp. New York: Oxford University Press.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, Mass.: The MIT Press.

Kahneman, Daniel, and Jackie Snell. 1990. "Predicting Utility." In *Insights in Decision Making*, edited by R. M. Hogarth, 295–310. Chicago: University of Chicago Press.

Kahneman, Daniel, and Robert Sugden. 2005. "Experienced Utility as a Standard of Evaluation." *Environmental & Resource Economics* 32: 161–81.

Kant, Immanuel. [1784] 2009. "Idea for a Universal History with a Cosmopolitan Aim." In *Kant's "Idea for a Universal History with a Cosmopolitan Aim": A Critical Guide*, edited by A. O. Rorty and J. Schmidt, 9–23. New York: Cambridge University Press.

———. [1785] 2002. *Groundwork of the Metaphysics of Morals*. Ed. Mary Gregor. New York: Cambridge University Press.

———. [1795] 1891. Perpetual Peace: A Philosophical Essay. In *Kant's Principles of Politics, including his essay on Perpetual Peace. A Contribution to Political Science*, ed W. Hastie. Edinburgh: Clark. http://oll.libertyfund.org/title/358/56096.

Kelly, Daniel, Stephen Stich, Kevin J. Haley, Serena J. Eng, and Daniel M. T. Fessler. 2007. "Harm, Affect, and the Moral/Conventional Distinction." *Mind & Language* 22 (2): 117–31.

Kennett, Jeanette. 2006. "Do Psychopaths Really Threaten Moral Rationalism?" *Philosophical Explorations* 9 (1): 69–82.

Kirchin, Simon. 2003. "Ethical Phenomenology and Metaethics." *Ethical Theory and Moral Practice* 6: 241–64.

Kniffin, Kevin M., and David Sloan Wilson. 2005. "Utilities of Gossip Across Organizational Levels—Multilevel Selection, Free-Riders, and Teams." *Human Nature* 16 (3): 278–92.

Korsgaard, Christine M. 2003a. "Realism and Constructivism in Twentieth-Century Moral Philosophy." In *APA Centennial Supplement toThe Journal of Philosophical Research*, 99–122. Charlottesville, Va.: Philosophy Documentation Center.

———. 2003b. *The Sources of Normativity*. New York: Cambridge University Press.

Kriegel, Uriah. 2002. "Phenomenal Content." *Erkenntnis* 57 (2): 175–98.

———. 2007a. "Intentional Inexistence and Phenomenal Intentionality." *Philosophical Perspectives* 21 (1): 307–40.

———. 2007b. "Moral Phenomenology: Foundational Issues." *Phenomenology and the Cognitive Sciences* 7 (1): 1–19.

———. forthcoming. "The Phenomenal Intentionality Research Program." In *Phenomenal Intentionality*, edited by T. Horgan and U. Kriegel. New York: Oxford University Press.

Kripke, Saul. 1972. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

———. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, Mass.: Harvard University Press.

Lewis, David. [1976] 1983. "Survival and Identity." In *Philosophical Papers*. Original edition, In *The Identities of Persons*, ed. A. Rorty, 55–78. Berkeley, Calif.: University of California Press.

Lo, Ping-cheung 1981. "A Critical Reevaluation of the Alleged "Empty Formalism" of Kantian Ethics." *Ethics* 91: 181–201.

Loar, Brian. 2003. "Phenomenal Intentionality as the Basis of Mental Content." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by M. Hahn and B. Ramberg, 229–58. Cambridge, Mass.: MIT Press.

Louise, Jennie. 2004. "Relativity of Value and the Consequentialist Umbrella." *The Philosophical Quarterly* 54: 518–36.

Luetge, Christoph. 2005. "Economic Ethics, Business Ethics and the Idea of Mutual Advantages." *Business Ethics: A European Review* 14 (2): 108–18.

Lycan, William. 2008. "Phenomenal Intentionalities." *American Philosophical Quarterly* 45 (3): 233–52.

MacIntyre, Alasdair. [1984] 1986. *After Virtue*. 2nd ed. Notre Dame, Ind.: University of Notre Dame Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin.

Mandelbaum, Maurice. [1955] 1969. *The Phenomenology of Moral Experience*. 2nd ed. Baltimore, Md.: Johns Hopkins Press.

Mandeville, Bernard. [1732] 1988. *The Fable of the Bees or Private Vices, Publick Benefits*. Indianapolis, Ind.: Liberty Fund.

Marx, Karl. [1845] 1969. "Theses on Feuerbach." In *Marx/Engels Selected Works*, 13–15. Moscow: Progress Publishers.

McDowell, John. 1984. "Values as Secondary Qualities." In *Morality and Objectivity*, edited by T. Honderich, 110–29. London: Routledge and Kegan Paul.

McDowell, John, and I. G. McFetridge. 1978. "Are Moral Requirements Hyopthetical Imperatives." *Proceedings of the Aristotelian Society* Supplementary Volume: 13–42.

Mesoudi, Alex, Andrew Whiten, and Robert Dunbar. 2006. "A Bias for Social Information in Human Cultural Transmission." *British Journal of Psychology* 97: 405–23.

Meyers, Diana Tietjens. 2004. "Narrative and Moral Life." In *Setting the Moral Compass: Essays by Women Philosophers*, edited by C. Calhoun, 288–305. New York: Oxford University Press.

Milo, Ronald. 1995. "Contractarian Constructivism." *The Journal of Philosophy* 92 (4): 181–204.

Moore, G. E. [1903] 1993. *Principia Ethica*. 2nd ed. New York: Cambridge University Press.

Morrow, David R. 2009. "Of the Terrible Doubt of Moral Appearances: An Essay in Moral Epistemology." Philosophy, The Graduate Center of the City University of New York, New York.

Nagel, Thomas. 1979. *Moral Questions*. New York: Cambridge University Press.

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

Nichols, Shaun. 2002. "How Psychopaths Threaten Moral Realism, or Is It Irrational to be Amoral?" *Monist* 85: 285–304.

———. 2004. *Sentimental Rules: On the Natural Foundation of Moral Judgment*. New York: Oxford University Press.

Nucci, Larry P. 2001. *Education and the Moral Domain*. Cambridge: Cambridge University Press.

Nussbaum, Martha C. 1990. *Love's Knowledge: Essays on Philosophy and Literature*. New York: Oxford University Press.

———. 1999. *Sex & Social Justice*. New York: Oxford University Press.

O'Gorman, Rick, David Sloan Wilson, and Kennon M Sheldon. 2008. "For the Good of the Group? Exploring Group-Level Evolutionary Adaptations Using Multilevel Selection Theory." *Group Dynamics-Theory Research and Practice* 12 (1): 17–26.

O'Neill, Onora. 2003. "Constructivism Vs. Contractualism." *Ratio (New Series)* 16 (4): 319–31.

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

Pautz, Adam. 2008. "The Interdependence of Phenomenology and Intentionality." *Monist* 91 (2): 250–72.

Phillips, D. Z. 1977. "In Search of the Moral 'Must': Mrs Foot's Fugitive Thought." *The Philosophical Quarterly* 27: 140–57.

Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature*. New York: Viking.

Plato. [380 B.C.E.] 1997. "Protagoras." In *Plato: Complete Works*, edited by G. M. A. Grube, 746–90. Indianapolis, Ind.: Hackett.

Pritchard, H. A. 1912. "Does Moral Philosophy Rest on a Mistake?" *Mind* 21: 21–37.

Proundhon, Pierre-Joseph. [1851] 1923. *General Idea of the Revolution in the Nineteenth Century*. London: Freedom Press.

Pufendorf, Samuel von. [1703] 1990. "The Law of Nature and of Nations." In *Moral Philosophy: From Montaigne to Kant*, edited by J. B. Schneewind, 170–80. Cambridge, Mass.: Cambridge University Press.

Putnam, Hilary. 1973. "Meaning and Reference." *Journal of Philosophy* 70: 699–711.

———. 1975. "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131–93.

Quine, Willard van Orman. 1948. "On What There Is." *Review of Metaphysics* 2 (5): 21–38.

———. 1960. *Word & Object*. Cambridge, Mass.: The MIT Press.

———. 1969. *Ontological Relativity*. New York: Columbia University Press.

Railton, Peter. 1986. "Moral Realism." *Philosophical Review* 95: 163–207.

———. 1989. "Naturalism and Prescriptivity." *Social Philosophy & Policy* 7: 151–74.

Rawls, John. [1971] 1999. *A Theory of Justice*. 2nd ed. Cambridge, Mass.: Harvard University Press.

———. 1980. "Kantian Constructivism in Moral Theory." *Journal of Philosophy* 77 (9): 515–72.

———. 1993. *Political Liberalism*. New York: Columbia University Press.

———. 1999. *Collected Papers*. Edited by S. Freeman. Cambridge, Mass.: Harvard University Press.

Rey, Georges. 1998. "A Narrow Representationalist Account of Qualitative Experience." *Philosophical Perspectives* 12: 435–58.

Rosenthal, David M. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Ross, Steven. 2004. "Real, Modest Moral Realism." *Philosophical Forum* 35 (4): 411–21.

Salmon, Nathan. 1996. "Trans-World Identification and Stipulation." *Philosophical Studies* 84 (2/3): 202–23.

Sayre-McCord, Geoffery. 1988. "Introduction: The Many Moral Realisms." In *Essays on Moral Realism*, edited by G. Sayre-McCord, 1–23. Ithaca, N.Y.: Cornell University Press.

Scheuring, Istvan. 2010. "Coevolution of Honest Signaling and Cooperative Norms by Cultural Group Selection." *Biosystems* 101 (2): 79–87.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24.

———. 1983. *Intentionality: An Essay in the Philosophy of Mind*. New York: Cambridge University Press.

———. 1992. *The Rediscovery of the Mind*. Cambridge, Mass.: MIT Press.

Sellars, Wilfrid. 1954. "Some Reflections on Language Games." *Philosophy of Science* 21 (3): 204–28.

———. 1962. "Philosophy and the Scientific Image of Man." In *Frontiers of Science and Philosophy*, edited by R. Colodny. Pittsburgh: University of Pittsburgh Press.

———. 1967. *Science and Metaphysics: Variations on Kantian Themes*. Atascadero, Calif.: Ridgeview Publishing Company.

———. 1973. "Actions and Events." *Noûs* 7: 179–202.

———. 1974. "Meaning as Functional Classification." *Synthese* 27: 417–37.

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

Shue, Henry. 1978. "Torture." *Philosophy and Public Affairs* 7 (2): 124–43.

Sidgwick, Henry. [1907] 1981. *The Method of Ethics*. Indianapolis, Ind.: Hackett Publishing Company.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9: 331–52.

Smetana, Judi. 1993. "Understanding of Social Rules." In *The Development of Social Cognition: The Child as Psychologist*, edited by M. Bennett. New York: Guilford Press.

Smith, Adam. [1759] 1986. "The Theory of Moral Sentiments." In *The Essential Adam Smith*, edited by R. L. Heilbroner, 57–147. New York: Norton.

———. [1776] 1986. "The Wealth of Nations." In *The Essential Adam Smith*, edited by R. L. Heilbroner, 149–320. New York: Norton.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Solomon, David. 2005. "Christian Bioethics, Secular Bioethics, and the Claim to Cultural Authority." *Christian Bioethics* 11 (3): 349–59.

Solomon, Robert C. 1973. "Emotions and Choice." *Review of Metaphysics* 27: 20–41.

Sommerfield, Ralf D., Hans-Juergen Krambeck, and Manfred Milinski. 2008. "Multiple Gossip Statements and Their Effect on Reputation and Trustworthiness." *Proceedings of the Royal Society B–Biological Sciences* 275: 2529–36.

Sommerfield, Ralf D., Hans-Juergen Krambeck, and Dirk Semmann. 2007. "Gossip as an Alternative for Direct Observation in Games of Indirect Reciprocity." *Proceedings of the National Academic of Sciences of the United States of America* 104 (44): 17435–40.

Stevenson, Charles L. 1937. "The Emotive Meaning of Ethical Terms." *Mind* 46: 14–31.

———. 1963. *Facts and Values: Studies in Ethical Analysis*. New Haven, Conn.: Yale University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* (127): 109–66.

———. 2010. "What is Constructivism in Ethics and Metaethics?" *Philosophy Compass* 5 (5): 363–84.

Sturgeon, Nicholas L. 1985. "Moral Explanations." In *Morality, Reason, and Truth*, edited by D. Copp and D. Zimmerman, 49–78. Totowa, N.J.: Rowman & Allanheld.

Thomson, Judith Jarvis. 2006. "The Legacy of *Principia*." In *Metaethics After Moore*, edited by T. Horgan and M. Timmons, 233–54. New York: Oxford University Press.

Tiffany, Evan. 2006. "How Kantian Must Kantian Constructivists Be?" *Inquiry* 49 (6): 524–46.

Timmons, Mark. 2003. "The Limits of Moral Constructivism." *Ratio (New Series)* 16 (4): 391–423.

Turiel, Elliot. 1979. "Distinct conceptual and developmental domains: Social convention and morality." In *Nebraska Symposium on Motivation 1977: Social*

Cognitive Development, edited by H. E. Howe and C. B. Keasey. Lincoln, Neb.: University of Nebraska Press.

Turiel, Elliot, Melanie Killen, and Charles C. Helwig. 1987. "Morality: Its Structure, Function, and Vagaries." In *The Emergence of Morality in Young Children*, edited by J. Kagan and S. Lamb, 155–244. Chicago: University of Chicago Press.

Turiel, Elliott. 1983. *The Development of Social Knowledge*. Cambridge: Cambridge University Press.

Tye, Michael. 2000. *Consciousness, Color, and Content*. Cambridge, Mass.: MIT Press.

———. 2002. "Representationalism and the Transparency of Experience." *Noûs* 36 (1): 137–51.

Velleman, David. 1992. "The Guise of the Good." *Noûs* 26 (1): 3–26.

Warnock, Mary. 1960. *Ethics Since 1900*. Oxford: Oxford University Press.

Welchman, Jennifer. 1995. *Dewey's Ethical Thought*. Ithica, N.Y.: Cornell University Press.

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers 1973–1980*. New York: Cambridge University Press.

———. 1985. *Ethics and the Limits of Philosophy*. Cambridge, Mass.: Harvard University Press.

———. 1995. *Making Sense of Humanity and Other Philosophical Papers 1982–1993*. Cambridge: Cambridge University Press.

Wilson, Catherine. 1983. "Literature and Knowledge." *Philosophy* 58: 489–96.

———. 2004. *Moral Animals: Ideals and Constraints in Moral Theory*. New York: Oxford University Press.

———. 2010. "Moral Progress Without Moral Realism." *Philosophical Papers* 39 (1): 97–116.

Wittgenstein, Ludwig. [1939] 1997. "Lecture on Ethics." In *Moral Discourse and Practice: Some Philosophical Approaches*, edited by S. Darwall, A. Gibbard and P. Railton, 65–70. New York: Oxford University Press.

———. [1953] 1999. *Philosophical Investigations*. 3rd ed. Upper Saddle River, N.J.: Prentice Hall.

Wolf, Susan. 1982. "Moral Saints." *Journal of Philosophy* 79 (8): 419–39.

Wright, Crispin. 1998. "Realism, Antirealism, Irrealism, and Quasi-Realism." *Midwest Studies in Philosophy* 12: 25–49.