Three Challenges to Moral Realism:
Evolution, Disagreement, and Moral Semantics

By

Justin Horn

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Philosophy)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: 7/10/14

This dissertation is approved by the following members of the Final Oral Committee:
Russ Shafer-Landau, Professor, Philosophy
Daniel Hausman, Professor, Philosophy
Robert Streiffer, Associate Professor, Philosophy
Elliott Sober, Professor, Philosophy
Paul Kelleher, Assistant Professor, Medical History and Bioethics

Contents

Acknowledgments

I am grateful to many people for helping to bring about this dissertation in its present form. The project was assisted by a Mellon/ACLS Dissertation Completion Fellowship, and my sincere thanks go to the folks at the American Council of Learned Societies for their generous support. Special thanks also go to Russ Shafer-Landau, Dan Hausman, Elliott Sober, Rob Streiffer, and John Mackay for providing helpful comments on previous drafts of some or all of these papers.

I'm grateful to my fellow graduate students at UW for many stimulating philosophical conversations, and would like to express my gratitude in particular to my closest metaethical co-conspirators: Daniel Crow, Jeff Behrends, Stewart Eskew, and Alex Hyun. Our many conversations and disagreements about moral realism have made this dissertation far better than it would have been otherwise. Thanks also to audiences at University of Colorado-Boulder, Grinnell College, St. Ambrose University, Colgate University, and Oklahoma State University for listening to previous versions of this material, and providing lively discussion.

My greatest academic debt, by an enormous margin, is to my advisor, Russ Shafer-Landau. Russ has worn many hats as my mentor: providing incisive feedback on many, many of my papers over the years; heartily encouraging me in moments of self-doubt; introducing me to brilliant people working in the field whom I otherwise might not have met; and, most importantly, serving as a model of clarity and rigor in philosophical inquiry and kindness and grace in social interaction. It is difficult to imagine a better advisor than Russ.

Thanks also to my parents, Jim and Nancy Horn, for helping me to become the person I am today, and for supporting me in my wacky dream of becoming a professional philosopher. I

wish that my mother were able to see me complete this degree—sadly, she passed away in November of 2012 after a long and courageous battle with cancer. But I carry her memory and influence with me every day, and know that any successes I achieve are largely made possible by the love and support she gave me during the first 28 years of my life.

Finally, my greatest thanks go to my wife, Anna Zeide, for uncountably many things. Throughout the long haul of graduate school, Anna has been—as I declared to our family and friends a little over five years ago—*my sounding board, my muse, and my sharpest interlocutor.* Among the many joys Anna has brought to my life, the greatest was born this past November: our daughter, Nancy Zeide-Horn. I want to thank both of them for filling my life with love and adventure, and for cheerfully putting up with the occasional pedantry that comes along with living with a philosopher.

Abstract

According to the philosophical position known as *moral realism*, morality is a robustly objective domain of fact about which many of us have justified beliefs. This dissertation consists of three papers, each of which presents an independent line of argument against this position.

In the first paper, I examine Sharon Street's "Darwinian Dilemma," which claims that realists can give no adequate account of the relation between the (supposed) objective moral truths and the evolutionary pressures that have influenced our moral judgments. I develop a general strategy for constructing a realist response that avoids both horns of Street's dilemma. Then, I argue that while such a response escapes the specific critique presented by Street, it fails to adequately rescue moral realism from the epistemological challenges raised by the (putative) fact of widespread evolutionary influence.

In the second paper, I consider whether widespread, intractable moral disagreement raises an additional epistemological challenge for moral realists. First, I isolate exactly what sort of disagreement would pose the most serious threat to justified beliefs about objective moral truths, and develop an account of such *fundamental* disagreements. Next, I examine several popular anti-realist arguments from disagreement, and argue that they fail to undermine the realist position. Finally, I develop a novel argument for the claim that moral disagreement of a particular sort would undermine our ability to attain justified beliefs about objective moral facts.

In the final paper, I once again explore the implications of widespread ethical disagreement, but this time through the lens of moral semantics. Realists hold that moral terms such as "good" and "right" refer to objective moral properties, and that different parties to serious moral disputes refer to the *same* properties as one another when they use these words. I

argue that we have excellent reason to doubt that co-reference obtains in cases of fundamental disagreement. The semantic challenge, if successful, undermines the realist's contention that there is a distinct *moral reality* that we are all attempting to accurately describe when we engage in moral thought and discourse.

Introduction

This dissertation is a work in metaethics. As a sub-discipline of philosophy, metaethics examines questions concerning the *nature* of morality, as opposed to its *content*. Those studying the content of morality are concerned to discover which actions are right or wrong, which character traits are moral virtues or vices, which states of affairs are morally valuable, and related matters. Metaethics, in contrast, is concerned with more fundamental questions about moral thought, talk, and practice. For example, metaethicists explore questions such as: Are there any moral facts? If so, are they (in some sense to be clarified) *objective*? How do we know what morality requires of us? Can we even have any moral knowledge in the first place? What does it *mean* to say that (for example) a particular act is morally wrong? Does such a claim purport to *describe* the act in question, or does it fulfill some other (non-descriptive) function, such as expressing a sentiment of disapproval? Metaethics deals with these, and other, questions about the metaphysics, epistemology, and semantics of morality.

After falling out of favor for much of the twentieth century, one position (or, perhaps more appropriately, a family of positions) in metaethics has made an impressive comeback, and now threatens to be the dominant view in the metaethics literature: *moral realism*.[1] This dissertation develops three independent lines of critique against moral realism, and aims to show that the challenges to the view are more significant than many authors have appreciated. Before diving into the project of critique, however, I must take a moment to precisely specify the target of my arguments. In this introduction, I'll explain exactly what I take moral realists to be committed to, and then proceed to briefly outline the plan of attack for the chapters to come.

[1] According to a recent survey, it is the dominant view among philosophers generally. See Chalmers and Bourget (2013)

What is Moral Realism?

Very roughly, moral realism is the view that there are objective moral truths, and when we engage in moral inquiry we aim to discover those truths. This rough characterization is inadequate to pick out moral realism with enough precision for our purposes, however, especially in light of the proliferation of increasingly subtle positions in metaethics over the last few decades. In this section, I will describe five theoretical commitments that jointly constitute the realist position.[2]

1) Cognitivism

Moral realists are *cognitivists* in the sense that they hold that moral judgments are beliefs, rather than some other mental state (such as desires). Beliefs are often said to have a "mind-to-world" direction of fit in the sense that they aim to *accurately represent* the state of the world.[3] Because the aim of beliefs is to accurately represent the world, beliefs are *truth-apt*. If I hold the belief that my shoes are under my bed, this belief represents the world as being a certain way: namely, such that my shoes are located under my bed. Such a belief is true if my shoes are in fact under the bed, and false otherwise.

Desires and related states (such as hopes and plans), in contrast, have a "world-to-mind" direction of fit: very roughly, they represent the world *as we'd like it to become*. Because desires

---

[2] I should note that some authors (for example, Sayre-McCord (1988)) prefer to use the term "moral realism" in a much more permissive way, such that any view that meets my first two conditions below counts as "realist". How we use our terms here is largely a matter of stipulation, but I do think that my thicker characterization fits better with contemporary philosophical practice in that a) the commitments I describe are accepted by nearly all prominent contemporary metaethicists who identify as realists and b) few, if any, of those who identify as moral anti-realists would be willing to accept all five.

[3] cf. Platts (1997), who draws on Anscombe (1957).

(unlike beliefs) do not aim to accurately represent the state of the world, desires are not truth-apt. It may be true or false that I have a certain desire, such as a desire to eat cake, but my desire to eat cake cannot be true or false.

So, the first commitment of moral realism is cognitivism, the view that moral judgments are beliefs and are therefore truth-apt. Until a few decades ago, the standard foil of cognitivism was known, fittingly enough, as *non-cognitivism.* Traditional non-cognitivists deny that moral judgments are beliefs, and thus deny that moral judgments are capable of being true or false. Such views correspondingly deny that our moral *statements* are truth-apt. According to traditional non-cognitivism, moral statements cannot be true or false because our moral talk simply isn't in the business of attempting to *describe* the moral features of things; rather, our moral statements are best seen as *expressing* our feelings or *prescribing* certain behavior to others.[4]

Today, the picture has grown a little bit more complicated. A number of heirs to the non-cognitivist tradition, now calling themselves *expressivists*, have sought to show that even if one holds that moral judgments "start theoretical life" as something other than beliefs, one may nonetheless "earn the right" to talk of moral belief and even moral truth.[5] One element of this strategy is to invoke a *minimalist* understanding of truth, according to which the claim that *p is true* amounts to nothing more than the claim that *p*.[6] Minimalists about truth thus hold that if one is willing to go on saying things like "torture is wrong" (as traditional non-cognitivists are willing to do), then one should also be willing to say things like "it is true that torture is wrong," for according to minimalism about truth this latter claim is simply equivalent to the first. But if

---

4 Cf. Ayer (1952), Stevenson (1937) and (1944), and Hare (1991).
5 See Blackburn (1984) p. 168, and Blackburn (1998) pp. 77-83
6 See Field (1986) and Horwich (1990).

one holds that it is true that torture is wrong, then it seems to follow that there are some moral truths. And if our moral judgments are capable of truth, it is natural to think of them as beliefs.

If expressivists are willing to talk of moral beliefs and moral truth, one might wonder what exactly sets their position apart from cognitivist views.[7] Contemporary expressivists aim to maintain a distinctive position for themselves by, in Simon Blackburn's words, "separating truth... from 'represents' and its allies."[8] While the expressivist may feel comfortable talking of moral beliefs and moral truth in a minimal sense, she denies that the constitutive aim of moral judgment is to *accurately represent* those truths. In light of this, we should supplement our initial characterization of cognitivism to include the claim that moral judgments are not merely beliefs in some minimal sense, but beliefs that are as robustly *representational* as our beliefs in other areas in which we tend to think of ourselves as describing reality, such as discourse about ordinary physical objects. So, the first commitment of realism should be stated thus:

> **COGNITIVISM:** Moral judgments are robustly *representational* beliefs: mental states that aim to accurately represent the state of the world, and are true or false solely in virtue of whether they succeed in doing so.

In general, when we attempt to form judgments that accurately represent reality, it is natural to think and talk of ourselves as attributing properties and relations to various entities (e.g., persons, objects, events, etc.). For ease of exposition, I will adopt this way of talking. So, I will treat cognitivists as holding that moral terms pick out moral properties, and our moral judgments attribute these properties to actions, people, events, and so on. Furthermore, I will say that when some action, policy, person, or state of affairs possesses a moral property, that this

---

[7] Indeed, some have suspected that expressivists lack a distinctive position at all. See, for example, Dworkin (1996), and Dreier (2004)

[8] Blackburn (1993), p. 185.

constitutes a *moral fact*. Finally, I will talk of the content of moral judgments in terms of *moral propositions,* and say that these moral propositions are true if and only if they correspond to the moral facts. At the same time, I want to respect the right of cognitivists to retain a general metaphysical nonchalance.[9] The fundamental claim of the cognitivist is that moral judgment and talk is just as much in the business of representing and describing reality as judgment and talk in other familiar domains. If, like some philosophers, one is uncomfortable with talk of properties, or facts, or propositions quite generally, one may translate what follows into the terms of one's favored ontology. The important point is that for the cognitivist, moral properties exist in the same sense that ordinary non-moral properties do, and our moral judgments represent and attribute these properties in the same sense that our non-moral judgments represent and attribute other properties.

2) Success theory

The second commitment of realism is easier to characterize succinctly. All cognitivists believe that moral judgments aim to represent moral reality, but some cognitivists deny that they ever succeed in doing so. According to *error theorists*, there is simply no moral reality to be accurately represented, and thus any judgment that attributes a moral property is bound to be untrue.[10] Realists, in contrast, hold that there is indeed a moral reality to be represented, and our moral beliefs sometimes succeed in accurately representing it. Thus, the second commitment of realism is:

[9] See Enoch (2011b), p. 5.
[10] See Mackie (1977) and Joyce (2001).

SUCCESS THEORY: Some (property-attributing) moral beliefs are true.[11]

3) Stance-independence

In addition to thinking that our moral beliefs sometimes accurately represent moral reality, realists characteristically believe that this moral reality has a certain character. Roughly stated, moral realists think that morality is objective. The notion of objectivity is unclear in various ways, however, so let me say a little more about the sort of objectivity I take realists to be committed to.

In general, the notion of objectivity is often explicated by invoking the notion of *mind-independence*. But since the moral status of an action plainly depends to some extent on the mental states of human beings—for example, the fact that an action would cause someone severe pain is clearly a morally relevant consideration—we must take some care in spelling out the relevant sort of mind-independence that the realist attributes to morality. The best characterization I know of puts the point in terms of *stance-independence.*[12] A domain is stance-independent if and only if the truths in that domain do not constitutively depend on the attitudes that any actual or hypothetical agent takes toward their content. A commitment to stance-independence contrasts realism with various versions of subjectivism, cultural relativism, and contractarian constructivism, according to which the moral truth is wholly determined by individual attitudes, social conventions, or the verdicts of hypothetical contractors in idealized circumstances, respectively. To hold the view that morality is stance-independent is to provide a

[11] The qualification about property-attributing beliefs is required to account for the fact that error theorists can admit the truth of certain beliefs that employ moral concepts, such as the belief that *John's act was not morally wrong*. The mention of property-attribution here is subject to the qualification above about metaphysical nonchalance.

[12] I borrow this term from Shafer-Landau (2005), p. 15, who reports that he got it from Ronald Milo.

definite answer to Euthyphro-style questions about morality: though it might sometimes be the case that an agent believes that something has a moral property because it *does* have that moral property, it is never the case that something has a moral property solely because some agent (even an idealized hypothetical agent) thinks that it does. So, the third commitment of moral realism is:

> **STANCE-INDEPENDENCE**: The truth of moral judgments does not constitutively depend on the attitude that any actual or hypothetical agents take toward their content.

4) Epistemic access

So far, I've said that moral realists hold that moral judgments aim to represent a stance-independent moral reality and sometimes succeed in this aim. This is sufficient, I think, to accurately characterize the metaphysical commitments of the view. But it is not sufficient to capture the complete set of commitments that moral realists standardly accept, or that (as I will now argue) realists *must* accept in order for their position to charitably capture the central intuitions that tend to motivate realism and make it a *prima facie* plausible position in the first place. In addition to the metaphysical commitments above, I shall also attribute to realists two further commitments, one epistemological and one semantic. Let's start with the epistemological commitment.

Perhaps the greatest attraction of moral realism is that it promises to provide an account of our moral practice that vindicates what Allan Gibbard has called the "objective pretensions" of morality. Moral discussion is a ubiquitous feature of human life, and bears certain characteristic features. Among the most striking of these features is the fact that, at least on the face of it,

ordinary moral discussion seems to presuppose that some moral views are true and some are false, that the moral truth is *the same for everyone*, and that this truth can be at least partially discovered via rational reflection and conversation. Moral realism has a very tidy explanation of these appearances, accommodating them easily without having to add epicycles or try to explain them away as resting on illusions.[13] On the realist view, our practices of moral inquiry and discussion, while perhaps not entirely perfect, make excellent sense.

The point I want to make here is that the realist account has the potential to vindicate ordinary moral practice in this way only if we are not entirely in the dark about moral matters. If it were to turn out that for some reason human beings were incapable of becoming justified in their moral beliefs, the realist would have to admit that our practices of moral inquiry would be in terrible shape after all, and a good deal of our moral thought would indeed rest on an illusion. Most realists concede that this would render the metaphysical commitments of realism, in the words of one realist, "a logically coherent position that contains about zero appeal."[14]

In light of this, I shall understand realists as taking on a relatively weak epistemological commitment, holding that we are not entirely in the dark about moral matters. The fourth commitment of realism is as follows:

> **EPISTEMIC ACCESS**: Our (property-attributing) moral beliefs are sometimes justified, and responsible moral reflection would lead to *many* of our moral beliefs being justified.

One might simply insist that "moral realism" denotes a metaphysical position, that moral epistemology is another matter entirely, and so imputing an epistemological commitment to

---

[13] For more on this sort of motivation for realism, see, for example, Brink (1989), Ch. 2.
[14] Shafer-Landau (2012) p. 1.

realists *qua* realists is illegitimate. I don't think anything too substantive hangs on this point about terminology. As far as I can tell, if radical moral skepticism is true, moral realism is an unmotivated and deeply unattractive position. What I'm interested in capturing here is the package of commitments that a) jointly constitute a *prima facie* plausible and attractive account of morality, and b) that captures those commitments that are held by (virtually) all metaethicists who identify as realists. If one insists on using "moral realism" to refer to only the metaphysical components of this package, one can simply interpret my arguments as a critique of *non-skeptical* moral realism. In any case, going forward, I will presuppose that the preceding weak claim about epistemic access is a commitment of the realist views I will be addressing.

5) Invariantism

The final commitment is less often discussed by realists, but I believe it is necessary to rule out views that meet the above four criteria, but intuitively do not make morality objective in the way that realists believe it to be. I suggest that in addition to the metaphysical and epistemological commitments discussed above, realists are also committed to a particular *semantic* claim regarding the reference of our moral terms.

Perhaps the easiest way to motivate this point is to consider a paradigmatic case of a moral dispute. Suppose that Jane and John are engaged in a dispute about the morality of the death penalty. Jane sincerely utters the sentence, "the death penalty is sometimes morally permissible," and John sincerely replies, "the death penalty is never morally permissible." We might helpfully contrast two different interpretations of such a dispute. According to a traditional non-cognitivist diagnosis, there need be no disagreement in belief between Jane and

John; the moral disagreement is fundamentally a clash of conative attitudes, and the statements in question are not truth-apt.[15] In contrast, realists would naturally give a different diagnosis of what is going on in such cases. Namely, we would expect a realist to say that Jane and John have a disagreement in moral belief; they have asserted incompatible propositions, and at most one of them can be correct.

There is a semantic assumption underlying this interpretation. The assumption is that, at least insofar as the speakers are being sincere and using their terms in standard and literal ways, Jane and John are referring to the same property as one another when they are using the term "morally permissible." Imagine that this turned out not to be the case. (Whether the position I'm about to describe is remotely *plausible* is not important at this stage; the point is merely a conceptual one.) Suppose it turns out that Jane is best interpreted as asserting the proposition that the death penalty is sometimes permitted by some standard *A*, and John is best interpreted as asserting that the death penalty is never permitted by some different standard *B*. If this is so, then Jane and John would not have a disagreement in belief at all, or at any rate, not a disagreement in belief about the morality of the death penalty.[16] And this would be very much contrary to the spirit of realism. For the realist holds that in paradigmatic moral disputes, there is a stance-independent fact of the matter that settles the issue. If speakers, using moral terms literally in standard ways, are in fact picking out different properties with their moral terms, there is no such resolution: both speakers might well be saying something true. Indeed, each speaker might well be saying something *stance-independently* true, and that he or she *justifiably believes* to be true. Despite this, it seems clear to me that any view that entails that parties to a

[15] cf. Stevenson (1963)

[16] Though they might still be correctly described as being involved in a genuine *dispute*. Perhaps Jane and John disagree about which standards our moral terms *ought* to pick out. If that is so, the disagreement would ultimately be a metalinguistic one. See Plunkett and Sundell (2013).

paradigmatic moral dispute might both be correct in their judgments should not be classified as a version of moral realism. Thus, I suggest that realists should be thought of as accepting a fifth and final commitment:

> **INVARIANTISM:** At least for 'thin' moral terms (such as "morally obligatory," "morally permissible" and "morally impermissible"), all competent speakers use their moral terms to pick out the same properties as one another when sincerely reporting their moral judgments.[17]

One might wonder whether the realist should really be saddled with this commitment. Why can't the realist simply admit that people may well use moral terms in many ways, but that her theory gives an account of the nature of the properties that *some* people use moral language to pick out? After all, anyone (or any community) could simply adopt a slang usage of moral terms and begin using them to refer to some different set of properties, and this would hardly be a refutation of realism. (One might note, for example, Michael Jackson's famous use of the term "bad" as a term of praise.)

The answer is that the realist seeks to give an account of *morality* and *moral judgment* generally, and so it is crucial that her theory have something to say about at least many of the cases that we'd pre-theoretically identify as paradigmatic instances of moral judgment and moral disagreement (for example, arguments about the morality of abortion, the death penalty, etc.). If the realist refuses to apply her analysis to these cases by denying that they are genuine cases of moral disagreement, then she is vulnerable to the objection that her theory merely changes the subject. In light of this, I think it's best to include a commitment to invariantism as part of the

[17] I do not insist that realists accept invariantism about 'thick' moral terms like "courageous" or "cruel". Holding that different communities use "courageous" to pick out different properties doesn't seem as flagrantly contrary to the spirit of realism in the same way as holding this about thin terms.

realist package.

The Plan

I've sketched five commitments that I take to jointly constitute the position of moral realism. The position has many attractions. It makes excellent sense of many of the surface features of our moral practice, including our talk of moral truth and moral progress, and our practices of moral deliberation and moral argument. Furthermore, it seems to have the potential to vindicate the idea that morality is something worth *taking seriously.*[18] Nonetheless, I will spend the remainder of this dissertation arguing that the position faces very serious difficulties. In particular, I will argue that given the realist's metaphysical commitments, it is extremely difficult for her to make good on the epistemological and semantic commitments. This is especially so, I will argue, when we reflect on the following phenomena in particular: the fact that our moral faculties were shaped by an evolutionary process of natural selection, and the fact that deep, intractable moral disagreement is ubiquitous both within and across cultures.

The body of this dissertation consists of three relatively free-standing chapters, each pressing a different line of objection against moral realism. Although I don't claim that any particular argument provides a decisive refutation of realism—such refutations being extremely rare in philosophy—I hope to show that the cumulative force of the three of them gives us strong *pro tanto* reason to reject realism and seriously explore rival metaethical positions. Here is a brief preview of the attractions to come.

1) *The Evolutionary Challenge*

[18] See Enoch (2011b)

Over the last few decades, a number of psychologists and philosophers have argued that many of our moral attitudes likely have evolutionary roots. While the details of such accounts remain speculative, the broad outlines are sufficiently plausible to warrant investigation into their philosophical implications. In the first chapter, I assess recent attempts to refute moral realism on evolutionary grounds. The most promising of these, in my estimation, is Sharon Street's "Darwinian Dilemma" for realist theories of value, which claims that realists can give no adequate account of the relation between the objective moral truths and the evolutionary pressures that have influenced our moral judgments.

I think that Street's particular version of the evolutionary challenge can be met, and I develop a general strategy for constructing a realist response that avoids both horns of Street's dilemma. Unfortunately for the realist, I then go on to argue that while such a response escapes the specific critique presented by Street, it fails to adequately rescue moral realism from the epistemological challenges raised by the (putative) fact of widespread evolutionary influence. I conclude that if evolutionary pressures have significantly affected the content of our moral judgments, then accepting the metaphysical commitments of moral realism ends up saddling us with the conclusion that all of our moral beliefs are unjustified.

2) The Epistemological Challenge from Disagreement

In the second chapter, I explore how to best understand the epistemological challenge raised by widespread moral disagreement. First, I isolate exactly what sort of disagreement would pose the most serious threat to the justification of our moral beliefs, and develop an account of such *fundamental* disagreements. Next, I examine two influential anti-realist arguments: the first

suggests that disagreement gives us good reason to believe there are no objective moral facts, while the second invokes disagreement to argue that we are required to suspend judgment about a very wide variety of moral questions. After diagnosing what I take to be fatal flaws in these arguments, I develop a novel argument for the claim that moral disagreement of a particular sort would undermine our ability to attain justified beliefs about stance-independent moral facts. The deepest problem raised by fundamental disagreement, I argue, is that it shows that even our best methods of moral inquiry are unreliable guides to any stance-independent moral truth. Once again, I argue that the metaphysical commitments of realism seem to leave us unable to make good on the claim of epistemic access.

3) The Semantic Challenge from Disagreement

The third chapter again explores the implications of widespread ethical disagreement, but this time through the lens of moral semantics. As I mentioned above, realists hold that moral terms such as "good" and "right" refer to stance-independent moral properties, and that different parties to moral debate (typically) refer to the *same* properties as one another when they use these words. In the third and final chapter, I argue that the extent and character of moral disagreement calls this into question. My argument takes the form of a dilemma for the realist. In giving an account of moral semantics, realists must either adopt an *internalist* account, according to which the reference of our moral terms is fixed by certain beliefs we have about the properties in question, or an *externalist* account, according to which the reference of our moral terms is fixed by facts external to the speaker. Moral disagreement raises a worry for realists who adopt internalist views, because the deep diversity of beliefs about the nature of moral

properties threatens to undermine the notion that there is any sufficient reference-fixing description that all competent speakers associate with moral terms. On the other hand, the most prominent externalist views are problematic for the reason that different natural properties seem to causally regulate the use of moral terms for different speakers and communities. The semantic challenge, if successful, undermines the realist's contention that there is a distinct *moral reality* that we are all attempting to accurately describe when we engage in moral thought and discourse.

By the end of the third chapter, I hope to have shown that moral realists face severe difficulties in defending their epistemological and semantic commitments in light of our evolutionary origins and the phenomenon of moral disagreement. Let us begin with the evolutionary challenge.

Chapter 1: The Evolutionary Challenge

1. Introduction

Human beings evolved from our primate ancestors by a process of natural selection.[19] The idea that this scientific fact threatens to undermine traditional ways of thinking about morality is not new. Recently, however, there has been a resurgence of philosophical interest in the idea that evolution might play a key role in answering metaethical questions. In particular, a number of authors have argued that an understanding of our evolutionary origins might give us compelling reason to reject moral realism.[20] The version of the challenge that I wish to examine centers on the following claim: given the way in which our moral faculties have been shaped by evolutionary processes, those faculties cannot provide us with justified beliefs about any stance-independent moral truths (at least once we are made aware of this evolutionary influence). Insofar as the combination of a realist moral metaphysics with a radically skeptical moral epistemology is deeply implausible, the evolutionary challenge would, if successful, give us strong *pro tanto* reason to abandon moral realism.

Evolutionary arguments in metaethics characteristically rely on empirical assumptions which are, admittedly, somewhat speculative. A complete assessment of evolutionary arguments against realism would require detailed examination of the evidence supporting various hypotheses concerning the evolution of our capacity for moral judgment. I shall not attempt this difficult task here. Rather, in this chapter I propose to examine the metaethical implications of one hypothesis that has been endorsed by a number of authors, both philosophers and scientists.

---

[19] In highlighting the importance of natural selection, I do not mean to deny the possible importance of other factors, such as genetic drift.

[20] See, for example, Kitcher (2005, 2011), Street (2006), Joyce (2007), Locke (2014).

The hypothesis is this: *evolutionary forces have had a significant influence on the content of our moral judgments*.[21]

Notice that this hypothesis is stronger than the plausible claim that our capacity to think morally is, in some general sense, the product of evolution. The hypothesis is committed to the stronger claim that evolution has "pushed" us in the direction of making certain moral judgments rather than others.[22] According to the hypothesis, just as the pressure of natural selection played a significant role in bringing it about that tigers have sharp claws (rather than dull ones), and zebras are speedy (rather than slow), so too did the pressure of natural selection play a significant role in bringing it about that human beings have a very strong tendency to regard certain things as morally valuable or disvaluable (rather than regarding radically different things as having such value or disvalue).[23]

Examples of evolutionarily favored moral attitudes might include the widespread positive moral regard enjoyed by activities such as caring for one's own children and reciprocating benefits provided by others, and the negative moral regard commonly held for defecting from agreements or casually harming one's kin. There are, of course, variations in the precise form that such "moral regard" takes in different individuals and cultures. But it is hard to deny that even these vague generalizations get at real patterns in moral judgment, patterns for which there

---

[21] For some empirical evidence supporting this hypothesis, see Hauser (2006), De Waal (1997) and (2006), and Joyce (2006). Street (2006) briefly argues for a very similar hypothesis.

[22] Just how powerful did this "push" have to be in order to count as "significant"? I have no precise answer to this question, but one requirement would have to do with the *centrality* of the relevant judgments to our web of moral judgments. If the influence of evolutionary pressure were found to be limited to a few judgments at the periphery of our moral commitments (say, a prohibition on incest or an instinctive discounting of the moral relevance of the interests of those we see as "outsiders"), this would not count as "significant" in the relative sense. So, one requirement for the influence to count as significant is that it would have to target some moral judgments that we regard as relatively central, such as those mentioned shortly in the main text.

[23] I use "things" here as a catch-all term, intended to include acts, properties of acts, states of affairs, character traits, and anything else that might be subject to moral evaluation.

must be some causal explanation.[24] The hypothesis in question holds that some such deep patterns in moral judgment are significantly due to the pressure of natural selection, such that had we evolved differently, we would make moral judgments with very different content.[25]

I hasten to note, however, that the hypothesis is not intended to serve as a complete explanation of why we make *all* of the particular moral judgments we do. The truth of the hypothesis is consistent with the fact that there are many other significant influences on the content of our moral judgments—influences that include rational reflection, as well as a variety of social, cultural, and historical factors. The claim is only that evolution has been one powerful causal influence on the content of our moral judgments.

To avoid confusion, I should take note a few of the consequences of the hypothesis. First, because the hypothesis is not intended to explain *all* moral phenomena, it is not threatened by the existence of certain moral phenomena for which the best explanation is not an evolutionary one —for example, the changes in moral attitudes towards women and racial minorities that have taken place in the U.S. over the last 150 years. The following analogy might be helpful here. We have excellent reason to believe that genetic factors have a significant influence on our personalities. This belief is in no way threatened by the existence of personality-related phenomena for which the best explanation invokes non-genetic factors, such as a case in which

---

[24] One might think that the relevant explanation is not causal but *conceptual* for the following reason: were the contents of our evaluative judgments to diverge too widely from these patterns, we would no longer be making *moral* judgments at all. I am skeptical of this line of argument, but will not argue against it here. (See Shafer-Landau 2012, pp.11-12, for a tentative defense of this sort of position.) Rather, I will simply note that one could accept such a claim and simply reformulate the hypothesis (much less elegantly) to read something like this: *evolutionary forces have had a significant influence on the content of the human evaluative judgments that play a central role in our social and practical lives, and are thus significantly responsible for the fact that many such judgments qualify as* moral judgments *at all.*

[25] Darwin himself found such a hypothesis attractive. In the *Descent of Man*, he speculated that had we evolved under the conditions of hive bees, "there can hardly be a doubt that our unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering." See Darwin (2004, p. 122).

identical twins have substantially different personalities. Such cases merely show that genetic factors are not the *only* causal factor contributing to personality development. Similarly, certain moral phenomena may have a non-evolutionary explanation, even if the influence of evolution on our moral beliefs is pervasive.

Furthermore, the claim that evolution has significantly influenced the content of our moral judgments does not entail that we will inevitably make all the moral judgments that would be adaptive. Nor does it entail that all of the judgments we do make will be adaptive. On the most likely model of the development of our moral capacities, evolution has endowed us with certain very general evaluative dispositions regarding harm, fairness, purity, and the like.[26] These dispositions in turn shape the content of our moral judgments significantly, such that if we had a different set of general evaluative dispositions, we would come to very different conclusions about a variety of moral matters. Nonetheless, these general evaluative dispositions do not wholly *determine* which moral judgments we make; rather, they can be channeled in a variety of ways by culture, learning, and rational reflection. Thus, the foregoing model is consistent with us making moral judgments that are in fact quite harmful to our fitness, such as the judgment that I am required to sacrifice my life for my country.

With these clarifications in place, I will henceforth assume for the sake of argument that the hypothesis is true. The aim of this chapter is to examine whether its truth would raise any serious challenge to moral realism. I begin by laying out Sharon Street's "Darwinian Dilemma" for realist theories of value, and argue that realists can avoid both horns of the dilemma by adopting a third position that Street fails to adequately address. I try to show that this position

[26] For evidence of this from primatology, see the work of Frans de Waal on the evolutionary "building blocks" of morality, especially de Waal (1997) and (2006). The work of psychologist Jonathan Haidt and others on "moral foundations" is also relevant. See Haidt & Joseph (2007) and Haidt (2012). See also Kitcher (2005, 2011) and Street (2006).

has significant advantages as it allows the realist to deny that evolutionary forces have distorted our moral judgments, and crucially does so without committing the realist to any controversial scientific claims. Nonetheless, I argue that this realist position suffers from a significant defect of its own: any attempt to establish such a view will rely on premises for which the recognition of evolutionary influence will have already defeated *prima facie* epistemic justification. In light of this defect, I argue that if our initial hypothesis is true, moral realists are saddled with the conclusion that all of our (property-attributing) moral beliefs are unjustified. Since this sort of radical skepticism about morality is implausible, I conclude that *if* evolutionary forces have strongly influenced the content of our moral judgments, then we have strong *pro tanto* reason to reject realism as a metaethical position.

2. Street's Darwinian Dilemma

Moral realists claim that our moral judgments aim to accurately represent stance-independent moral truths. In an influential paper, Sharon Street claims that the following question raises a dilemma for realists: *What is the relation between the stance-independent moral truths and the selective forces that have influenced the content of our moral beliefs?*[27] On the one hand, the realist could admit that there is *no* relation between the two, in the following sense: the forces of natural selection have "pushed us in evaluative directions that have nothing whatsoever to do with evaluative truth."[28] Alternatively, Street holds, the realist could assert that there *is* a relation between the two, namely, that an ability to grasp moral truths was adaptive to our ancestors, and was therefore selected for.[29] Street goes on to argue that both of these alternatives are

[27] Street (2006) Street's target is in fact evaluative realism generally, but I restrict my consideration here to moral realism.

[28] *Ibid.*, p. 121.

[29] As Street acknowledges, this is oversimplifying a good deal. Much of the selection that has plausibly shaped the

unattractive, and that we therefore have reason to abandon the moral realism that gives rise to the dilemma.

I agree that both of these responses raise serious difficulties for the realist, for reasons to be explained shortly. Street's contention that she has posed a *dilemma* for the realist is more problematic, however, because the above dichotomy of possible relationships between the selective forces and any stance-independent moral facts is not exhaustive. In fact, the most plausible realist responses to Street's question neither assert that evolutionary forces have pushed us in a direction completely independent of the moral truth, nor that an ability to "grasp" the moral truths was selected for among our ancestors. Rather, this response holds that evolutionary pressures have pushed us *towards* moral beliefs that are mostly true, but not *because* those beliefs are true. Street dismisses this type of response on the grounds that it would require a "fluke of luck" that would be "extremely unlikely".[30] As I'll explain, I believe such an account cannot be dismissed so easily, and merits our attention when developing an evolutionary argument against moral realism. Let us begin, though, by considering the two alternatives that Street discusses.

3. The Distortion Hypothesis

The first possible reply to Street's question is that there is no positive correlation whatsoever

---

content of our moral judgments took place on organisms that lacked a full-fledged capacity of moral judgment. Furthermore, it is not likely that moral judgments themselves are the sort of thing that are genetically heritable. Street's actual view is that there were strong selection pressures on what she calls *basic evaluative tendencies*, non-linguistically infused dispositions to see certain behaviors as "called for". Street claims that selection strongly influenced the content of our ancestors' basic evaluative tendencies, and these basic evaluative tendencies strongly influence the moral judgments we make. I don't believe that these details are crucial to my discussion, however, so in what follows I shall continue to simplify by talking of "selection for moral judgments."

[30] Street, p. 122.

between the stance-independent moral truths and those moral judgments that evolution has "pushed" us toward. Let us call this the *distortion* hypothesis.[31] On the face of it, the distortion hypothesis seems to represent a worst case scenario for the realist. If the distortion hypothesis is true, then evolutionary pressures are no more likely to have pushed us towards true moral beliefs than, say, were we to have based our moral beliefs on a random drawing from a hat containing all logically possible moral judgments. The hat-drawing method is obviously very unlikely to consistently yield true moral beliefs. But if the distortion hypothesis is true, one considerable influence on the content of our moral judgments is, in all relevant respects, exactly like hat-drawing. If evolutionary influences on our moral judgments are sufficiently deep, and we have no way of correcting this distorting influence, then many of our moral beliefs are very likely to be false.

According to the hypothesis under consideration, evolutionary influences on our moral beliefs are in fact quite deep. Although certainly other factors may influence our moral judgments, the hypothesis suggests that many of our central moral commitments are largely the result of evolutionary pressures.[32] Thus if this hypothesis is true, a realist seeking to embrace the distortion hypothesis without succumbing to skepticism must therefore argue that we possess the tools for weeding out and correcting even deep and widespread errors among such moral

---

[31] One might think the label inappropriate, thinking that *distortion* suggests a *negative* correlation, rather than simply a lack of positive one. Ultimately, it it doesn't matter much what we call this hypothesis, but it's worth mentioning why I think *distortion* is an apt characterization. On the realist picture, our moral judgments are attempts to faithfully represent objective moral reality. Analogously, we can think of a radio as "attempting" to faithfully represent the original broadcast that is being transmitted. Just as interference with a radio signal needn't *negatively* correlate with the original signal (in terms of, say, pitch or volume) in order to badly distort the original signal, extraneous influences on our moral judgments needn't *negatively* correlate with the moral truth in order to knock us badly off-track in our quest to form true moral beliefs. All that is required in either case is a lack of positive correlation between the "signal" (the reality being represented) and the "noise" (the extraneous influence).

[32] It is important to remember that the hypothesis doesn't require that this pressure take the form of direct selection for these judgments. See the qualifications in section 1.

judgments. One way to argue for this conclusion would be to postulate a special faculty of moral intuition. It is not clear that this would solve the problem, however. Unless the realist also holds that we can reliably distinguish the outputs of our special faculty of intuition from those moral judgments that have been conditioned by evolutionary forces, the most that a faculty of intuition will achieve is to mix some true beliefs in with those mostly false beliefs that we have due to evolutionary pressures. Without any tool to reliably separate the two, a large percentage of our moral beliefs will still be false.

Michael Huemer argues that we have a way of determining which of our moral judgments have been heavily influenced by evolutionary pressures, and which are safe from such influence. According to Huemer, we need only look to the content of a particular judgment to figure out whether it is likely to have been influenced by evolutionary pressure. This is so, Huemer claims, because "biological evolution would be expected to produce a bias toward favorable evaluations of things that promote one's own inclusive fitness; intuitions that do not imply favorable evaluations of things that promote one's own inclusive fitness are not candidates for being products of this particular bias."[33]

This method of sorting is problematic, however. First, since evolutionary change takes time and environments can change rapidly, natural selection can produce organisms with traits that no longer promote their own inclusive fitness. Consider, for example, human tastes in food. As it happens, most of us find ourselves attracted to sweet, salty, and fatty foods. In environments where food is plentiful, these preferences can be quite maladaptive, leading people to consume more calories than would be healthy and often leading to early death. From the point

[33] Huemer (2008), p. 381. See Shafer-Landau (2012), pp.5-8 for a similar argument. *Inclusive fitness* is a measure of evolutionary success that accounts not only for survival and reproduction, but for an organism's increasing the representation of it's genes in the gene pool by promoting the fitness of kin (with whom organisms share their genes).

of view of survival, it would probably be far better for those of us living today if we were to crave whole grains and leafy green vegetables, and to find large quantities of fatty red meat repulsive. Nonetheless, there is a straightforward and plausible evolutionary explanation of our cravings for fats and sugars: in the ancestral environment, in which food was scarce and starving was a very real threat, it was adaptive to be motivated to consume the most calorific foods available. We inherited from our ancestors a palate that helped them survive, but which often leads us today to heart problems and obesity. Contrary to Huemer, then, natural selection can produce traits that are, in our current environment, quite detrimental to our own inclusive fitness.

Furthermore, it is not plausible that evolution influenced the content of our moral judgments by directly selecting for some particular judgments over others. On the contrary, the influence of evolution was likely much more indirect: certain general evaluative tendencies were selected for over time, and these evaluative tendencies in turn strongly influence which moral judgments we end up making. Given this sort of influence, it wouldn't be surprising if such dispositions sometimes "misfired" to produce particular judgments that turn out not to be reproductively advantageous, especially given significant differences between the ancestral environment and our own. Consider just one (admittedly speculative) example. It clearly promotes one's own inclusive fitness to ensure that one's own offspring—and to a lesser extent the offspring of one's close relatives—survive to reproductive age. If, in the ancestral environment, most of the small children that one came across were closely related to oneself, a standing disposition to be gentle towards all small children and refrain from harming any of them would be highly adaptive, and could in principle be selected for. In our present environment, we often come across small children to whom we are not related. Perhaps, in some circumstances,

harming them would promote one's inclusive fitness. Nonetheless, there could very well be an evolutionary explanation for why we are strongly disposed to regard harming small children as forbidden, even when it would be advantageous to do so.

This last example is, of course, a just-so story, and I am not advancing it as a complete and accurate explanation of our attitudes towards small children. Nonetheless, it does not seem implausible, and Huemer offers no argument against the possibility that something like this in fact partially explains our protective attitudes towards small children. If this is so, then clearly Huemer's criterion for identifying which moral judgments have been shaped by evolutionary forces is inadequate. The mere fact that a judgment does not currently promote one's own inclusive fitness is consistent with the hypothesis that the judgment is substantially the product of natural selection. Furthermore, to the extent that evolutionary influence on our moral judgments took place largely at the level of very deep and general evaluative dispositions (for example, by inclining us to regard certain very general features such as harm, fairness, loyalty, and purity as having positive moral relevance), it will be very difficult to find substantive moral judgments that are plausibly entirely isolated from such dispositions, even when we consider judgments that happen to be detrimental to our fitness in the present environment.

Even if we cannot identify exactly which moral beliefs have been shaped by evolutionary forces, one might think that the widely-endorsed method of reflective equilibrium would allow us to correct for any potentially distorting effects that evolution has had on our moral judgments.[34] Perhaps by "testing" our judgments about moral principles against our judgments about particular cases and vice versa, while also seeking coherence between our moral judgments and background theoretical considerations, we can root out even deep moral errors generated by

[34] On reflective equilibrium, see Rawls (1971), Daniels (1979), DePaul (1993), and Kelly and McGrath (2010).

evolutionary pressure.

If the distortion hypothesis is true, however, it is doubtful that such a method will be of much help. To see why, it may be helpful to distinguish narrow from wide versions of reflective equilibrium. Narrow reflective equilibrium involves simply working back and forth between judgments about moral principles and judgments about particular cases, adjusting each in light of the other until an adequate degree of coherence in achieved. If the set of initial moral judgments with which we begin inquiry is sufficiently corrupt, however, such a process of mutual adjustment is unlikely to be promising as a way of arriving at stance-independent moral truth.[35] As Street points out, if the distortion hypothesis is correct, then this sort of reasoning will simply involve "assessing evaluative judgments that are mostly off the mark in terms of others that are mostly off the mark."[36]

On the other hand, we might insist that a method of *wide* reflective equilibrium will fare better. Wide reflective equilibrium takes into account not only judgments about cases and moral principles, but also relevant background theories, including theories about which influences on our moral judgments might render them suspect. But if the distortion hypothesis is true, the attempt to employ such views faces a dilemma: we can either take suspected evolutionary influence on a moral judgment as a reason to discount that judgment, or we can accept that evolutionarily influenced judgments are legitimate starting points for the inquiry. If we take the first route, then unless we have some way of knowing which moral beliefs have been influenced

---

[35] Here we may notice the importance of the assumption of stance-independence to the argument at hand. If moral truth were held to be stance-*dependent*—for example, as being constituted by the judgments we would settle on after fully applying the method of reflective equilibrium—no such challenge can arise. Indeed, it would be hard to see how the distortion hypothesis could even be coherently stated under such an understanding of moral truth. For if moral truth is simply a function of our considered judgments, then it is hard to see how evolutionary pressures could render those considered judgments "off track" in any meaningful sense. See Street (2006), Section 10.

[36] Street, p. 124.

by evolution (and I've argued above that we do not have this), we will end up discounting all of our initial moral judgments. In this case, we will simply have nothing to work with, and the result of wide reflective equilibrium will be the suspension of judgment about moral matters.[37] If, instead, we accept evolutionarily influenced moral judgments to play the role of our moral starting points, then it is hard to see how widening the reflective equilibrium will help us to escape from the problem that we're simply trying to render mostly false moral judgments coherent with other mostly false judgments. Thus it seems that if the distortion hypothesis is true, a large percentage of our moral beliefs are very likely to be false, even if we prune them so as to bring them into reflective equilibrium.

The distortion hypothesis is therefore an unattractive option for any realist who believes that some of us are reliable moral judges. For this reason, realists might be encouraged to notice the following peculiar feature of that hypothesis: it would be very difficult to establish that the hypothesis is actually correct. To establish that there is indeed *no* positive correlation between the moral truth and the evaluative judgments that were selected for, it seems that we would need to compile a rough list of some moral truths, and then compare the moral truths to those evaluative judgments favored by natural selection.[38] Only by having some information about the contents of each list could we provide evidence that these contents were not correlated. If we

---

[37] Or perhaps, ultimately, a rejection of the realist commitments that give rise to such an unhappy result. Indeed, this is precisely where I shall argue that we should end up. The arguments of this chapter can be viewed as a (partial) attempt to bring our metaethical views into reflective equilibrium.

[38] Elliott Sober makes a similar point in Sober (1994), p. 107. It's worth noting that there are *conceivable* cases in which we could know that an influence was distorting without knowing much in particular about the moral truth. For instance, if the influence was *entirely random*, we could be confident the influence was distorting provided that we were confident that moral truths were not distributed randomly. For any plausible picture of evolutionary influence, however, the point in the text holds. For evolutionary influence on the content of our moral judgments was certainly not random; rather, the influence of evolution has plausibly disposed us to regard a certain cluster of natural features (in particular, those bearing certain relations to our inclusive fitness) as possessing moral relevance. In order to determine whether such influence was distorting, we would need to know whether such features really do possess the moral relevance we attribute to them.

possessed the information required for this task, however, then clearly any skeptical argument would be very hard to get off the ground. After all, in such a situation, we would already have a rough list of at least some moral truths! Much like global skepticism, it seems that the distortion hypothesis is not one that can be coherently asserted with confidence.

Yet it remains a troubling possibility, for an obvious reason. If the distortion hypothesis is correct, very many of one's moral beliefs are likely to be false, since there is no correlation between those moral beliefs that natural selection has pushed us towards (and thus, many of the moral beliefs that humans tend to have) and the moral truth. It seems to be a very plausible epistemological principle that if one has undefeated reason to think that one's beliefs in a domain have a high probability of being false, one cannot be justified in holding those beliefs. Thus, if realists have reason to believe that there is even a fairly high *probability* that the distortion hypothesis is correct, realism faces a serious epistemological challenge. Of course, I have not yet given any reason to believe that there actually is a high probability that the distortion hypothesis is correct. At this point, the thing to notice is merely that the *distortion* hypothesis could, in principle, threaten to undermine the justification of our moral beliefs even if it cannot be firmly established as correct.

4. The Direct Tracking Hypothesis

Realists might want to forestall this skeptical possibility by arguing that the evolutionary influence on our moral beliefs has been largely benign. One way of doing so would be to grasp the second horn of Street's dilemma. According to this view, evolutionary pressures have pushed us towards the stance-independent moral truth, because "natural selection favored ancestors who

were able to grasp those truths."[39] Let us call this the *direct tracking* hypothesis. Such a hypothesis, if true, would not only save the realist from epistemological objections based on evolutionary grounds, but would in fact provide the realist with a powerful tool for defending our general moral reliability.

As Street argues, however, this hypothesis is unacceptable on scientific grounds. In particular, it is inferior to a competing hypothesis, which she calls the *adaptive link* account. According to the adaptive link account, "tendencies to make certain kinds of evaluative judgments rather than others contributed to our ancestors' reproductive success not because they constituted perceptions of independent evaluative truths, but rather because they forged adaptive links between our ancestor's circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous."[40]

The main problem with the direct tracking account is that the most promising explanations of the evolutionary influence on the content of our moral beliefs simply needn't make any reference to the existence of moral facts. Indeed, it's not clear how postulating such facts would contribute anything to such an explanation. In contrast, consider the best explanation of the origins of our capacity for detecting mid-sized physical objects. Any acceptable explanation of our perceptual abilities will invoke the fact that non-veridical perceptions of mid-sized physical objects (say, predator or prey) would tend to be detrimental to the fitness of an organism. If an organism tends to form beliefs to the effect that it is being chased by predators when this is not so, it will end up wasting a lot of valuable time and energy running and hiding. Still worse, if an organism tends *not* to form beliefs that it is being pursued

---

[39] Street (2006) p. 109.
[40] *Ibid.*, p. 127.

by a predator on those occasions when it is in fact being pursued, that organism's genes are likely to be swiftly removed from the gene pool. In short, when it comes to avoiding predators, the *truth* of one's perceptual beliefs is of paramount importance.

In contrast, it is not at all clear how the *truth* of one's moral judgments can play any analogous role in an evolutionary explanation of our moral abilities. Other things being equal, it seems it would be adaptive for an organism to believe that it ought to take care of its offspring, and maladaptive to believe that it ought to kill them. But the adaptiveness (or lack thereof) of these judgments would remain exactly the same if it were to turn out, quite surprisingly, that we have a fundamental moral obligation to kill our own offspring. In morality, the adaptiveness of a judgment does not seem to depend on its truth.[41] Thus, we should expect selection for moral judgments which form adaptive links between circumstances and behavior, regardless of whether such beliefs are true or false.

One should note the limitations of the preceding remarks. I have not argued (nor does Street argue) that the mere fact that stance-independent moral facts play no role in scientific explanations justifies eliminating them from our ontology.[42] The present claim is much more modest. Given that we can explain everything worth explaining *about the evolutionary influences on moral judgment* without postulating moral facts, considerations of parsimony give us a reason to prefer the *adaptive link* account to the *direct tracking* account. Thus, we may

---

[41] That is to say, such judgments are not adaptive *in virtue of* being true. It's worth noting that if moral standards are necessary (as most realists hold), it might well be the case that adaptiveness and truth are *necessarily correlated* when it comes to morality. (Indeed, I explore this possibility at length in Section 5, below.) If this is so, there is a weaker sense in which the adaptiveness of a judgment depends on it's truth. But this weaker sense is not what is at stake when assessing the direct tracking hypothesis. The (plausible) direct tracking account of our visual capacities claims that the ability to make (roughly) true visual judgments about mid-sized physical objects was adaptive *in virtue of* those judgments being true. Direct tracking accounts in morality make an analogous claim.

[42] For such an argument, see Harman (1977).

conclude that while the *direct tracking* account would save the realist from a skeptical conclusion, it is unacceptable on scientific grounds.

5. Indirect Tracking and Pre-Established Harmony Explanations

Street claims that there are no other alternatives for the realist besides the *distortion* hypothesis and the *direct tracking* hypothesis. As she puts the point, "the only way for realism *both* to accept that [human evaluative] attitudes have been deeply influenced by evolutionary causes *and* to avoid seeing these causes as distorting is for it to claim that these causes actually in some way *tracked* the alleged independent truths."[43] But there is an ambiguity in the notion of *tracking* here, which obscures a host of important realist responses. At some points, Street identifies the notion of *tracking* with "the view that selective pressures pushed us *toward* the acceptance of the independent evaluative truths."[44] In all of her negative arguments, however, Street is clearly arguing against what I have called the *direct tracking* account, according to which true moral judgments were selected for *because they were true*.[45]

There is another possibility, however. It may be that there is a strong correlation between the stance-independent moral truths and those moral judgments that were selected for, such that the moral judgments that were selected for *are* mostly true, but were not selected for *because they were true*. I say "mostly" because the realist needn't insist that evolutionary pressures have pushed us towards the truth in every case in order to surmount the epistemological challenge. As David Copp points out, the realist can resist a skeptical conclusion, provided that "our beliefs

---

[43] Street (2006), p. 135.
[44] Street (2006), p. 135.
[45] Again, it is important to bear in mind that I am slightly oversimplifying things for the purposes of ease of exposition. As discussed previously, what were likely to be directly subject to selection pressure were not moral judgments, but more basic evaluative tendencies in human ancestors.

tend to do *well enough* in tracking the moral truth that rational refection can in principle correct sufficiently for any distorting influence."[46] The position we must consider, then, is one which accepts the adaptive link account as an explanation of *why* certain moral judgments were selected, while still holding that the moral judgments selected are *close enough* to the truth. I think that this view, which I will call the *indirect tracking* hypothesis, is the most promising avenue for the realist.

How might one defend the hypothesis that true moral beliefs were not selected for *because they were true*, but that nonetheless evolutionary influences have pushed us towards mostly true moral beliefs? The most promising explanation appeals to the widely accepted principle that any moral facts that exist supervene on natural facts: natural facts *fix* the moral facts in the sense that, necessarily, any two states of affairs that are exactly alike in all natural respects must be exactly alike in all moral respects. According to the evolutionary hypothesis presently under consideration, evolutionary forces have pushed us towards the acceptance of moral beliefs that are appropriately related to certain natural facts (namely, facts about survival and reproduction). If the natural facts that our moral beliefs tend to track are systematically related to the moral facts, this opens the door for a "pre-established harmony" explanation of the correlation between the moral judgments that were selected for and the realist's stance-independent moral truths.[47]

Suppose, then, that the realist accepts that certain moral beliefs were selected for, as described by the adaptive link account. The realist might proceed to argue that these moral beliefs are (mostly) true, because the features that moral judgments were selected to track either

---

[46] Copp (2008), p. 194.

[47] I borrow the use of the term "pre-established harmony" in this context from David Enoch (2010), whose views I discuss below. See also Skarsaune (2011).

constitute or closely correlate with moral features. We can explore how such a strategy would work by considering a simple form of naturalistic realism: hedonistic utilitarianism.

The hedonistic utilitarian might admit that moral beliefs were selected not for their truth, but for their tendency to motivate individuals to behave in ways that increased reproductive success. But the utilitarian might then claim that the moral beliefs that evolution has conferred on us are for the most part reliable. The utilitarian needn't simply see this as a convenient coincidence, but could argue for it as follows. Pleasure is intrinsically good and pain is intrinsically bad. Given this, one can imagine why natural selection would, to a considerable degree, favor true moral beliefs rather than false ones. After all, pain is typically an indicator of bodily harm, so organisms that tend to view pain as bad would tend to survive longer than those who do not. Likewise, pleasure is often an indicator of bodily benefit (or in the case of sexual pleasure, of reproductive success), and therefore organisms that see pleasure as good would tend to have greater reproductive success than those that do not. Thus, while evolutionary forces may have led us astray in some cases (for instance, the widespread belief that we have only very weak obligations to distant strangers), it is no accident that it has given us mostly true moral beliefs.[48]

I use utilitarianism as an example, but it is important to note that this sort of explanatory strategy could in principle be used for a wide variety of normative theories. It needn't be limited to reductive accounts, or even to naturalist accounts. Any view that claims that the moral facts supervene on natural facts could in principle tell this sort of story, by first linking certain natural features of the world with moral features, and then arguing that it was (for the most part) adaptive for our ancestors to regard those natural features as good, even though the explanation of why this is adaptive makes no reference to the truth of their judgments.

---

[48] Skarsaune (2011) considers a similar argument, which I will discuss below.

The non-naturalist realist David Enoch adopts this sort of strategy in order to respond to Street's Darwinian Dilemma. According to Enoch, what I have called the *indirect tracking* hypothesis can be adequately supported if we merely accept that "survival or reproductive success (or whatever else evolution "aims" at) is at least somewhat good."[49] This claim is not intended as a reductive account of what goodness *is*; it is merely a rough and ready claim that in most circumstances, survival has value. Enoch argues that if survival has value, and viewing survival as valuable was selected for, then evolution might have left us with mostly true moral beliefs, even if the truth of these moral beliefs plays no role in the explanation of why they were selected for.

It is not clear that such a modest assumption is sufficient to explain the correlation Enoch aims to explain. The claim that survival is at least somewhat good is compatible with the claim, for instance, that the beauty of nature is of far greater value, and that we are all obligated to sacrifice our own survival in order to maximize natural beauty. Likewise, Enoch's normative claim is compatible with the view that while survival is good, this goodness is outweighed by the goodness of excruciating suffering. An indefinite number of logically possible, internally coherent ethical systems are compatible with the claim that survival is at least somewhat good, and many of these systems differ dramatically from our moral intuitions in far-reaching, systematic ways. Thus, even assuming that survival is somewhat good, the realist still needs an explanation of why *our* system of intuitive moral judgments (which incorporates this assumption) approximates the stance-independent moral truth while all other such internally coherent sets incorporating it do not.

In general, though, indirect tracking accounts seem attractive because they have the

[49] Enoch (2010), p. 430.

potential to provide an explanation of a correlation between those moral beliefs favored by natural selection and the stance-independent moral truth, and all this without giving up the scientifically preferable adaptive link hypothesis. Street seems to think that the truth of any such account would rely on a fluke or a coincidence, and for that reason may be dismissed. It is not clear what exactly the fluke is supposed to be, however. Terms like "fluke" and "coincidence" suggest contingency. If realism is true, any fundamental moral standards which exist presumably do so necessarily. Furthermore, it is doubtful that in any world close to ours, evolution "aims" at something dramatically different than it does in ours.[50] The *relata* of an indirect tracking account are thus a necessary truth (i.e., a principle linking the moral to the non-moral), and a principle which although not strictly necessary, is probably true in most worlds that are remotely similar to ours (e.g., "evolution favors organisms that favor survival and avoid pain"). If some *indirect tracking* account is true, it would be odd to think that the correlation between the moral truths and those moral judgments that were selected for is a mere "fluke" or "coincidence."

Nonetheless, such accounts suffer from a serious defect. When presented with a claim linking the moral to the non-moral, we are entitled to ask what evidence or justification is on offer for the claim. The realist answer, it seems, will typically rely on substantive normative ethical views. This was clearly the case in the utilitarian example above, as well as the case of David Enoch's more modest bridge principle. In a similar vein, Erik Wielenberg attempts to vindicate our moral judgments in the face of evolutionary influence by assuming the normative claim that there are "moral barriers" that surround all creatures with sufficient cognitive capacities. These cases do not seem to be exceptions to the rule. As realist David Brink writes, ‘‘determination of just which natural facts and properties constitute which moral facts and

[50] Cf. Enoch (2010), p. 434.

properties is a matter of substantive moral theory."[51] The problem is that invoking substantive normative ethical views at this point in the dialectic begs the question, since the reliability of these views is exactly what is at stake.

The question for the realist is whether evolutionary influences have left us with the capacity to form moral beliefs that (at least roughly) track a stance-independent moral truth. Supposing we have ruled out the direct truth tracking account, we are left with two options: the *indirect tracking* hypothesis or the *distortion* hypothesis. If the distortion hypothesis is correct, then most of our intuitive moral judgments are false. If the indirect tracking hypothesis is correct, then a large number of our intuitive moral judgments are true, at least enough such that rational reflection could (in principle) weed out the bad apples. The trouble for the realist is this: how do we figure out which of these two possibilities obtains?

If we are at all unsure, it simply will not do to invoke substantive normative judgments at this point. Consider the following analogy. Suppose you discover that you've been brainwashed by a cult leader, who has given you all sorts of supernatural beliefs, which are based on visions he experienced while taking a brand new Miracle Drug. Further suppose that you are genuinely unsure whether Miracle Drug visions are a reliable guide to the supernatural truth, and are trying to ascertain whether or not this is so. Clearly it would not do to "test" the beliefs that the leader formed when using Miracle Drug against your own convictions about the supernatural. After all, you *know* that your beliefs about the supernatural are the result of the cult leader's brainwashing, so of course his supernatural beliefs will pass this "test," whether Miracle Drug visions are reliable or not.[52]

---

[51] Brink (1989) p. 177-178.

[52] David Copp considers a similar point. Cf. Copp (2008) p. 197.

Analogous things could be said about the evolutionary influences on our moral beliefs. If we are trying to determine whether evolutionary forces have pushed us towards the moral truth (as the indirect tracking hypothesis says) or not (as the distortion hypothesis says), it will be of no use to "test" the moral beliefs that would be selected for against our intuitive moral judgments. For we *know* (or so we are supposing) that our moral judgments have been heavily shaped by evolutionary forces. For this reason, the moral beliefs that have been selected for would be very likely to pass this test, even if the distortion hypothesis were correct.

One might be tempted to deny that invoking normative beliefs to defend realism against evolutionary challenges begs the question. Erik Wielenberg argues for such a view by noting the dialectical situation:

> [Epistemological debunking arguments] are not aimed at showing that there are no moral truths. Rather, such arguments are aimed at showing that even if there are moral truths, human beings lack knowledge of such truths. In arguing against this conditional claim, it is not question-begging to assume the truth of its antecedent (that there are moral truths).[53]

I have no problem with Wielenberg's characterization of the aim of evolutionary challenges (though I've been focusing on justification rather than knowledge), and he is also surely correct that in order to argue against the relevant conditional, one must be permitted to assume the truth of the antecedent. Notice, however, that *indirect tracking* theorists (Wielenberg included) assume much more than this. The antecedent of the conditional is that *there are (stance-independent) moral truths*. But what the indirect tracking theorist assumes is that these stance-independent truths have a particular content. It is this further assumption that seems question-begging.

Consider another analogy. Suppose an atheist and a theist are talking about religious

[53] Wielenberg (2010) p. 447.

matters, and the atheist presents some argument for the following claim: *even if God exists and there is something that He wants us to do, surely we have no way of knowing what it is*. In seeking to rebut the argument, it seems perfectly legitimate for the theist to assume that God exists, and that there are things He wants humans to do. What is not legitimate is to make substantive assumptions about the content of God's desires; this would make the original argument, however powerful, far too easy to refute. For instance, if one could simply assume without argument that God wants us to obey all the commandments in the Bible, then from this one could infer that we have a reliable means of figuring out many of the things God wants us to do. But this assumption about the content of God's desires is exactly what is at stake, and thus cannot be assumed without begging the question.

Of course, there is a disanalogy between the theological case and the moral case. While any claims about God's desires are bound to be extremely controversial, the moral claims invoked by defenders of *indirect tracking* are typically ones that enjoy widespread assent among those who hold widely diverging metaethical views. Given that the acceptance of such first-order moral claims need not depend on any specific metaethical way of viewing them, one might think that invoking such claims at this point in the dialectic would beg the question only against error theorists or (some) expressivists, who deny that there are any true moral claims.

But this thought, too, would be mistaken. It is surely correct that the claim that (e.g.) survival is usually good does not presuppose any metaethical account. And it is true that such a claim is quite plausible, and unlikely to be denied by most theorists, either realist or antirealist. It is another matter, however, whether the realist is in a position to assert such a claim at this point in the dialectic. The argumentative strategy of this chapter provisionally *assumes* that

moral realism is true, and then proceeds to explore whether realists can give a satisfactory account of the relation between the stance-independent moral facts and the evolutionary influences that have shaped our moral beliefs. If the realist invokes a substantive normative view as a premise in an argument against the distortion hypothesis, one may reasonably ask whether the realist is justified in believing that the normative claim is true. And it seems that these judgments can be regarded as justified only if we are in a position to justifiably suppose that the distortion hypothesis is probably false. For this reason, such judgments cannot be used in an argument for the indirect tracking hypothesis over the distortion hypothesis without begging the question.

Is there some other way of establishing the indirect tracking hypothesis that does not rely on any first order normative views? David Copp thinks so. According to Copp, it is possible to establish the indirect tracking hypothesis on the basis of "second-order philosophical intuitions, including the idea that morality has the function of making society possible."[54]

The claim that morality has the function of making society possible is ambiguous, however. One reading of this claim is perfectly acceptable to the antirealist and does not support the indirect tracking hypothesis, while the other is plainly a normative ethical view. On the first reading, we could interpret Copp's claim as meaning (very roughly) that the moral practices of individuals and societies exist because they make society possible. This clearly coheres very well with a plausible evolutionary account of the origins of morality, but it says nothing about the truth conditions of moral claims. For this reason, this reading cannot support the indirect tracking hypothesis. The second reading, which seems to be what Copp intends, would entail that people *actually* morally ought to do what makes society possible. But interpreted in this

[54] Copp (2008), p. 203.

way, the claim is clearly normative, and so cannot be utilized at this point in the dialectic without begging the question.

There is a general point to be made here. If we accept some version of Hume's dictum that moral claims cannot be established by arguments which invoke no moral premise whatsoever, we must admit that any attempt to establish the indirect tracking hypothesis will rely on normative judgments. I've argued that relying on normative judgments to establish the truth of the indirect tracking hypothesis over the distortion hypothesis begs the question. Thus, any argument for favoring the indirect tracking hypothesis over the distortion hypothesis will be question-begging.

Given all this, the realist might like to find some defense that doesn't rely on taking a stand as to which hypothesis is correct. Knut Skarsaune has offered a realist reply to Street that does not require choosing one hypothesis over the other.[55] Skarsaune begins in a way similar to my imagined utilitarian indirect tracking theorist above, by asking us to consider the proposition that pleasure is usually good and pain usually bad (for the person who is experiencing these states). At this point Skarsaune and the indirect tracking theorist diverge, however. In my example above, the utilitarian asserts the truth of a substantive normative claim. I have argued that such an argumentative strategy is unpromising, as it begs the question against the distortion hypothesis. Skarsaune, by contrast, remains (for the purposes of the argument) officially neutral as to the truth of the normative claim in question. His strategy is to argue by cases. If the normative claim is true, he claims, then the indirect tracking account is tenable and evolutionary arguments raise no epistemological challenge. If the normative claim is false, however, then most of our moral judgments are likely off track. But since realists have independent reason to

[55] Skarsaune (2011).

accept the conditional that *if Skarsaune's normative claim is false, then most of our moral beliefs are wrong,* Skarsaune claims that evolutionary arguments raise no new challenge for realism.

Again, I am skeptical that such a modest normative premise, even if granted, is sufficient to establish the indirect tracking thesis. The mere claim that my pleasure is good for me and my pain is bad for me is compatible with an enormous number of widely divergent but internally coherent moral views. Establishing that evolution has landed on one of the members of this set that is tolerably close to the stance-independent moral truth would require considerably more argument. Further, such argument must employ other, substantive normative assumptions, reliance on which would be problematic in this context.

I think there is a deeper problem with Skarsaune's argumentative strategy, however. Recall the previous example in which one discovers that one has been brainwashed by a cult leader, and is genuinely unsure whether the beliefs with which one has been inculcated are reliable or not. Some philosopher *could* come along and find some central belief at the core of the cult's ideology, and point out that *if* this claim is true, then one's beliefs are reliable, and *if* it is false, then one is hopelessly off track. But this does nothing to assuage the epistemological worries that come with realizing that one has been brainwashed by a source whose reliability one is unsure of. When one realizes that one has been brainwashed, one comes to realize that one's best-confirmed explanation of how one came to hold the beliefs one holds in a domain is entirely consistent with those beliefs being dramatically off track. The epistemological worry, in both the brainwashing case and the moral case, is not that we learn some new conditional of the form *if core belief x is false, most of my beliefs in this domain are off track*. The worry, in both cases, is that the antecedent of such a conditional leaves the realm of idle speculation and becomes a

serious contender for truth. As I will argue in the next section, this can have far-reaching epistemological implications.

6. Moral Realism and Skepticism

I've argued thus far that arguments for the indirect tracking hypothesis are question-begging, while the distortion hypothesis cannot be coherently asserted confidently. One might be tempted to conclude that without any way of resolving which hypothesis is the correct one, the epistemological challenge to moral realism flounders. After all, given everything that I've said, perhaps the indirect tracking hypothesis is correct and our epistemological situation is pretty good. So nothing I've said can be thought to undermine realism.

Furthermore, one might press the following line of argument. Suppose one were to call into question the justification of our perceptual judgments by challenging us to show that they themselves were not distorted in some deep way. One natural reply to such a challenge is to point out that the most plausible account of our basic perceptual capacities will be (to a significant extent) a *direct tracking* account, according to which the ability of those capacities to yield *true* judgments was essential to their being selected for. And this reply does seem adequate to vindicate our perceptual capacities to some degree. But notice: we can only establish a *direct tracking* story about the evolutionary origins of our perceptual capacities by relying on those capacities from the outset. Without the input of sensory observations, scientific theorizing about the nature of evolutionary influence on our perceptual capacities could never get off the ground.[56] And yet we do think that we are justified in believing things on the basis of our senses. So it seems plausible to suppose that our perceptual judgments have some *pro tanto* justification from

[56] cf. Schafer (2010) and Vavova (forthcoming)

the very beginning.

The moral realist might insist that similar considerations allow us to justifiably accept the *indirect tracking* hypothesis over the *distortion* hypothesis in the case at hand. If we are willing to grant *pro tanto* justification to our perceptual judgments from the outset, there seems to be no reason not to allow that our moral judgments enjoy a similar degree of *pro tanto* justification at the outset of inquiry. Once we grant this, though, it seems that the realist can rely on her *pro tanto* justified moral beliefs to rule out the *distortion* hypothesis and find in favor of an *indirect tracking* view.[57]

This is an elegant line of argument, but I think it can be resisted. We should grant the first point: we should, at the outset of inquiry, regard our intuitive moral judgments as having some modest degree of *pro tanto* justification. The question we must ask is whether this justification is undercut by the time we face the question of whether to prefer the *indirect tracking* hypothesis to the *distortion* hypothesis. And it seems to me the answer is yes.

The first thing to notice is that in seeking an evolutionary vindication of our perceptual judgments, there is never a moment at which we have an explanation of the origins of our perceptual faculties that completely leaves open the question of whether they are reliable in tracking stance-independent facts about our surroundings. We begin with *pro tanto* justified perceptual judgments, we do a lot of scientific inquiry, and we wind up with additional reasons to trust our perceptual faculties: our best explanation of their emergence vindicates their (approximate) reliability. But the second thing to notice is that we can imagine things being different. And in such imagined cases, the wrong kind of genealogy of our perceptual judgments could undermine their justification, even while leaving it open whether or not such judgments

[57] This is the line of argument advanced by Schafer (2010).

were actually reliable.

Imagine you were to discover something shocking about your perceptual judgments: they are never caused by external physical objects. Further, imagine you discover this in a manner completely independent of your perceptual capacities—perhaps God directly imparts this knowledge to you. It turns out that all of your perceptions are constantly being caused directly by some supernatural creature. This supernatural creature is akin to Descartes' evil demon, with one important difference: we have no idea whether he is evil. (For some reason, God neglects to tell us this part.) Indeed, we have no indication of the being's intentions whatsoever. Call this creature the Demon of Unknown Intentions (DUI).

With this new and disturbing information about your perceptual capacities, you start to worry about your ability to reliably form beliefs about external physical objects. You reason as follows: on the one hand, it's consistent with your newfound knowledge that the DUI is benevolent, and only gives you veridical perceptual experiences. Perhaps you only have the experience of a tree when there is indeed an external physical tree in your vicinity. Perhaps, in fact, the demon is necessarily benevolent, and so couldn't possibly deceive you in any deep and undetectable way. On the other hand, it's also entirely consistent with your newfound knowledge that the DUI is entirely deceiving you. Perhaps, as far as physical reality goes, you are just a brain in a vat, or an eight-armed slimy creature, or perhaps there is no external physical world at all.

What would be reasonable to conclude if (somehow) you were to learn that, as a matter of fact, all of your perceptual experiences were caused by the DUI? You could hope for a kind of indirect tracking explanation. Perhaps you could find some central regularities in the world of

your experience and assume that these correspond to physical reality (perhaps relying on our *a priori* entitlement to trust our perceptual capacities to justify this claim), and then deduce that your initial perceptual beliefs were close enough to veridical to correct any distorting influence through reasoning. For example, you might note that the DUI has made your experiences such that Newtonian mechanics seems roughly true of macroscopic objects. Since, you insist, Newtonian mechanics *is* roughly true of such objects, the DUI has probably not led you *too* far astray.

But here this reply seems totally unconvincing. Once you learn that your perceptual judgments are caused by something wholly distinct from any physical objects they seem to report, something which you have no independent reason to regard as a reliable source, the initial *pro tanto* justification provided by your perceptual judgments is defeated. Absent any other way of finding out about a world of external physical objects, it seems that all of your beliefs about them would be rendered unjustified.[58]

This remains so even if we weaken the case a bit, so that the DUI is not wholly responsible for your perceptual beliefs. Suppose you are informed that the DUI is only *one* significant influence on the content of your perceptual judgments. Nonetheless, you learn, this influence is such that a) you have no way of isolating any perceptual judgments that are known to be free of the influence of the DUI, and b) the influence of the DUI is sufficiently powerful that the following is true: had the DUI influenced you differently, you would make radically different perceptual judgments. It seems to me that learning of even this more modest influence

[58] It is worth noting that if you took a phenomenalist view, according to which talk of physical objects is simply talk about what perceptual experiences you would have under various circumstances, this skeptical problem would disappear. Such a view might become quite tempting if you were to find out that all your perceptions were implanted by the DUI. At any rate, the point is that the skeptical problem, both in the evolutionary case and in this one, arises only when we assume *realism* about the domain in question.

of the DUI on your perceptual judgments has deep skeptical consequences. Continuing to believe that your perceptual beliefs accurately represent an external physical reality in such a case requires trusting that the influence of the DUI has pushed you toward, rather than away from, the truth. But this is exactly what you have seem to have no reason to believe, and no way of figuring out.

In the case of the DUI, what defeats our initial *pro tanto* justification for our perceptual beliefs is that we justifiably accept an account of the origins of those beliefs that a) rules out a direct tracking explanation and b) gives us no reason to prefer an indirect tracking account to a distortion account. Once we have this, the initial warranted confidence we had in regarding those beliefs as faithful representations of an external physical reality disappears. Yet if our initial hypothesis about evolutionary influence on the content of our moral judgments is true, then—at least once we come to realize its truth—we seem to be in an analogous epistemic situation when it comes to moral matters. For we will have identified one deep influence on the content of our moral judgments, where the best explanation of the nature of this influence a) rules out a direct tracking account and b) gives us no reason to prefer an indirect tracking account to a distortion account. Once we have this in hand, it seems that—at least insofar as we regard our moral beliefs as attempts to represent a stance-independent moral reality—all of our moral beliefs will be unjustified.

7. Conclusion and Preview

Let us briefly review what I have tried to accomplish in this chapter. I began by considering a general phenomenon (i.e., human evolution) that has long been thought to raise a serious

challenge to moral realism, and stated a plausible (though somewhat speculative) hypothesis about it: evolutionary forces have had a significant influence on the content of our moral judgments. I did not attempt to establish this hypothesis empirically, but chose it both for its plausibility and because it is thought by some to have metaethical implications. Second, I considered an influential argument that attempts to move from this hypothesis to an anti-realist conclusion, and I argued that this argument is inadequate as it stands. Third and finally, I argued that the hypothesis in question really does raise a serious challenge for the moral realist, and in fact (if true) undermines the realist's claim that we have justified beliefs about stance-independent moral truths.

In the next chapter, I will focus on a distinct phenomenon that has also long been thought to raise a serious challenge to moral realism: the existence of widespread, intractable moral disagreement. My general strategy there will be very similar. First, I will articulate a hypothesis about the character of existing moral disagreement. Second, I will examine several influential arguments that attempt to move from a premise about widespread moral disagreement to an anti-realist conclusion, and will argue that these arguments are inadequate as they stand. Third and finally, I will construct a novel argument from disagreement against realism. Like my evolutionary argument, this argument from disagreement will target the epistemological commitment of realism, and I will argue that, if my hypothesis about the nature of existing moral disagreement is correct, then we do not have any justified beliefs about stance-independent moral truths. By the end of the next chapter, then, I will have presented two entirely distinct arguments for this anti-realist conclusion. Let us turn now to the phenomenon of moral disagreement.

Chapter 2: The Epistemological Challenge of Disagreement

It is a striking sociological fact that there is quite a bit of disagreement about moral matters. This is perhaps most obvious when looking at different cultures and different time periods, but even within contemporary American society, intractable disagreements about moral issues are far from rare. The extent and persistence of moral disagreement has long been thought to pose a challenge to conventional ways of thinking about morality. The literature on this topic is large, but most attempts at fleshing out the challenge can be divided roughly into two categories. *Metaphysical* arguments from disagreement contend that the phenomena of (actual or possible) moral disagreement give us reason to think there are no stance-independent moral facts. *Epistemological* arguments from disagreement, in contrast, allow for the sake of argument that there are such facts, and try to show that we cannot be justified (or warranted) in any of our judgments about them once we become aware of widespread and intractable disagreements.[59] Both types of arguments raise a direct challenge to moral realism.

I will aim to accomplish three tasks in this chapter. The first is to identify exactly what kind of disagreement would raise the most serious challenge to moral realism. Whether the kind of disagreement I give an account of here is in fact instantiated, and if so to what extent, is a complicated and largely empirical question that is beyond the scope of this chapter. Nonetheless, I think that specifying the nature of this kind of disagreement and examining its philosophical implications is a worthwhile endeavor for two reasons. First, if I am right that the kind of disagreement I discuss raises the *strongest* challenge to moral realism, then insofar as the realist can adequately respond to the challenges it raises, she can rest assured that other kinds of

[59] There is a third type of argument from disagreement that does not fall neatly into either category, namely *semantic* arguments from disagreement. See Loeb (1998) and Tersman (2006). I discuss such arguments in Chapter 3.

disagreement will not pose any serious threat to her views. On the other hand, if the kind of disagreement I describe *would* raise a serious challenge to moral realism, then the theory is vulnerable to empirical refutation, and must defend itself on empirical as well as philosophical grounds.

My second aim is to show that two influential attempts at formulating anti-realist arguments from disagreement are unsuccessful. First, I consider metaphysical arguments that claim that the best explanation of moral disagreement entails that there are no stance-independent moral facts. While realists have responded to such arguments at length, many such replies rely on speculation about the character of existing moral disagreement, and thus remain vulnerable to empirical challenge. I present a different reply to this kind of skeptical argument, one that realists should find preferable insofar as it involves no such speculation. I then I examine an influential epistemological argument—inspired by some passages from Henry Sidgwick—for the conclusion that we are rationally required to suspend judgment about a very wide variety of moral questions due to widespread disagreement. This argument, too, suffers from a fatal flaw, as the general epistemological principle it relies upon turns out to be indefensible.

Third, and finally, I develop a novel epistemological argument for the claim that widespread moral disagreement of a particular sort would undermine our ability to gain justified beliefs about any stance-independent moral facts. As I mentioned before, whether disagreements of the sort I shall discuss are actual (and if they are, the extent to which they occur) is a difficult question that I do not attempt to settle. Here I simply attempt to determine the extent to which such an empirical discovery would threaten the epistemological commitment of moral realism.

Fundamental Moral Disagreement

So let us begin by asking: what sort of disagreement would raise the most serious challenge for moral realism? Plainly, the mere fact that people disagree about an issue does not entail that there is no objective fact of the matter, or that knowledge of that fact is impossible. The mere fact that some people deny that the earth is round clearly does not suggest that the earth has no objective shape, or that this shape is unknowable. If moral disagreement is to raise a serious challenge, it must be in virtue of some particular feature that renders the disagreement in question especially problematic. Perhaps the easiest way to determine what such a feature might be is to enumerate some of the unproblematic types of moral disagreement, and then see what (if anything) remains.

As realists have pointed out, many actual moral disagreements can be explained by the fact that at least one party is irrational, biased, or is ignorant or misinformed about some underlying non-moral fact.[60] Such cases do not raise any pressing skeptical worries. We should expect that bias, irrationality, and misinformation will frequently lead to mistaken verdicts, not only in the moral domain but quite generally. Thus, if we have good reason to think that the moral views of those who disagree with us on particular occasions can be attributed to these factors, the fact that they disagree with us does not provide any reason to doubt the reality of the relevant domain, or the justification of our own beliefs. If any moral disagreements are to raise worries for realism, they will have to be disagreements that could not be resolved simply by providing the parties with relevant non-moral information, or by correcting logical errors, or by removing some source of bias.

---

[60] Cf. Brink (1989), Thomson (1996), and Shafer-Landau (2005).

Disagreements meeting these criteria certainly seem possible, and indeed are almost certainly actual. Consider longstanding disagreements about the moral status of non-sentient nature. Some people believe that non-sentient living things, such as trees, have intrinsic moral value that gives us reason to preserve them for their own sakes. Others believe that communities, such as ecosystems, have intrinsic moral value that is not reducible to the value of their members. Still others deny these claims, and hold that only sentient organisms have such intrinsic value.

Similarly intractable disagreements occur concerning the moral permissibility of capital punishment. Some hold that it is always morally impermissible to kill a defenseless person who poses no clear threat to others, while others claim that this is permissible, and some retributivists claim that the death of a murderer is in fact morally good in itself. Examples of this sort could be multiplied. The crucial feature of such disagreements is that they often persist after both parties are informed of the relevant non-moral facts, even in cases in which neither side seems to be unduly influenced by self-interest or related forms of bias. Furthermore, it seems that parties to such disagreements needn't hold their position on the basis of any logical error.

What could explain disagreements of this sort—moral disagreements in which neither party is ignorant of or misinformed about non-moral facts, guilty of logical error, or objectionably subject to bias? It seems that such cases must ultimately trace back to a difference in the *moral intuitions* of the parties involved.[61] To some people, certain properties (such as sentience) *seem*, upon reflection, to be of paramount moral importance, while it seems to others

[61] There are a number of competing accounts of the nature of moral intuitions. I take it to be fairly uncontroversial that intuitions are seeming states with propositional content. There is significant debate about how else to characterize them—for example, whether they are a species of belief or are *sui generis*. Such disagreements are not relevant to the arguments of this chapter. I will assume that moral intuitions are fallible (i.e., a mental state can count as an intuition even if its content is false), but beyond this, I will not take a stand on the nature of intuitions.

that different properties (perhaps, having interests) are crucial to whether a being has moral status.

Although some have suggested that disagreements arising from differences in intuition are simply rationally unresolvable,[62] this is surely too quick. For in such situations we still have a powerful tool at our disposal: the method of reflective equilibrium.[63] It might be that one party's judgment about the particular case at hand fails to cohere with her other moral judgments and the moral principles that she accepts upon reflection, or with other of her background beliefs. If so, these other judgments might be used as premises in reasoning, with the end goal of bringing one's judgments about cases and principles into alignment with one another, and with one's broader philosophical and scientific views. Such coherentist reasoning might produce agreement even when moral intuitions initially diverge.

But then again, it might not. We should expect it to do so in cases where the two parties have largely similar intuitions but one party has an aberrant intuition that does not cohere well with the remainder of her intuitions. There might well be cases, however, where differences in intuition are *systemic*, such that the intuitions of the two parties differ in such a way that even perfectly scrupulous coherentist reasoning would lead the parties to diverging states of reflective equilibrium. These cases, I believe, raise the most serious epistemological challenge to the moral realist. For the purposes of this chapter, I shall refer to disagreements that are ultimately explained not by lack of non-moral information, logical error, or bias, but by systemic differences in moral intuition, as *fundamental moral disagreements*.

On the face of it, the widespread existence of fundamental moral disagreements would

---

[62] Cf. Ayer (1952), p. 147.
[63] Cf. Goodman (1983), Rawls (1999), and Daniels (1979).

seem to raise a skeptical challenge for anyone who believes in non-relative moral truth. If I have a disagreement with an equally rational and (non-morally) informed interlocutor and this disagreement is ultimately traceable to a systemic difference in our intuitions, then at least one of us must have systemically misleading intuitions. Given that we are equally rational and non-morally informed, why should I think that my intuitions happen to reflect the truth while my opponent's do not? And if there's reason to think that even our most careful moral judgments fail to reliably track any stance-independent moral reality, is there any remaining reason to believe in such a reality at all?

Inference to the Best Explanation

Such rhetorical questions are not yet an argument, of course, and it is far from clear how to turn them into one. One popular way of doing so goes briefly as follows.

Inference to the Best Explanation (IBE)

1) There is widespread moral disagreement, much of which is intractable.
2) The best explanation for this is that there are no objective moral facts.
3) Therefore, (probably) there are no objective moral facts.

J.L. Mackie has this sort of argument in mind when he writes that "the actual variations in the moral codes are more readily explained by the hypothesis that they reflect ways of life than by the hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values."[64]

Realists have spilled a great deal of ink responding to *IBE* and closely-related arguments. In particular, many have proposed alternative explanations of the "actual variations in moral

[64] Mackie (1977), p. 36.

codes" that are perfectly consistent with realism. For instance, a good deal of actual disagreement can be attributed to the effects of self-interest, ideology, prejudice, logical error, and non-moral factual disagreement, influences that quite commonly distort our beliefs in domains about which we are inclined to be realists. Many realists express confidence that taking account of such influences will suffice to explain most actual disagreement. David Brink, for example, insists that "most moral error is in principle correctable by coherentist reasoning."[65] Richard Boyd writes, even more boldly, that "agreement on nonmoral issues would eliminate *almost all* disagreement about the sorts of moral issues which arise in ordinary moral practice."[66] These realists suggest that since most moral disagreement is explained by these commonplace phenomena, explaining moral disagreement raises no particular worry for the realist.

If *fundamental moral disagreement* is widespread, however, this strategy will not do. For fundamental disagreement is precisely the sort of disagreement that *cannot* be resolved by removing non-moral ignorance, logical error, or the influence of self-interest and other forms of bias. Insofar as realists base their arguments on the assumption that these account for all (or nearly all) moral disagreement, they leave their position vulnerable to empirical refutation. This vulnerability would be removed if the realist could show that *IBE* can be answered without appealing to such speculations about the character of moral disagreement.

I believe that it can. To see why, we must dig more deeply into *how* exactly the absence of objective moral facts is supposed to explain disagreement. The basic argument, as David Brink puts it, relies on the following counterfactual: "If there were objective moral facts to be discovered, one would expect convergence of moral belief at least over time."[67] But we do not

---

[65] Brink (1989), p. 204.
[66] Boyd (1988) p. 213
[67] Brink (1989), p. 197.

see such convergence, the argument goes, and thus we should not believe in any such objective moral facts.

There are a number of points to make here. First of all, it should be noted that the presence of objective facts about a domain certainly does not *entail* that people will converge in their beliefs about that domain over time. It could be that a particular area of inquiry is simply very complicated, or that we lack sufficient evidence on which to base our beliefs. As Brink is careful to note, the matter is one of *expectation.* In domains of objective fact, intractable disagreement is, in some sense, surprising.

But this is not yet enough to warrant even an abductive inference from lack of convergence to anti-realism. At the very least, we would need an additional premise to the effect that lack of convergence would be *less* surprising if we assume anti-realism. And it's not entirely clear that this is so. There are many domains about which most of us are reluctant to believe in objective facts, but in which we nonetheless find a striking degree of convergence. Human beings converge to a large degree in their judgments about what general kinds of food are tasty, and to an even greater degree in finding certain things disgusting (e.g., feces and other bodily fluids, rotting food, etc.). Yet this convergence is not particularly surprising, and does not incline most of us towards realism about these domains. The point here is that insofar as there is a correlation between realism and convergence, it is a fairly weak one. Many domains of real fact are subject to protracted disagreement, while many domains about which anti-realism is plausible are areas in which our judgments converge significantly.

This suggests that the prospects for explaining moral disagreement *simply by citing a metaethical hypothesis* are unpromising. The absence of objective moral facts would not explain

why moral claims are subject to protracted disagreement, rather than to the kind of convergence we see about the disgustingness of feces. Indeed, this is presumably why anti-realists typically invoke mundane causal facts in their explanations. Mackie, as we've seen, attributes differences in moral codes not simply to the absence of moral facts, but to diverging "ways of life." Brian Leiter, in a similar vein, attributes to Nietzsche the view that the best explanation of moral disagreements among philosophers "is the absence of any objective fact of the matter about foundational moral questions *conjoined with... the psychological needs of philosophers which lead them to find compelling dialectical justifications for very different basic moral propositions.*"[68]

But there's something odd about explanations like the one Leiter attributes to Nietzsche: the metaethical component—namely, the denial of moral facts—*simply doesn't seem to do any work*. Consider a simple analogy. Suppose I arrive at the office one morning and see that the ground is wet outside. I wonder what explains this, and speculate that perhaps the gardener has run the sprinkler system the night before. But then you provide me evidence that it rained last night, and the fact that it rained is sufficient to explain the moisture on the ground. You might also tell me, as an afterthought, that there is no sprinkler system installed. But it would be very odd if you were to say: the best explanation of the moisture on the ground is that it rained last night *and there is no sprinkler system*. It is simply very hard to see how this addition about the absence of a sprinkler system contributes anything to the explanation of the moisture.

The moral case seems similar. Suppose that Mackie is correct that moral disagreements are often explained by the fact that parties to moral disputes adhere to different ways of life and

[68] Leiter (2014). (My emphasis.) Note that although Leiter attributes this argument to Nietzsche, he also himself endorses the argument.

this adherence influences their moral judgments. This fact (if it is a fact) seems perfectly sufficient to explain the disagreement in question. Adding a qualification to the effect that *there are no objective moral facts* simply doesn't seem to add anything to the explanation.

One might object here by pointing out that such negative existential claims do sometimes play a legitimate role in explanations. For example, if a plane were to fall from the sky and crash into the ground, we might well explain this by citing the various forces on the plane *and the fact that there was no pilot in the cockpit at the moment*. This explanation is perfectly sensible in virtue of the fact that had there been a pilot in the cockpit, things would have happened differently.

It is possible to argue that, similarly, if there *were* moral facts, they would somehow causally override the distorting factors cited by the anti-realists, and thereby generate convergence. But this is a rather bold metaphysical claim concerning what moral facts would have to be like, and I see no reason that the realist should have to accept it. Not only are there prominent realist views according to which moral facts are wholly non-causal[69], but even those who believe in the causal efficacy of moral properties could simply hold, in a way that does not seem ad hoc, that the causal powers of such facts are often outweighed by the many other forces that causally influence our moral judgment.

It does seem plausible to me that the best explanation for moral disagreement will probably not *make reference* to any moral facts. But it also seems unlikely that such an explanation will involve *denying the existence of such facts*, any more than the best explanation of the wet grass will involve denying the existence of the sprinkler system. At most, the anti-realist might say this: the best explanation of moral disagreement entails that our moral

[69] See, for example, Huemer (2005), Enoch (2011), and Parfit (2011).

judgments are seldom or never caused by the moral facts. If this raises a challenge, however, it does not fundamentally arise from the fact of disagreement, but from the causal claim.[70] Indeed, if it is true that our moral judgments are seldom caused by the moral facts, the challenge this presents would seemingly be just as significant even if moral agreement were *universal*. And as I mentioned before, a number of reputable realists are antecedently committed to the causal inefficacy of moral properties, and thus already hold that our moral beliefs are *never* caused by the moral facts.

For these reasons, it seems to me that *IBE* and arguments like it are unpromising as challenges to moral realism. The guiding idea behind such arguments is that disagreement seems to show that our moral beliefs are often caused by something other than the moral facts, whether ways of life, psychological needs, or something else. But it is difficult to see why the realist cannot simply accept whatever causal story the anti-realist presents in its entirety (minus, of course, the proviso that there are no moral facts). If the moral beliefs of many people are being influenced by the forces the anti-realist mentions, it's no wonder that many of us often fail to arrive at the moral truth. There's no need to speculate that the causal influences in question are of a particular kind, as Boyd and Brink do. For although the best explanations of moral disagreement may make no mention of moral facts, they also make no mention of their non-existence.

Sidgwick's Principle

I've argued against the claim that the best explanation of disagreement is that there are no stance-independent moral facts. But perhaps the problem with disagreement is not that it shows directly

[70] For arguments of just this sort, see Harman (1977), Street (2006), and Bedke (2009).

that there are no such facts, but that it undermines the justification of our beliefs about them. After all, if our moral beliefs are heavily shaped by the sorts of forces anti-realist mention, how could we have any reasonable confidence that we are getting it right, especially when intelligent people of good will disagree with us? Several authors have tried to develop this thought by leaning on some key passages from Sidgwick's *Methods of Ethics*. Sidgwick writes:

> [I]f I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me to a state of neutrality.[71]

Sidgwick's wording suggests a descriptive reading of this remark: he seems to be saying that given human psychology, "a state of neutrality" is the inevitable outcome of reflecting on such disagreements. Whether or not Sidgwick intended it, however, many authors have found a normative reading of Sidgwick's remarks to be a more plausible one. Roger Crisp, for example, endorses what he calls "Sidgwick's principle," which states, "A person who judges that *p*, if she finds that some other person judges that *not-p*, and if she has no reason to believe that other person to be in a worse epistemic situation than her, should suspend judgment on *p*."[72] Let us leave aside the interpretive question of whether Sidgwick actually meant to endorse such a normative principle, and follow Crisp in referring to this as "Sidgwick's Principle." According to Crisp, the truth of Sidgwick's principle, along with actual widespread disagreement in moral intuitions, requires that we suspend judgment about all controversial questions of ethical theory.[73]

Like many conciliatory positions in the epistemology of disagreement, Sidgwick's

---

[71] Sidgwick (1981) p. 342.

[72] Crisp, (2012). McGrath (2008) and Wedgwood (2010) consider similar principles.

[73] McGrath (2008) argues for the weaker conclusion that our beliefs about such questions do not amount to *knowledge.*

principle has some intuitive appeal.[74] If we disagree about the truth of a proposition and I have no good reason to think that you are more likely to be mistaken than I am, it seems to be mere dogmatism to retain my belief in the face of our disagreement. Just as it would be irrational, in a case in which our watches showed different times, to trust my watch over yours simply because it was *mine*, it seems equally irrational to trust my belief rather than yours, without having any grounds for thinking that I'm more likely to be correct than you are.

Nonetheless, skeptical arguments based on Sidgwick's principle face serious problems. The first question we must ask is this: what kinds of considerations can count as reasons for thinking that the other party to a disagreement is in a worse epistemic situation than oneself? Suppose I meet a stranger on the street, and we begin a conversation in which I tell her that I believe some proposition *p*. She replies that she believes *not-p*. Knowing nothing about her grounds for holding *not-p*, do I have any reason to believe that she is in a worse epistemic situation than I am? If I do have such a reason, it seems that this reason must be *the very fact* that she believes *not-p.* And indeed, we can think of cases in which this is very plausible. If I strike up a conversation with a stranger and she tells me that she believes that the Earth is flat or that demons cause mental illness, I can be pretty confident that either she is not acquainted with the relevant facts or that she is to some significant degree irrational, and therefore is in an epistemic situation that is inferior to mine.

Suppose that any time you believe a proposition *p,* you could always reasonably count the fact that another person believes *not-p* as evidence that that person is in an epistemic situation inferior to yours. If this were so, Sidgwick's principle would not have any skeptical implications,

---

[74] For similarly conciliatory positions regarding disagreement generally, see Christensen (2007), Elga (2007), Kornblith (2010), and Fumerton (2010).

for you would always have *some* reason to believe that those who disagree with you are in an epistemic situation inferior to your own. This reason could of course be overridden; for example, if the person who disagrees with you is an acknowledged expert and you are not, this might give you much stronger reason to think that she is in a better epistemic position than I am. But such cases are very rare when it comes to morality, as there are very few moral disagreements in which one party is an acknowledged moral expert while the other is not.

It might seem objectionably circular, however, for Mary to count the mere fact that John disagrees with her as evidence that John is in an inferior epistemic situation concerning the issue at hand. Certainly, it would seem unjustified for me to epistemically demote an eminent scientist simply because she reports findings that conflict with my current beliefs. Given this, one might want to stipulate that Sidgwick's principle requires *independent* grounds for thinking that the other party's epistemic situation is inferior to one's own, grounds that make no reference to their belief that not-*p*.[75] But this faces two problems. First, it seems to allow the mere fact of disagreement with complete strangers to undercut the justification of even our deeply held beliefs. Consider again the person on the street about whom I know nothing, who suddenly tells me (apparently quite sincerely) that the Earth is flat. Knowing nothing else about the epistemic situation of the person, I have no grounds for believing her to be in an inferior epistemic position, other than the fact that she has just endorsed a patently false claim. If her endorsement of the claim cannot give me reason to think that she is in an epistemically inferior position, then Sidgwick's principle requires me to suspend judgment about the shape of the Earth, at least until I find out more about her grounds for holding that the Earth is flat. But this is absurd. The mere fact that someone about whom I know nothing disagrees with me should not suffice to

[75] Cf. Christensen (2007) p. 198, Elga (2007) p. 489.

undermine the justification of my deeply held and well-supported beliefs.

Furthermore, if Sidgwick's principle requires that we have independent grounds for holding those with whom we disagree to be in an epistemically inferior position, it seems likely that no philosopher is in a position to reasonably believe Sidgwick's principle. That's because a number of very well-respected philosophers reject Sidgwick's principle.[76] Unless proponents of the principle can identify some compelling reason to believe that those philosophers are in an inferior epistemic position to evaluate the merits of Sidgwick's principle, Sidgwick's principle actually requires them to suspend judgment about the principle itself. And clearly if we must suspend judgment about Sidgwick's principle, no argument which relies on it can be rationally persuasive.

Those who attempt to draw skeptical conclusions from Sidgwick's principle therefore face a dilemma. Consider a case in which a very smart philosophical colleague whom I generally regard as my equal disagrees with me about some moral issue. Either the mere fact that such a person disagrees with me can count as a reason to believe that she is in an epistemic situation inferior to mine, or it cannot. If it can, then Sidgwick's principle has no skeptical implications, for then we may surely reasonably conclude that those with whom we have fundamental moral disagreements are in epistemically inferior situations. If it cannot, then Sidgwick's principle not only has implausible implications about disagreements with strangers, but it also turns out to be self-undermining. Either way, Sidgwick's principle provides no basis for suspending judgment about our moral judgments. If fundamental moral disagreement is to give us grounds for skepticism, we must look elsewhere.

---

[76] For example, see Kelly (2006) and (2010), and Weatherson (2012).

Fundamental Disagreement and Reliability

While Sidgwick's principle focuses on particular moral beliefs in isolation, I suggest that we shift our focus to consider the implications of the existence of fundamental disagreements for the reliability of our moral faculties more generally. Even the most robustly anti-skeptical realist will admit that many of us often form moral beliefs in unreliable ways, and come to moral verdicts that are epistemically unjustified. In order to focus the discussion, let us focus on those moral beliefs that are the most likely to be justified, namely moral beliefs that are the outcome of our best method of moral inquiry, responsibly applied. While debates rage on about the structure of *justification* in ethics—whether we should be foundationalists, coherentists, reliabilists, contextualists, etc.—there seems to be a remarkable degree of consensus about how moral inquiry should proceed in practice. Despite deeper epistemological disagreements, most philosophers now agree that our best method of moral inquiry involves trying to bring our moral judgments into reflective equilibrium in light of all the relevant non-moral facts. Doing so involves revising our particular moral judgments in light of the principles we accept, and revising those principles in light of other moral judgments and still other principles, with the goal of achieving a set of moral views that is both internally coherent and fits comfortably within the framework of our non-moral beliefs about the world.

Moral intuitions necessarily play a role in applying this method. For if we are to get the process of reflective equilibrium off the ground, we need some moral starting points. And there seems to be no more plausible place to start than our moral intuitions, or at least some subset thereof.[77] If there is widespread fundamental moral disagreement, however, then many people's

[77] This seems to be true regardless of whether we conceive of intuitions as non-inferential moral beliefs, or as non-doxastic *seeming* states, on the basis of which moral beliefs may be (non-inferentially) formed.

moral intuitions are unreliable. And given the above characterization of fundamental moral disagreement as involving *systemic* differences in intuitions, even the application of our best methods of moral inquiry will not be able to correct for this unreliability. Thus the existence of widespread fundamental disagreement would give us very strong reason to believe that the way that many humans come to form moral beliefs about controversial issues, even when they responsibly apply our best method of moral inquiry, is unreliable. Sufficient evidence that many humans form moral beliefs unreliably even when they're being maximally epistemically responsible would, in the absence of any evidence that I am an exception to this general trend, undermine my claim to be justified in believing disputed moral propositions.

More formally, the argument would go like this:

The Reliability Argument from Disagreement (RAD)

1) If there is widespread fundamental moral disagreement, then many people would have false moral beliefs even if they were to flawlessly employ our best method of moral inquiry.
2) If many people would have false moral beliefs even if they were to flawlessly apply our best method of moral inquiry, then that method is not reliable for many people.
3) If our best method of moral inquiry is not reliable for many people, then one cannot be justified in believing the outputs of such a method unless one has special reason to believe that it is reliable in one's own case.
4) It is not the case that any of us has special reason to believe that the best method of moral inquiry is reliable in our own case.
5) Therefore, if there is widespread fundamental moral disagreement, then we cannot be

justified in believing the outputs of our best method of moral inquiry.

Although the conclusion of this argument only concerns a particular method of moral inquiry, it plausibly generates a general epistemological challenge for moral realism. For if we cannot form justified beliefs on the basis of our best methods of moral inquiry, it is difficult to see how we could form any justified moral beliefs at all.

One reply to the argument would simply be to concede that widespread fundamental disagreement *would* undermine moral realism, but deny that we currently have sufficient evidence for the existence of widespread fundamental moral disagreement. I will have very little to say about this response. Although I suspect that fundamental disagreements play a role in explaining a number of moral disagreements (including, for example, issues about the moral status of fetuses, non-human animals, and non-sentient nature), my main concern here is with what the philosophical implications would be *if* fundamental moral disagreement were shown to be widespread. So for now, I'll simply note that serious further empirical work would be required to determine whether this response is successful.

So how else might the realist respond? Premise 2) is intended to state a conceptual truth about reliability. Premise 1) also seems fairly secure, provided one accepts the consensus view that some version of wide-reflective equilibrium is our best method of moral inquiry. For fundamental moral disagreements, on my construal, are precisely those disagreements which would withstand the application of such method in the limit of inquiry. Thus, it seems to me the most natural places to resist the argument are premises 3) and 4).

One way of resisting 3) would be to invoke an internalist account of justification. Justification, on internalist views, is not a matter of whether our beliefs reliably "hook up" with

an external world, but whether we respond in epistemically responsible ways to our consciously accessible internal states, such as beliefs, seemings, and the like. An internalist might argue that while the existence of widespread moral disagreement shows that many people have *false* moral beliefs, it does not necessarily show that these beliefs are unjustified. False beliefs can be justified, as in cases where the parties have misleading evidence from which they reason flawlessly.

Clearly, there are cases of justified false belief. In many cases, the justification of false beliefs rests on an agent's non-culpable ignorance of a *defeater*—a consideration that would undermine the justification for the agent's belief were she to become aware of it. But the fact (if it is a fact) that even maximally epistemically responsible agents following our best methods of moral inquiry often fail to arrive at the moral truth would seem to be just such a defeater. While an internalist might plausibly hold that we are justified in trusting some belief-forming method in the absence of proof of its reliability, it is far less plausible to claim that we are justified in continuing to trust such a method after it has been shown to be unreliable in many cases. So, it seems that the strongest claim that this internalist reply can establish is that we are justified in our moral beliefs only prior to discovering that these beliefs are subject to fundamental moral disagreement.

Given that this is dissatisfying, the internalist might shift her focus to premise 4) and claim that we (or at least many of us) have reason to trust the results of *our own* moral inquiry, even in the face of evidence that the method we are employing is generally unreliable. One way of supporting such a view would be to claim that we are entitled to—indeed one might go so far as to claim that we *must*—place a level of fundamental trust in our own beliefs and belief-

forming processes that we need not grant to the beliefs of others. The thought here is that in cases of disagreement, one cannot simply treat oneself and one's interlocutor as a pair of "truthometers" that give different readings about a particular issue, for one cannot escape adopting an ineliminably first-person perspective. As David Enoch puts the point, "Whenever you try to decide how much trust to place in someone, or indeed, when deliberating epistemically about anything at all, your starting point is and cannot but be your own beliefs, degrees of beliefs, conditional probability, epistemic procedures and habits, and so on."[78] Similarly, Ralph Wedgwood argues that because I can base my beliefs *directly* on my own intuitions, while I can take the intuitions of others into consideration only via my beliefs about them, it is permissible to grant my intuitions a kind of epistemic priority that I do not give to the intuitions of even my most respected peers.[79]

These considerations have some force. It is impossible to treat one's own beliefs and the beliefs of others as playing exactly the same epistemic role, if only because the task of assessing whether another's views deserve epistemic respect inevitably requires relying on one's *own* judgments. *Perhaps* such considerations can be used to provide a compelling argument against "equal weight" views in the epistemology of disagreement. But the argument of this section does not depend on the truth of any such view. The argument does not attempt to apply a principle concerning how to respond to an isolated case of disagreement. The question is not whether I am entitled to favor my moral belief *over yours*, but rather whether, if provided with evidence that even our best method of moral inquiry is often an unreliable guide to the truth, I am entitled to rely on the outputs of this method at all.

---

[78] Enoch (2011), p. 980.
[79] Wedgwood (2010), pp. 237-244.

A different way of resisting premise 3) relies on externalism about justification, the view that whether a belief is justified does not depend solely on the accessible mental states of the agent, but also on the connections between the agent's mental states and the world. An externalist might point out that widespread fundamental disagreement is consistent with the possibility that *some* of us have reliable moral intuitions, and are thereby reliable at determining the moral truth. For perhaps, an externalist might argue, widespread fundamental disagreement would not show that our best method of moral inquiry is *generally* unreliable, but only unreliable in certain circumstances. We can avoid the skeptical conclusion simply by individuating our methods of moral inquiry more finely. Perhaps *my* method (reflective equilibrium from *my* starting beliefs) is reliable, even if yours (reflective equilibrium from *your* starting beliefs) is not. In the absence of a conclusive reason to think *my* method is unreliable, there is no defeater here. In light of this, one might argue that (at least) those who have reliable intuitions are justified in continuing to trust their own intuitions, and the moral judgments based on them.

Only the most extreme externalists should accept this, however. Even if we accept that there is an externalist component to justified belief, the following principle seems very plausible: if one possesses evidence of the widespread unreliability of a type of source of belief, one is not justified in trusting a source of that type without independent evidence of its reliability. As an example, consider the following case. Suppose I were to pick a watch at random out of a basket of 100 watches, 50 of which worked perfectly and 50 of which were systematically misleading. Suppose, too, that I know that half of these watches are unreliable. If I ended up with a reliable watch and formed my beliefs about the time on the basis of this watch, I would be able to reliably tell the time. Absent some sort of confirmation, however, it seems that my beliefs about

the time would not be justified, as I lacked any reasonable grounds for believing that I ended up with a reliable watch. If there are widespread fundamental moral disagreements, we seem to be in a similar situation regarding moral knowledge. Some of us have intuitions that are sufficiently reliable that the method of reflective equilibrium will allow us to arrive at the truth. Others have systematically misleading ones. Absent any mechanism of checking whose intuitions are reliable, it seems doubtful that we would be warranted in trusting our intuitions in such circumstances.

Walter Sinnott-Armstrong invokes a similar analogy[80] in order to argue against the claim that our moral intuitions can be *non-inferentially* justified. Citing empirical evidence that our intuitions are unreliable in a number of circumstances, Sinnott-Armstrong argues that we must have inferential confirmation before we can trust our moral intuitions about cases. But the argument of this chapter cuts deeper: if there is fundamental moral disagreement, even inferential confirmation cannot justify our moral beliefs. For such confirmation will inevitably involve checking some moral judgments against others and against our non-moral beliefs. In cases of truly fundamental disagreement, however, this will not save us, for many of us will be unreliable even after we do this fully.

Does the RAD suffer from the Generality Problem?

The basic worry behind the RAD is that if fundamental moral disagreement is widespread, then even our best method of moral inquiry is unreliable. The argument's focus on reliability might call to mind *reliabilist* theories of justification and knowledge. Reliabilists hold that a belief is justified (or, alternatively, constitutes knowledge) if and only if the belief is the product of a

[80] Sinnott-Armstrong (2010), p. 71.

reliable belief-forming process.[81] Although the RAD invokes the notion of reliability in order to establish a conclusion about justification, it is worth noting that the argument does not presuppose a reliabilist view of justification. Whereas reliabilists traditionally hold that reliability is both necessary and sufficient for justification, all RAD needs to get off the ground is the claim that the unreliability of a method of inquiry is a *defeater* for beliefs formed by that method. Despite this significant difference, one might still worry that the RAD drifts *close enough* to reliabilism to be beset by a classic problem for reliabilism: the generality problem.[82]

The generality problem can be stated fairly simply. According to reliabilism, the justification of a belief depends on the reliability of the process by which that belief was formed. But the process by which a given belief was formed can be described at an indefinite number of levels of generality. So, my belief that there is a cat in front of me might be aptly described as the result of perception, of seeing, of seeing at night, of seeing at night in a well lit room, and so on. But these different descriptions might pick out processes that are reliable to different degrees. (Perhaps seeing at night generally is fairly unreliable, but seeing at night in a well lit room is quite reliable.) The question for reliabilists is how we go about figuring out which level of description is the epistemically relevant one. The problem is that we seem to have no non-arbitrary way of doing so.

One might level the same objection against the RAD. One might concede that there is bound to be some description under which our moral beliefs were formed by an unreliable process.[83] Still, one might contend, for any true moral belief, there will also be *some* description under which it was formed by a reliable process. For, in the limit, one could simply craft a

[81] See Goldman (1979), Swain (1981) and Alston (1988).
[82] See Feldman (1985) and Conee and Feldman (1998).
[83] Given enough moral disagreement of any sort (fundamental or not), the process *forming a moral belief* might be held to be generally unreliable.)

description so specific that it accurately describes only the formation of that single belief. Such a process would, trivially, have a 100% success rate of forming true beliefs. So, regardless of whether fundamental moral disagreement is widespread, it appears that any true moral belief is going to be formed by a belief-forming process that is reliable under some descriptions, but unreliable under others.

In order to respond, I think that we should first notice that there are two components to the worry that can be teased apart. The first is that any belief-forming process will fall under many descriptions. The second is that the selection of one of these descriptions as epistemically relevant threatens to be arbitrary. In order to respond to the challenge, the reliabilist about justification *must* insist that there is exactly one such description that is relevant when it comes to justification. For to fail to do so would be to invite contradictions: after all, the same belief-forming process can be reliable under one description but unreliable under another. If *both* descriptions were allowed to be epistemically relevant by the reliabilist, a single belief would turn out to be both justified and unjustified, which is unacceptable.

Since I am not committed to reliabilism—indeed, in the case of the watches above I explicitly rejected the claim that reliability is sufficient for justification—I can defuse both components of the generality problem in a way the reliabilist cannot. First, since I do not claim that the justification of a belief depends solely on the reliability of the process that produced it, I can happily admit the multiplicity of descriptions of belief-forming processes without giving up on my central epistemic claim. Consider again the case in which I believe that there is a cat in front of me, because I see the cat in a well lit room at night. The epistemic principle behind premise three of the RAD holds that if I know that some belief was formed by an unreliable

process, I cannot be justified in holding that belief unless I have some special reason to think that the process was reliable in my own case. In the case at hand, this condition is met. The mere fact that vision at night is generally unreliable does not undermine the justification of my belief that there is a cat in front of me, because I have reason to believe that cases of seeing at night *in well lit rooms* are exceptions to this general unreliability.

Furthermore, while reliabilists might struggle to find a compelling reason to select any particular level of description of belief-forming processes as the epistemically relevant one, my selection of the relevant description of the method involved in forming our moral beliefs is far less arbitrary. I want to focus on the class of moral beliefs that have the *best claim* to being justified, and there is a very broad consensus that some version of wide reflective equilibrium is our best method of moral inquiry. Because of this, it is not unreasonable to think that focusing on it does pick out an epistemically relevant way of carving up methods.

Can We Know Whether Disagreement is Fundamental?

The conclusion of the RAD is conditional in form: it makes a claim about what we *would* be justified in believing if fundamental moral disagreement turns out to be widespread. I've admitted that the force of the argument against moral realism will depend on complicated and largely empirical questions about the extent of fundamental moral disagreement. But here one might raise the following objection. It's not simply the case that we don't *currently* know whether many of the moral disagreements that we observe are truly fundamental. Rather (the objection goes), we have no practical way of *ever* determining whether existing disagreement is fundamental or not. That's because fundamental disagreements as I've characterized them above

are the result of *systemic* differences in moral intuition—differences that would persist even if one were to flawlessly apply our best method of moral inquiry. But, imperfect creatures that we are, not one of us has ever flawlessly applied our best method of moral inquiry. We are so far from doing so, one might argue, that we simply have no way of figuring out whether the moral disagreements we find in the world around us are fundamental or not.

If this objection goes through, then the RAD is truly idle. For if we could *never* determine whether the antecedent of the conclusion is true, then the argument could never rationally persuade a realist that her moral beliefs (insofar as they are taken to be about stance-independent moral facts) are unjustified.

I think this objection is excessively pessimistic about our ability to gain knowledge about the character of existing moral disagreement. As a general matter, we can often have justified beliefs about highly idealized counterfactuals that never obtain in our world. We can know things about how particular physical objects would behave on a frictionless plane, despite the fact that no planes are frictionless. We can determine how such objects would behave if subjected to exactly one force, despite the fact that no object is subject to only one force. The mere fact that a condition (whether a frictionless plane or the idealized limit of moral inquiry) does not obtain and will never obtain does not entail that we cannot know what would happen under that condition.

In the moral case in particular, we have a pretty good idea of what kinds of features make a disagreement more or less likely to be fundamental. Furthermore, we can easily identify a large number of disagreements as non-fundamental: precisely those that rely on disagreement about the non-moral facts, or logical error, or bias, or in which one party has incoherent moral

beliefs. In some cases, however, none of these factors seem to be responsible for the persistence of disagreement. When examining such cases, if we were to search for such factors long enough to no avail, at some point it would become reasonable to conclude that the disagreement in question is fundamental. I have not attempted to such work here, but it seems to me that such work could be done, and that it could yield the (defeasibly justified) verdict that a significant number of actual moral disagreements are fundamental. It is this possibility that gives the RAD its bite.

Conclusion and Preview

The existence of widespread fundamental moral disagreement—disagreement that is explained by a systematic difference in moral intuitions—would raise a deep challenge for the moral realist. This is not, I have argued, because the best explanation of such disagreement would involve denying the existence of stance-independent moral facts. Nor is it, as some have claimed, because we are generally required to suspend judgment in the face of disagreement with those we regard as our epistemic equals. Rather, the existence of widespread fundamental moral disagreement would show that even our best method of moral inquiry is deeply unreliable, in which case we could not justifiably believe that the outputs of such a method correspond to any stance-independent moral truth.

As mentioned previously, determining the extent to which actually existing moral disagreement is fundamental is a difficult task, requiring both empirical research and philosophical judgment (concerning, among other things, what counts as "bias" or "irrationality"). Nonetheless, I hope to have shown that its existence is something worth

worrying about for moral realists. Widespread fundamental moral disagreement would present a deep challenge to their view, and an adequate defense of the position would require providing good reason to think that such disagreements are not often instantiated in our world.

But suppose for a moment that my arguments fail, and the epistemological challenge from disagreement can be met. Would this show that the prospect of widespread fundamental disagreement poses no threat to moral realism? In the next chapter, I will argue that it would not. For the prospect of widespread fundamental disagreement raises an even deeper challenge to realists: it threatens to undermine the realist's claim that in our deepest moral disputes, we are all talking about a single topic of *morality*. In what follows, I will argue that moral disagreement raises not only epistemological challenges for the realist, but semantic challenges as well.

Chapter Three: The Semantic Challenge

Moral realists hold that when we make moral judgments, we are attributing stance-independent moral properties to actions, agents, policies, and so on.[84] In the previous chapter, I argued that disagreement of a certain kind would undermine the realist's contention that we frequently have justified beliefs about when such properties are instantiated. This chapter explores a different challenge that arises from the deep differences in moral view that we find in the world. The worry, in short, is that given the diversity of moral judgments across individuals, cultures, and time periods, there may not be any unique set of properties—*the moral properties*—which all of us are attributing when we make moral judgments. This way of framing the worry puts things in terms of the contents of our moral judgments. But since the primary way that we know about the contents of other people's judgments is via their utterances, a natural way to approach the worry in question is by thinking about moral semantics. From a semantic perspective, the concern is that widespread moral disagreement may ultimately cast doubt on the notion that different speakers are using moral terms to pick out the same properties as one another.

Strictly speaking, the mere fact that some speakers fail to co-refer when employing moral terms would not *necessarily* raise a problem for moral realism. If a small subgroup of English speakers were to begin using the term "morally wrong" to refer to the property of *being square*, this would of course have the result that not all speakers co-refer when using the term "morally wrong." In such a case, the natural thing to say is simply that the speakers in the subgroup have

[84] As I mentioned in the introduction, I employ property talk for ease of exposition, but I want to respect the right of realists to main a general metaphysical nonchalance. What is essential to the realist position is that moral truths are every bit as stance-independent as truths in other robustly objective domains (such as truths about our immediate physical surroundings). If one has quite general reservations about properties, my arguments in this chapter may be translated into the terms of one's favored ontology.

started using moral terms for some purpose other than reporting their moral judgments. Given this diagnosis, we would not regard a speaker from this group who utters "my kitchen table is morally wrong" as disagreeing in any significant sense with someone outside of the group who utters "kitchen tables are not the sort of thing that can be morally wrong." In cases of this sort, absence of co-reference raises no challenge for the moral realist.

However, I will argue that in a wide range of other cases, the absence of co-reference of our moral terms really would undermine moral realism. The argument begins by noting that there is a range of cases that we would naturally identify as *paradigmatic* moral disputes. Such cases include long-standing disputes about a range of controversial issues in practical ethics such as abortion, capital punishment, euthanasia, and same-sex marriage, as well as disputes between proponents of rival ethical theories, such as various versions of utilitarianism, Kantianism, virtue ethics, and so on. In these paradigmatic moral disputes, all parties are sincere and none intend to employ moral terms in some unusual, idiosyncratic sense. Furthermore, the parties to such disputes see themselves as being involved in a disagreement that is, in some important sense, *genuine*, rather than merely resting on a semantic misunderstanding.

Moral realists have a characteristic diagnosis of such disputes: as difficult as they may be to resolve in practice, settling such disputes is ultimately a matter of figuring out the *truth* of the claims in question. Consider the following schematic example, where X is some action subject to moral evaluation, and Bill and Jill are engaged in a paradigmatic moral dispute:

> **Bill:** X is morally wrong.
> **Jill:** No, X not morally wrong.

On the characteristic realist diagnosis, resolving disputes of this form is (in principle) a matter of figuring out which speaker has spoken truly. According to realists, the truth of the utterances in

question will depend on whether X possesses the property of *moral wrongness*, or not. And since presumably nothing can simultaneously possess and not possess the property of *moral wrongness*, at most one speaker will have spoken truly.[85]

If, however, it turns out that in a range of paradigmatic moral disputes of this form speakers are in fact picking out different properties with their moral terms, then the realist diagnosis of such cases is mistaken. For if Bill uses "morally wrong" to pick out some property A, while Jill uses "morally wrong" to pick out some distinct property B, both speakers may well speak truly. If this is so, such disputes cannot be resolved simply by determining who is speaking truly. And if we take the view that parties to paradigmatic disputes can all be speaking truly despite taking seemingly opposed moral positions, then we no longer have a position that seems worth describing as *realism*.[86]

The notion that parties to moral disagreements might fail to co-refer with their moral terms might seem, on the face of it, to be a very strange concern to raise. In ordinary cases of disagreement, we do not conclude that speakers are using their terms to pick out different properties from one another. I would disagree deeply with someone who claims that the earth is flat or is only a few thousand years old, but this disagreement wouldn't incline me in the slightest to think that such a person must be using the words "flat" or "years" in idiosyncratic ways. Indeed, one could make a stronger claim: it seems that it is only when we use our terms to pick out the same things as one another that genuine *disagreement* is possible in the first place. For consider what happens in cases where we fail to co-refer. Suppose I sincerely utter the sentence "Tom went to the bank;" in response, you utter, "Tom did not go to the bank." If I am using

[85] I leave open the possibility that the issue is indeterminate, so that neither speaker has spoken truly. See Shafer-Landau (1994).

[86] This is why, as I explain in the Introduction, I include *invariantism* among the commitments of realism.

"bank" to refer to a financial institution and you are using the word to refer to the edge of a river, there is no genuine disagreement here; we have simply failed to understand one another.

Reflecting on cases such as this one, one might argue as follows. Unless we are all picking out the same property with the expression "morally wrong," apparent disagreements about the moral wrongness of an action would be *merely* apparent. If you and I are using "morally wrong" to pick out different properties from one another, then trying to resolve a moral dispute between us would be as senseless as arguing about whether Tom *really* went to the bank, once we realize we are using the term "bank" in different ways. But surely it is not senseless to try to resolve moral disputes, even in cases of deep disagreement. So, one might conclude, instead of raising a pressing semantic challenge for moral realists, the phenomenon of moral disagreement actually gives us excellent reason to think that we are using our moral terms to pick out the same properties as one another.[87] For co-reference is precisely what makes genuine disagreement possible.

The argument of the preceding paragraph has been very popular.[88] I will argue, however, that the semantic worries raised by widespread moral disagreement are not so easily dispatched. My argument will involve several parts. First, I will offer a sketch of how deep divergence between different speakers' judgments involving the application of some term "T" *could* give us evidence that they are using "T" to pick out different properties.[89] In the course of offering this

---

[87] This assumes, of course, that "morally right" picks out a property in the first place. If one takes an expressivist view of moral language, one would likely deny that moral disagreement involves shared reference to properties, understanding such disagreements instead as disagreements in attitude. In what follows, however, I will set such views aside and follow the realist in assuming that moral terms pick out properties.

[88] See, for example, Brink (1989), Sturgeon (1994), Smith (1994), and Huemer (2005).

[89] The use/mention distinction will be important in what follows. Notation is slightly awkward because, of course, "T" is intended as a variable, rather than an actual term. Nonetheless, for purposes of notation, I will proceed as though "T" were an actual term in our language. Thus, "T" will appear in quotation marks when I am mentioning it, and without when I am using it.

sketch, I will make several distinctions in an attempt to clarify exactly where the challenge for the realist lies. It will emerge that the fundamental problem is an explanatory one: in the face of deep differences of moral belief on several levels, the realist must provide a plausible account of *how* it is that our moral terms could nonetheless co-refer.

Next, I will present a dilemma for the moral realist. The realist must provide either a semantically *internalist* or a semantically *externalist* account of how it is that our moral terms could refer to a single distinct set of properties. I will then argue that on the most prominent versions of either sort of view, it is very difficult to see how co-reference to a distinct set of moral properties could obtain in a range of paradigmatic moral disputes. I conclude that there are strong reasons to believe that in at least some paradigmatic moral disputes, different speakers fail to co-refer in their use of moral terms.

Third, and finally, I will respond to the objection that there is simply no credible account of *what we are doing* in paradigmatic moral disputes if our moral terms do not co-refer, because genuine disagreement *requires* co-reference. I will argue that even if we end up holding that some moral disputes involve parties "talking past one another" in one sense, such disputes need not be silly or senseless. On the contrary, I suggest that there are other plausible examples of worthwhile disputes in which speakers do not co-refer with their key terms. But let us first tackle some preliminaries, starting with a question already suggested above: given that disagreement seems to *require* co-reference, how could disagreement possibly give us evidence of its absence?

Preliminaries

To answer this question, we must first distinguish between two different possible conceptions of disagreement. One might understand disagreement as being essentially constituted by *mental states* or by *utterances.*[90] On the first conception, disagreement is thought of as a *state of affairs*: for two people to disagree is for them to hold attitudes that conflict in content in an appropriate way. Such disagreements may obtain between beliefs whose conjunction cannot be true (as when economists disagree about the causes of a recession) or between conative attitudes whose conjunction cannot be satisfied (as when we disagree about where to go for dinner). For a state of disagreement to obtain, it is not necessary that such conflicts in attitude ever be vocalized, or even that the parties be aware of them. For example, there are likely many disagreements between my beliefs and those of a typical 12th century European peasant, even if I am mostly ignorant about what such a person believed.

Alternatively, disagreement might be thought of as an *activity*: some people utter (or inscribe) sentences, and others respond with utterances (or inscriptions) that function as *rejections* of the original utterances. On this conception, two people are engaged in a disagreement whenever they display a familiar syndrome of linguistic behaviors. Paradigmatic instances would include utterances greeted sincerely with responses like "That's not the case" or "No, I don't think so." While both the state and the activity can be aptly described in ordinary language as "disagreements," it is useful to keep the two of them distinct. Following David Plunkett and Tim Sundell, let us henceforth use the term "dispute" to refer to the linguistic activity whereby people utter/inscribe certain sentences and others respond by rejecting them, and reserve the term "disagreement" to designate a state of conflicting attitudes (including, but

[90] On these two conceptions of disagreement, see Plunkett and Sundell (2013) pp. 10-11, who draw on MacFarlane (in progress) and Capellan and Hawthorne (2009) pp. 60-61.

not limited to, conflicting beliefs).[91]

With this distinction in hand, let us make the following observation: if a *dispute* turns out to have a certain character, this can sometimes provide excellent reason to believe that speakers are in fact picking out different things by their terms. Consider the following simple example. Suppose Charles walks into the philosophy lounge and hears Albert and Betty engrossed in conversation.

> **Albert** (to Betty)**:** Socrates really was a rather friendly fellow.
> **Charles** (butting in)**:** Friendly? I wouldn't say he was particularly friendly! Admirable, but not particularly friendly, in my judgment.
> **Betty:** Well, he always enjoyed interacting with people, even strangers.
> **Charles:** Yes, but those interactions usually involved telling people they were wrong!
> **Albert** (laughing)**:** Actually, as I recall, they involved him wagging his tail and panting.

At this point in the conversation, it becomes clear that Albert and Charles have rather different views about the individual they take themselves to be referring to with the word "Socrates." Charles clearly means to be referring to the philosopher, while Albert and Betty have presumably been talking about a particular four-legged creature. While the example may seem a bit contrived, I suspect that many of us have had this experience of mistaking the intended referent of a proper name. What's important to note here is that in cases like this, the content of what is asserted (e.g., that the individual in question wagged his tail) can provide us with evidence that we're not all talking about the same thing.

This phenomenon can occur not only with proper names, but with property or kind terms as well. Suppose the American Denise is talking with her British friend Everett about the items on offer at some restaurant.

> **Denise:** If you'll remember, the breakfast came with a biscuit.
> **Everett:** No, I don't believe it did**.** As I recall it came with a flaky, buttery roll.

---

[91] Plunkett and Sundell (2013), p. 10.

Denise: Yes, that buttery roll was a biscuit!
Everett (perplexed): But biscuits are dense and sweet...

The confusion here is caused by the fact that Denise and Everett fail to realize that "biscuit" picks out a different kind of baked good in American English than in British English. Let us note a few features of this example. At the outset of the conversation, Denise and Everett would each agree that they use the term "biscuit" to pick out a particular kind of baked good, and that the property of *being a biscuit* supervenes on "lower-level" descriptive properties about the appearance, texture, flavor, etc. of a baked good. What Denise and Everett would come to realize if they continued their conversation is that they (and their respective speech communities) each employ quite different standards concerning which combinations of subvening properties yield something that is aptly described as a "biscuit."

The existence of different dialects makes it easy to multiply cases of this sort—for example, we might consider the terms "chips" and "football"—where the cases in question all have the following form. We begin with a dispute about whether a term "T" truly applies to a particular object or activity—for example, whether "biscuit" aptly describes the baked good served with breakfast. Over the course of the conversation, it emerges that speakers accept very different standards specifying the conditions under which something counts as a T. These standards needn't be explicitly believed *analyses* or *definitions*—I myself would be hard pressed to give necessary and sufficient conditions for something being a biscuit—but would include a range of beliefs about what features *make* something a T (e.g., that biscuits are baked goods of a certain type). When a difference in standards is sufficiently large, and persists even after critical reflection, speakers acquire some reason to suspect that they are using the term "T" to pick out different properties or kinds from one another. In the case above, when Denise and Everett

realize the magnitude of the difference between their standards concerning what counts as a "biscuit," they acquire reason to believe that "biscuit" picks out a different kind of baked good in Denise's dialect than in Everett's. And indeed, that turns out to be the correct diagnosis in this case.

It would be premature, of course, to conclude that deep differences in standards concerning what counts as a T *entails* that speakers are not using "T" to pick out the same property. To take a classic counterexample to such a claim, consider the term "water." Suppose that my friend Empedocles believes that the term "water" refers to an element, whereas I think it refers to a compound. This is a quite significant disagreement concerning what it takes to qualify as "water"; indeed, it turns out that nothing which satisfies my conception can satisfy Empodocles's, and vice versa! Nonetheless, it's intuitively quite clear that we could both manage to use the term to refer to *water*, that is, to H_2O. If Empedocles and I were to have a dispute about whether there is a bottle of water in the fridge, this would most plausibly be interpreted as a genuine disagreement about *water*, rather than (as in the case of Denise and Everett) a case in which we're talking past one another. Furthermore, if there *is* a bottle of water in the fridge, then Empedocles can truly believe and assert that there is, even if he falsely believes that water is an element.

There are now familiar explanations of how co-reference is possible between speakers who have sharply different beliefs concerning what makes something a T.[92] As applied to the current tale, such stories would involve the fact that Empedocles and I (or previous speakers from whom we borrow the term "water") have had some kind of interaction with a certain clear and odorless liquid, and we use the term "water" with the intention of referring to that kind,

[92] See, for example, Putnam (1975) and Kripke (1980).

whatever its nature turns out to be. Notice, however, that on this sort of story, co-reference is made possible in part by significant agreement about what count as samples of water. If we didn't already agree on at least a range of paradigmatic cases about what counts as "water" and what does not, then investigation of the nature of the kind in question would be impossible. For consider the following case.

Suppose that I were to stumble upon an isolated community of scientifically untutored people who seem to speak English, and I find that they all use the word "water" to refer to the stuff in rivers, lakes, and streams. Excited to share my knowledge of chemistry with them, I declare to them, "Water is H_2O." I am surprised, however, when one of them gives the following response, "This hypothesis that water is H_2O would account for why the stuff in our rivers and streams count as "water." But it is totally deficient in other regards, for it fails to account for the fact that pineapple juice and cow milk are also water!" If others in the community were to find this response compelling, and if I were to find that it did not rest on misconceptions about the nature of these other liquids, this would give me reason to believe that the members of this community were using the word "water" to pick out a different property than typical English speakers. (Perhaps they use "water" to refer to any drinkable liquid.)

Let us draw some preliminary conclusions from these cases. Suppose we come across a dispute concerning the application of some term "T", where "T" seems to pick out a property or kind. What the cases above suggest is that dramatic differences in standards concerning what something has to be like in order to count as *T*, especially when accompanied by dramatic differences in belief about which things are (or are not) T, provide us with reason to believe that speakers are not using "T" to pick out the same property as one another. Of course, there will

always be *some* disagreement about virtually any interesting topic, so there will likely be some modest threshold below which divergences of belief do not provide any evidence of the absence of co-reference. Beyond this threshold, however, our reason for suspecting the absence of co-reference gets stronger as divergences of both kinds of relevant belief get wider. When there are extremely deep and intractable differences in judgment about both what *makes* something a *T*, and about which particular things possess *T*, we have strong reason to suspect that different parties fail to co-refer in their use of "T." In such cases it becomes incumbent upon those who think that we are all using "T" to pick out the same property to give an account of how this could be, in spite of our differences.

In what follows I argue for several things. First, I note that moral disputes between sincere speakers often involve quite radical differences in belief of both relevant sorts: not only do individuals have sharply divergent views about (for example) which actions are aptly described as "morally wrong," but they also have sharply different beliefs about what kinds of features could *make* something morally wrong. The magnitude of these differences puts significant pressure on the realist to give an account of how, in spite of them, such speakers may be picking out the same properties with their moral terms after all. The problem, I will suggest, is that the realist has no plausible story to tell about how this is possible. Furthermore, there are plausible accounts of such disputes that do not require that speakers co-refer with their moral terms. The upshot of this chapter is that we have very good reason to doubt that the realist's semantic diagnosis is correct for all paradigmatic moral disputes, and excellent reason to explore alternative semantic proposals.

Two Levels of Moral Dispute

Let us very briefly take note of just a few of the intractable moral disputes within contemporary American society. There is of course the broad range of particular issues taught in applied ethics courses: abortion, euthanasia, affirmative action, animal welfare, and the like. We might also consider the deep differences in belief concerning what we owe to each other more broadly, reflected in the range of moral perspectives from extreme libertarians who deny that we ever have positive duties to one another based solely on need, to utilitarian views demanding that all parties' interests count equally in determining our moral obligations, and the enormous number of views between these extremes.

In addition to these deep differences in moral *verdict,* there are also quite significant differences in belief about *moral standards,* or beliefs about what *makes* an action right or wrong. To take one salient example, an enormous number of ordinary people believe the Divine Command Theory. In more conservative parts of the United States, it is quite common for people to sincerely believe that if there is no God, there can be no morality, for there is simply *nothing that could make actions right or wrong* besides God's commands.[93] Imagine that a person who held such a view were to engage in moral discussion with a philosopher who is a moral realist of the naturalist, reductionist stripe:

> **Divine Command Theorist:** The death penalty is morally permissible. For God clearly permits it. And by the way, if I'm wrong and God does not exist, then nothing could possibly be right or wrong, for morality is created by God's commands.
> **Naturalist Realist:** The death penalty is morally wrong. For it possesses natural property N, which is identical to wrongness. And by the way, since wrongness just *is* property N, morality does not depend on God.

[93] Conservative theists are not the only ones who find claims of this form attractive. I once had a conversation with several economists who all agreed that if moral claims were anything other than claims about what would bring about the most good, they simply *didn't know what we're talking about* when we talk about morality. Jeremy Bentham makes a very similar claim in Bentham (1781) Ch. 1, as does G.E. Moore (1903) Ch. 1.

Clearly these two speakers do not merely disagree about *which* actions are morally wrong; they also disagree quite sharply about what sorts of features can make it the case that something counts as "morally wrong." The Divine Command Theorist believes that "morally wrong" refers to a property that can only be realized in virtue of God's commands, while the Naturalist Realist sees her own moral judgments as attributing a natural property. Others may see themselves as attributing the property of *being the sort of thing that an ideal observer would prohibit*, while still others may see themselves as attributing a non-natural *sui generis* property.[94]

When faced with such diversity both at the level of moral *verdicts* and at the level of (normative and metaethical) *standards,* we have two basic interpretive options. On the one hand, we could hold that there is a common subject matter that is fixed by a single set of properties—*the moral properties*—that all speakers are referring to. One consequence of this view would be that many parties to moral disputes are mistaken not merely about when such properties are instantiated, but about *the sort of features* that could make moral judgments and moral claims true. On the other hand, we might conclude that some speakers are using moral terms to pick out different properties from one another (while granting that there are likely many speakers who do co-refer with their moral terms). Those who opt for the first option acquire the burden of explaining how such co-reference is possible in the face of deep moral diversity. Those who opt for the second acquire the burden of giving an account of moral disputes that accounts for the seemingly genuine nature of moral disagreements, even across wide gulfs in moral belief. In the next three sections, I will argue that the burden of explaining how co-reference is possible proves very difficult for the realist. In the final section, I will argue that the burden of providing a

[94] Of course, for most people, these metaethical views are more or less inchoate. Nonetheless, in my experience, a number of ordinary people do have strong pre-theoretical metaethical leanings, and that these vary widely from person to person. For the purposes of this chapter, I set aside those who do not see themselves as using moral terms to attribute properties at all.

plausible account of worthwhile moral disputes in the absence of co-reference is not as heavy as one might initially think.

A Metasemantic Dilemma

Do all parties to paradigmatic moral disputes co-refer in their use of moral terms? This is a first-order semantic question about moral terms. Attempts to answer this question in a principled way, however, require us to delve into second-order questions about *metasemantics*. Metasemantic theories attempt to answer the question: in virtue of what do our words have the semantic features that they do?

While our moral terms plausibly have a number of semantic features, the feature that I focus on here is *reference.* While I shall have a fair bit to say about what it is *in virtue of which* a term refers to a property, in what follows I'll assume that we all have an intuitive grasp of this relation. To take a few simple examples: English speakers typically use the word "square" to refer to the property of being a four-sided equilateral polygon, but sometimes use it to refer to the property of being boringly conventional. We typically use the word "green" to pick out or refer to the color shared by grass and ripe peas, but sometimes use the same word to pick out the property of being naïve due to inexperience. Most of the time we refer to properties—and figure out which properties others are referring to—effortlessly and largely unconsciously, simply as a result of being competent speakers of a language. Nonetheless, questions about which properties our terms pick out are of great philosophical importance, in large part because the truth of our utterances will depend crucially on which properties are the ones we are attributing.

In providing an account of how moral co-reference is possible in deep and intractable

moral disputes, the realist must confront the following question: in virtue of what do our property terms pick out the properties that they do? There are two general types of answer to this question. *Internalist* views hold that the reference of the relevant terms is wholly fixed entirely by the mental states of speakers. The classic internalist view is *descriptivism*, according to which the reference of a term is fixed by some *description* that competent speakers associate with it. Semantic *externalist* views, in contrast, reject the view that the reference of our terms is fixed solely by the mental states of speakers. On such views, the reference of such terms is fixed in part by facts *external* to the speaker, such as facts about the causal interactions between speakers and their environment, or facts about the history of the term in question.

In what follows, I will develop a metasemantic dilemma for those who defend the view that all parties to paradigmatic moral disputes co-refer in their use of moral terms.[95] In attempting to give an account of how co-reference is possible despite deep differences in belief about moral verdicts and standards, invariantists must opt for either an *internalist* or *externalist* metasemantic theory of such terms. But on the most prominent versions of either type of metasemantic theory, it is very difficult to see how co-reference is possible. Therefore, I conclude, we have strong reason to doubt that all parties to paradigmatic moral disputes co-refer in their use of moral terms.

Some terminology from the philosophy of language will be useful in what follows. Assuming that speakers use moral terms to pick out properties, then each moral term will have an *extension* consisting of the set of actual things (e.g., actions, policies, etc.) which possess that property, and hence which can be truly described by the term in a speaker's idiolect. Of course,

[95] It is worth noting that this commitment has commonly been accepted not only by realists, but by constitutivists, ideal observer theorists, and some other constructivist views as well. To the extent that my arguments are successful, traditional versions of all such views will be undermined.

we often apply moral terms to actions and policies that are *not* actual, and these judgments are just as truth-apt as moral claims about actual actions and policies. Given this, we should also think of moral terms as having an *intension*, which for the purposes of this paper can be understood as a function from possible worlds to extensions. The intension of a term specifies which *possible* things could be truly described by that term in the speaker's idiolect.

Though there is some controversy about what conditions are *sufficient* to show that two terms pick out the same property, the following necessary condition seems unassailable: Term A picks out the same property as term B only if A and B have the same intensions. That is, if a possible action, intention, state, etc. could be truly described by term A, but not by term B, then term A and term B do not pick out the same property.[96] Therefore, if a metasemantic theory yields the verdict that a term "T" has different intensions when used by different speakers, then the theory has the consequence that the speakers do not use the term to co-refer to a single property. In what follows, I will try to show that given plausible assumptions, the most prominent metasemantic theories will sometimes assign different intensions to the same moral terms when used to sincerely express the moral thoughts of different linguistically competent agents. If this is so, the realist's contention that all speakers co-refer with their moral terms in paradigmatic moral disputes is mistaken.

The First Horn: Semantic Internalism

Suppose the realist tries to secure the co-reference of our moral terms by invoking some version

[96] Terms A and B may of course be orthographically identical, as will be the case if different speakers use "morally wrong" to pick out different properties than one another. The important issue is whether Term A, *as used by a particular speaker on a particular occasion*, has the same intension as Term B, *as used by a particular speaker on a particular occasion.* If they do not, then terms A and B are not being used to pick out the same property in the relevant context.

of semantic internalism. We can get a sense of the difficulties facing such an approach by beginning with a very simple version of internalism, according to which the reference of a moral term is fixed by whatever criteria a speaker actually employs when forming the judgments that she expresses by using that term. On such a theory, if a speaker regards all and only those actions that maximize happiness as being morally right, then the extension of "morally right" will simply be all and only those actions that maximize happiness. If a community of speakers all employ this criterion, they will co-refer in their use of "morally right." If this community were to encounter a different community, say, one in which people were disposed to regard certain actions as absolutely morally prohibited regardless of consequences, it would follow from the metasemantic theory under consideration here that they would fail to co-refer in their use of "morally right."

It is this sort of simple internalist theory that R.M. Hare seems to have in mind when he presents his famous case of the cannibals.[97] In Hare's case, a Christian missionary finds a community of cannibals who, like us, apply the word "good" to various actions as a term of commendation. The missionary notes, however, that the cannibals apply the word in ways that are systematically different from his own use of the term. Hare concludes that since the criteria being used by the missionary and the cannibals when judging something to be good are different, they cannot be using the term to pick out the same property as one another.[98]

Hare's case shows that realists run into problems if they allow the reference of our moral terms to be fixed too straightforwardly by the criteria we actually employ when using those terms. For it is plainly the case that speakers sometimes attend to quite different concrete

---

[97] Hare (1952) p. 148. See also Hare (1986).

[98] Hare then argues that this case supports an expressivist account of the meanings of moral terms, but I set this point aside here.

features of actions in deciding whether something counts as "morally wrong." Thus, if the reference of "morally wrong" were fixed in the way that the reference of "biscuit" is plausibly fixed in the case above—simply by consulting the criteria that speakers use when employing the term—then it would follow rather straightforwardly that many speakers fail to co-refer in their use of "morally wrong."

Fortunately for the realist, there are more sophisticated versions of internalist metasemantics on offer. The most promising internalist approach, developed by Frank Jackson, Philip Pettit, and Michael Smith (among others), is often called the "Canberra Plan," in homage to its place of origin.[99] According to this approach, the reference of moral terms is not fixed by speakers' beliefs about which concrete features of actions are morally relevant. Rather, Canberra planners hold that all competent speakers accept a large set of platitudes governing the use of moral terms. On their view, whatever property best satisfies the platitudes associated with a particular moral term is the property that the term picks out, provided that the property satisfies these platitudes to a sufficient degree.

Such a view, if it is to make good on the promise of securing co-reference, must meet several constraints. First, it must locate a number of platitudes accepted by all competent speakers who employ moral terms. Second, these platitudes must be *platitudinous;* that is, they must be such that their rejection would indicate not merely moral error, but that an agent is either linguistically incompetent, or else talking about something else besides morality. This is because, for the Canberra planner, not *every* moral belief is involved in fixing the reference of our moral terms—that would render the view vulnerable to Hare's case of the cannibals. Rather, the idea behind the Canberra plan is to locate a subset of moral beliefs—those that are

[99] See Smith (1994), Jackson and Pettit (1995), and Jackson (1998).

constitutive of competence with moral terms—and let these fix the reference of the terms in question. For the Canberra plan to succeed, the identified platitudes must be sufficiently rich to secure determinate reference for the terms in question.[100]

There is good reason to doubt that these constraints can be jointly met. First of all, note that in crafting any such list of platitudes, the Canberra planner must accommodate the following datum: with the exception of very young children and the severely mentally disabled, nearly every single human being makes moral judgments. Canberra planners aim to propose a list of platitudes, such that those who reject a significant number of them cannot be talking about morality at all. But in doing so, they run the following risk: if the list of platitudes that allegedly fix the reference of moral terms does not have nearly universal acceptance, then the Canberra planners themselves will have changed the subject: they are no longer talking about *moral judgment*, for "moral judgment" picks out a kind of psychological state that nearly all humans enter into with some regularity.[101]

Again, consider the baffling array of moral views that we find in the world. The world is full of people with moral views rarely considered (perhaps with good reason) by academic philosophers: devotees of Ayn Rand who believe that we never have fundamental moral reason to promote the well-being of others, devoutly religious individuals who believe that virtually nothing in the natural world has any meaningful value because all true value is to be found in the afterlife, environmentalists who believe that humans are morally obligated to undergo voluntary extinction for the good of nature, and so on. I do not mean to suggest that any of the foregoing

---

[100] This oversimplifies the matter slightly. On Jackson's version of the view, the relevant platitudes are not those that competent speakers *currently* accept, but those that they *would* accept after subjecting their views to rational reflection. I consider this variant of the view below.

[101] Of course, "moral judgment" can also be used to refer to a kind of *proposition*, but it is the mental state that is relevant for our purposes here.

positions are plausible moral views. But since each of these characters are plainly making moral judgments (rather than changing the subject), any alleged platitude concerning the content of morality must be consistent with the wide range of views they represent. These examples suggest that our practice of classifying people as making moral judgments incorporates what Folke Tersman has called "the latitude idea." This is the notion that, in Tersman's words, "we may attribute a specific moral conviction to a person, whether or not we share it, even in the absence of shared criteria and extensive overlap in basic values and norms, and even if it is based on quite different considerations than those we take to be relevant."[102] Only by accepting the latitude idea can we easily accommodate the datum that nearly all human beings make moral judgments. But if we embrace the latitude idea by insisting all content-related platitudes must be consistent with the diversity of moral views mentioned above, then it seems that content-related platitudes will be insufficient to secure any determinate reference for our moral terms at all.

Perhaps we can get closer to securing co-reference by incorporating platitudes concerning the *process* by which we form our moral judgments, either in actual fact or ideally? Consider, for example, the following platitude proposed by Michael Smith: "Whether or not ϕ-ing is right can be discovered by engaging in rational argument."[103] If we were looking for platitudes accepted by all accomplished moral philosophers, this would likely pass muster. But if we hold firmly in mind the desideratum that nearly all people make moral judgments, and that platitudes must be accepted as platitudinous by all relevant users of moral terms, the claim is considerably harder to defend. For one thing, as naturalist Nick Sturgeon notes, we have ample evidence that "for many appraisers moral claims are barely distinguishable from theological ones."[104] And a

[102] Tersman (2006), pp. 40-41.
[103] Smith (1994), pp. 39-40.
[104] Sturgeon, (1994), p. 106.

substantial subset of such appraisers might be inclined to direct us toward the book of Proverbs, which instructs that in practical matters one should "lean not on [one's] own understanding." Some religious believers think of moral knowledge as resting on *revelation* rather than reason, and even strongly believe that reason *will not* to lead people to converge on the moral truth. Instead, they believe, as one Christian philosopher once put the point to me, "Without revelation, we don't know our heads from our rears when it comes to morality." Again, the important point here is not whether such a view is ultimately defensible. The crucial point is that individuals who hold views like this are still correctly described as making moral judgments. Tersman's latitude idea can thus be extended beyond the *content* of moral judgments to the way in which they are formed, and subjects' beliefs about how they *ought* to be formed. And this renders any platitude in the ballpark of Smith's deeply suspect.

There is a range of platitudes of a third sort that do seem to be partially constitutive of linguistic competence when it comes to moral terms. These are platitudes stating relations between normative concepts. Here is one plausible candidate: *an action is morally permissible if and only if it is not morally wrong*.[105] If a speaker were to deny this, it would be reasonable to suppose that she was either conceptually confused or using moral terms in some idiosyncratic sense. Perhaps, the internalist might hope, a range of such platitudes could do significant work in fixing a determinate reference for our moral terms. But there are in fact good reasons to doubt that they will be able to do so. First, consider the toy case in which the previously mentioned platitude is *the only* relevant platitude. Plainly, this would not suffice to identify one definite intension as the correct one to assign to the term "morally wrong." It would, of course, rule out

[105] If moral indeterminacy is possible, this would need the qualifier: "*Except in cases of indeterminacy...*" I set this issue aside in what follows.

assigning intensions to "morally wrong" and "morally permissible" according to which some possible action falls under both terms. But there would remain an indefinite number of referential assignments which satisfy this constraint.

Of course, this is merely a toy case: there are quite plausibly many more platitudes that constitute competence with moral terms. We might add platitudes to the effect that "morally wrong" denotes that which we are *required* to refrain from doing, that morally wrong actions *deserve* condemnation, and so on, bringing other practical normative concepts into the fold. And we might even add some further platitudes the sort that Michael Huemer dubs "formal intuitions," such as: "If x is better than y, and y is better then z, then x is better than z," and "If it is wrong to do x, and it is wrong to do y, then it is wrong to do both x and y."[106]

If we confine ourselves to platitudes accepted by the vast range of speakers that we count as making moral judgments, however, it is doubtful that we will be able to assemble a list capable of determinately fixing the reference of our moral terms. At the most, it seems such platitudes will place a number of *holistic* constraints on candidate referential assignments, such as that all actions that count as "wrong" also count as "not permissible" and as "deserving condemnation." But a wide range of referential packages assigning referents to each of the relevant normative terms are likely to be consistent with the complete set of platitudes.

Similar considerations suffice to undermine a strategy introduced by David Brink. Brink proposes that the reference of moral terms is fixed by the intention of speakers "to use moral language to pick out those properties, whatever they are, that make objects of assessment interpersonally justifiable."[107] It is indeed plausible that, at least on one interpretation,

---

[106] Huemer (2008), p. 19.
[107] Brink (2001), p. 175.

competent moral speakers have such an intention. Just as there's a platitude stating the relationship between rightness and wrongness, there may well be a similar conceptual link between moral permissibility and interpersonal justification. The relevant question for our purposes is: exactly what property is picked out by the expression "interpersonally justifiable?" Note that it's not simply the notion of epistemic justification, but an essentially practical and social notion.

The problem, in short, is that the link between moral permissibility and interpersonal justifiability is simply too tight for Brink's purposes. We typically regard an action as capable of interpersonal justification if and only if it is not wrong. But the problem under consideration is that different speakers regard very different actions as morally wrong, and very different features as *making* actions morally wrong. Given the link between wrongness and interpersonal justifiability, we should expect similar divergences in judgment about which actions count as "interpersonally justifiable." So, even if we regard it as a platitude that morally permissible actions can be interpersonally justified, this platitude can no more fix a determinate reference for our moral terms than can the platitude that *an action is morally permissible if and only if it is not morally wrong*. For there are once again a wide range of packages assigning referents to the relevant terms, including "interpersonally justifiable," that respect the relevant platitudes.

What all of this suggests is that if the internalist hopes to rely solely on platitudes accepted by all linguistically competent users of the relevant terms, she will not have sufficient materials to fix any determinate reference for moral terms. Of course, this is not to deny that the internalist *could* provide an account of how moral terms could have a determinate reference: she could always revert to counting controversial claims as platitudes, such as Smith's principle that

moral truths are discoverable via reasoned argument, or more substantial claims about the content of morality. But this will yield the unacceptable result that a number of speakers with unconventional moral views are no longer making moral judgments at all. This result seems unacceptable in itself given the plausibility of the latitude idea, and at any rate, it concedes the point that parties to paradigmatic moral disputes sometimes fail to co-refer with their moral terms.

Given the actual dearth of agreed-upon platitudes, one might instead invoke platitudes that *would* be agreed upon in idealized conditions of reflection. Perhaps there is a set of claims that all parties to paradigmatic moral disputes *would* accept after suitable reflection, while also being strong enough to fix a determinate reference for our moral terms. This is the hope expressed by Frank Jackson. Jackson imagines a "mature folk morality," on which moral opinion would converge "after it has been exposed to debate and critical reflection."[108] If parties to moral disputes *would* converge on a sufficient range of moral propositions after engaging in careful moral reasoning, Jackson holds, *these* propositions can suffice to fix a determinate reference for moral terms, a reference that can also be attributed to current, non-idealized speakers' use of the terms.

There are two primary worries for such a strategy. First, as Jackson admits, it rests on a rather bold speculation about the character of existing moral disagreement, namely that none (or very little) of it is *fundamental* in the sense described in the previous chapter. If there *is* such disagreement, Jackson admits that on his view:

> [T]here will not be a single mature folk morality but rather different mature folk moralities for different groups in the community; and, to the extent that they differ, the adherents of the different mature folk moralities will mean something

[108] Jackson (1998), p. 133.

> different by the moral vocabulary because the moral terms of the adherents of the different schemes will be located in significantly different networks.[109]

So the first worry is that on Jackson's view, the possibility of moral co-reference rests on a questionable prediction about the character of existing moral disagreement.

But imagine that this worry can be met, and that no (or very little) existing disagreement is fundamental. There still seems to be a problem for Jackson's view. For, as I will discuss at greater length in the next section, it seems fairly easy to imagine possible agents who, although like us in most ways, have moral views that would *not* converge with ours even in the ideal limit of rational reflection. Perhaps, to borrow an example from Terry Horgan and Mark Timmons, they would all converge on some deontological moral theory, while we would converge on a consequentialist one. Nonetheless, it seems that if their judgments about which actions were (in their idiolect) morally right and wrong played the same practical role as ours in terms of regulating their attribution of praise, blame, and other reactive attitudes, we would naturally be willing to describe them as making moral judgments. This suggests that even if all *actual* agents pick out the same properties with their moral terms, there are *possible* cases in which all agents make what are recognizably moral judgments and engage in paradigmatic moral disputes, and yet would not co-refer with their use of moral terms. It seems to follow that reference to some particular set of properties—those we would call *the moral properties*—is not essential to making moral judgments and sincere moral assertions. This alone suggests that we should explore accounts of moral disputes that allow for lack of co-reference.

The internalist semantic views I've considered seek to find some set of mental states that is sufficiently general as to be shared by all competent moral speakers (either actually or under

---

[109] Jackson (1998), p. 137.

idealized conditions of reflection), but specific enough to fix a determinate reference for our moral terms. I've argued that it is very difficult to thread this needle. I haven't, of course, considered all possible candidate internalist views. But I hope to have shown that such views face difficulties severe enough to place the burden of proof squarely on the shoulders of the internalist to provide an account of moral co-reference in light of the diversity of moral belief.

The Second Horn: Semantic Externalism

In light of the difficulty of providing an account of how speakers' *beliefs* about morality could fix a common, determinate reference for our moral terms, one might be inclined to look to facts external to speakers to do the job. Semantic externalists adopt precisely this strategy, holding that the reference of our terms is not fixed exclusively by our mental states, but is partly determined by facts external to speakers, such as social, causal, and historical facts about the use of the terms in question. The most prominent attempts to develop an externalist account that accounts for (and explains) the co-reference of our moral terms have grown out of the work of Richard Boyd. Boyd attempts to take a causal theory of reference of the sort that many find attractive for natural kind terms and apply it to moral terms. According to Boyd's metasemantic theory, "Roughly, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k*."[110] On Boyd's view, moral terms refer to whatever properties *causally regulate* their use in the relevant way.

Causal regulatory semantics seems promising because it has the potential to secure co-reference even in cases of deep theoretical disagreement, provided that speakers are in

[110] Boyd (1988), p. 195.

appropriate causal contact with the relevant kind. The basic semantic worry that moral diversity presents for causal regulatory theories is that different individuals or communities might well have their use of moral terms causally regulated by different properties. As the theory is refined to confront this problem, it ends up facing problems very similar to those faced by the Canberra plan.

Begin with a toy version of the theory, according to which whatever property (or property cluster) is causally responsible for a speaker's use of the term "morally wrong" is the referent of that term. Plainly, such a view could not account for the co-reference of our moral terms, as people apply the label of "wrong" to actions on the basis of many different and incompatible lower-level properties. One natural response is to make the relevant sort of causal regulation *counterfactual*, rather than actual: moral terms refer to the properties that *would* causally regulate their use in the epistemically ideal conditions of full non-moral information and adequate reflection. As David Brink characterizes this version of the theory, "On this view, a natural property N causally regulates a speaker's use of moral term 'M' just in case his use of 'M' would be dependent on his belief that something is N, were his beliefs in dialectical equilibrium."[111]

Again, such a view rests the possibility of co-reference on the speculative hypothesis that none (or almost none) of the existing moral disagreement we find in the world is *fundamental*—that is, explained not by any non-moral ignorance or failure of epistemic rationality, but based on a systemic difference in moral intuitions.[112] To the extent that this hypothesis is doubtful, even counterfactual causal regulatory theories leave the possibility of co-reference hostage to empirical fortune. Furthermore, as mentioned in the previous section, even if no *actual* cases of

[111] Brink (2001), p. 169.
[112] See Chapter 2.

moral disagreement are fundamental, there are plainly possible cases that meet this description. The most discussed such cases in the context of causal regulatory semantics are the *Moral Twin Earth* cases developed by Terry Horgan and Mark Timmons.[113] Such cases imagine that we discover a group of people whose use of moral terms is causally regulated in the relevant respect by a different property than our own. In Horgan and Timmons' classic case, it turns out that our own use of moral terms is causally regulated by some consequentialist theory, while this other group's moral terms are regulated by a deontological theory. Nonetheless, utterances of moral sentences, and the judgments that these sentences express, play extremely similar functional roles among the two groups, in terms of guiding action and regulating praise, blame and other reactive attitudes.

If we were to come across such a group, it seems quite clear that a paradigmatic moral dispute could arise. And in such a case, it seems that our ordinary standards of attributing moral judgments would lead us to characterize members of the other group as making moral judgments. Horgan and Timmons argue that such cases provide a *reductio ad absurdum* of causal regulatory semantics for moral terms. I think this verdict is too strong, however. What such cases show is that in at least some possible (and, I think, very likely a number of actual) cases of paradigmatic moral disputes, causal regulatory semantics yields the verdict that speakers fail to co-refer with their moral terms. Horgan and Timmons assume that this result is unacceptable, as it fails to account for the intuitive verdict that the disputes in question are *genuine*. But, as I shall try to show in the next section, co-reference needn't obtain in order for disputes to worth having.

Still, *Moral Twin Earth* cases do seem to show that according to the most prominent

[113] See Horgan and Timmons (1991), (1992a), (1992b), and (1996).

externalist view in the literature, co-reference will fail to obtain in a range of possible (and likely actual) moral disputes. Unless some better version of externalism is in the offing, externalism doesn't seem to be a promising route for the realist to defend her commitment to co-reference. Other potential versions don't seem promising, however. Moral terms don't seem to trace their reference back to some initial baptism in the way that proper names plausibly do.[114] (Certainly speakers don't feel constrained in their moral judgments to respond only to those properties picked out by their remote ancestors who coined moral terms.) Furthermore, there seems to be no linguistic division of labor in ethics, of the sort we find in, say, botany or medicine. While speakers may commonly defer to acknowledged experts to fix the reference of terms such as"elm" or "arthritis," our moral practice contains no such commitment to deference, and at any rate uncontroversial moral experts are extremely hard to come by.[115] This makes *social externalism* of the sort defended (in a different context) by Tyler Burge an unpromising model for moral semantics.[116]

Finally, one might consider an externalist view of the sort developed in the work of Ruth Millikan, according to which reference is partially determined by a term's biological or evolutionary function.[117] Millikan's views are quite complex, and I can't hope to address them entirely adequately here. But I will note here that although it is fairly plausible that moral thought and language has an evolutionary function—perhaps, extremely roughly, that of coordinating and promoting social cooperation—it seems unlikely that any such function will solve the problem of fixing a determinate reference in many cases of fundamental moral disagreement. Again, consider a case in which one community finds that its use of moral terms

[114] See Kripke (1980)
[115] On the linguistic division of labor, see Putnam (1975).
[116] Burge (1979)
[117] See Millikan (1987), (1989), (2010)

is causally regulated (actually or ideally) by some consequentialist property, while another finds that's its own use is causally regulated by a deontological property. Given that both properties seem to be quite well-suited to fulfill any plausible function of morality (such as social coordination), it is very difficult to see how some such function could make it the case that, for example, members of the first group could all be referring to deontological properties in spite of their tendency to avow consequentialist theories in the limit of responsible moral inquiry.

Of course, I cannot hope to survey every possible externalist view. Still, I hope to have shown that the prospects of the most prominent externalist views for securing moral co-reference are rather dim. If my arguments have succeeded, they have placed the burden of proof on the externalist to provide an account of how speakers in paradigmatic moral disputes can co-refer with their moral terms even in the face of deep and even fundamental disagreement..

The Only Game In Town?

One might take all of the foregoing considerations into account, and admit that the moral realist has her work cut out for her in providing a metasemantic theory that allows for co-reference even in cases where moral beliefs differ sharply. But one might nonetheless insist that this can't really put significant pressure on the realist to think that we fail to co-refer. For if we did not co-refer, moral disputes would not be genuine or worth having; they would merely be cases in which we misunderstand one another and talk past one another. But moral disputes *are* worth having. Therefore, the argument goes, we must be co-referring with our moral terms, even if it may be difficult to explain *how* we manage to do so. I take this to be the most important objection to the arguments of this chapter, and I have two responses to it. The first involves an appeal to ordinary

moral thought and practice, while the second is more philosophical.

The first is merely to push back against the claim that it's *obvious,* or at any rate, simply a piece of common sense, that moral disputes involve a common set of properties to which we are all referring. As realists often note with dismay, skepticism about moral objectivity is ubiquitous in large segments of contemporary American society. For example, the intuitionist Michael Huemer relates the story of polling a class of forty undergraduates, and getting the result that *every single student* denied the claim that morality is objective.[118] Denial of moral objectivity is compatible with co-reference, of course, but if morality is a subjective matter, it wouldn't be surprising if I used moral terms to pick out *my* subjective standards, while you used them to pick out yours. And while many philosophers dismiss such an account out of hand, there is some evidence that many ordinary speakers think that this is exactly what is going on in moral disputes. Perhaps my favorite illustration of this comes from the moral philosopher Jonathan Dancy's appearance on *The Late Late Show* with Craig Ferguson in 2010. Consider the following exchanges:

> **Dancy:** "Moral philosophers are interested in... which actions are right and which actions are wrong and how they get to be so."
> **Ferguson:** "How do you define "right" and "wrong"?"
> **Dancy:** "You don't."
> **Feguson** (perplexed)**:** "You don't define right and wrong!?"

And a bit later:

> **Dancy:** "Some actions are right and some are wrong."
> **Ferguson:** "By whose definition?"
> **Dancy:** "We started out agreeing that we weren't going to do definitions."
> **Ferguson:** "Well that seems like a bit of a cheat, to be honest!"

My point in displaying this exchange is not to suggest that Ferguson's view—seemingly, that

[118] Huemer (2005), xxii.

moral claims can be true or false only relative to some agreed upon "definition" or explicit standard—is the correct one. Rather, it is to provide one illustration that this non-invariantist way of thinking about moral disputes is more common among ordinary folk than philosophers sometimes like to admit.[119] (I've found that the question "by whose definition?" is also popular among some students.) Ferguson and those who share his folk-metaethical proclivities might not balk at all at the proposal that different speakers use their moral terms to pick out different properties from one another (each in accordance with their own "definition"). We must not be careful to confuse common presuppositions among philosophers with widely shared "common sense."

Of course, even if denying co-reference turns out to be intuitive for many speakers, it might be that this particular segment of folk opinion is just deeply confused and philosophically untenable. This brings us back to the question of whether there is any defensible account of what we are doing in serious moral disputes if we are not co-referring with our moral terms. It would be philosophically unsatisfying and a move of last resort to say that such disputes are ultimately senseless and rest on simple linguistic confusion. Fortunately, we do not need to say this. In fact, such disputes might be perfectly sensible, even indispensable.

We have a strong collective interest in coordinating our use of language with one another. In many cases, it is quite sensible to put pressure on others to use their terms in the same way that we do, especially when things that matter to us deeply are at stake. And, as David Plunkett

---

[119] One might deny that Ferguson is denying invariantism by arguing that he is simply advocating for a version of ethical relativism. But there are two common ways of understanding such a view. One interpretation holds, roughly, that an agent's action is morally right if and only if (and because) the agent approves of the action. But on this interpretation, Ferguson's question about definitions would be rather strange; on this view, any particular action will be right or wrong *simpliciter*, rather than right or wrong *by someone's definition.* The other way to understand the relativist view is to view each speaker's use of "morally right" as roughly equivalent to "approved by me" or "approved by my culture." This view is not an invariantist view—the property of being approved by me is a different property from that of being approved by you—so the point in the text holds.

and Tim Sundell have shown at length in a recent paper, we need not conduct this coordinating activity by *mentioning* the terms in question.[120] Rather, we often coordinate the use of our terms *metalinguistically,* by *using* the terms in question as a means of negotiating their reference.

Consider two of Sundell and Plunkett's examples. First, they consider a case in which two people are involved in a dispute about whether a particular dish that they have both tasted is "spicy." We could, of course, suppose that there is some objective threshold at which something begins to count as "spicy." It seems far more plausible, however, to interpret the speakers as *negotiating* where this boundary will be placed for the practical purposes of the conversation. Such a negotiation is far from senseless or silly. As Sundell and Plunkett explain:

> [I]t is worth engaging in such a dispute because how we use words matters. For Oscar and Callie, as for many of us, an agreement amongst all the cooks in the kitchen that the chili can be described as "spicy" plays an important role in collective decision-making. In particular, it plays an important role in decision-making about whether to add more spice. This may have nothing at all to do with what is analytic about 'spicy'. Rather, it derives from sociological facts about how people in kitchens act when their creations earn that label. Why should Callie have to refrain from further seasoning the chili when it cannot even be described as "spicy"?[121]

A second example, which Plunkett and Sundell borrow from Peter Ludlow, is even more relevant for the purposes of this paper since it does not involve a gradable adjective. Ludlow describes hearing on the radio a heated argument concerning the greatest athletes of the 20th century. At issue was whether the racehorse Secretariat should be included on such a list. Following Plunkett and Sundell's simplification of the case, we can imagine a dispute in which one party utters "Secretariat is an athlete" and another responds "No, Secretariat is not an athlete." Again, we could insist that there is some prior fact of the matter as to whether the term

---

[120] Sundell and Plunkett (2013).
[121] *Ibid.*, p. 15.

"athlete" applies to non-human animals. But it is far more plausible in this case to see the disputants as *negotiating* precisely what property the disputed term will pick out.

I think is quite plausible upon reflection that many discussions involving the social world take this form. Consider, for instance, talk about romantic *love*. Clearly different speakers (especially across cultures and epochs) have extremely different views of what is involved in being "in love" with someone. Let us consider two speakers, Conservative Cal and Romantic Ron. Conservative Cal reflectively applies the expression "in love" only to couples who are deeply committed to one another, who are willing to make sacrifices for one another, and who plan to stay together indefinitely. Romantic Ron, in contrast, applies the expression "in love" primarily on the basis of *feelings*; according to Ron, someone counts as being "in love" just in case one has overwhelming feelings of romantic attraction to another individual, regardless of one's intentions or level of commitment.

Given the radical differences between their uses of the terms, there seems to be no good reason to deny that Cal and Ron pick out different relations with their use of the term "in love." At the same time, I don't think we'd be surprised to find Cal telling Ron that a love rooted in feelings alone is not *really* love. Even when they realize they are using the expression "in love" in different ways, Cal and Ron might well continue to engage in a dispute about, say, whether Ron is in love.[122] Given the lack of co-reference, however, this dispute should not be understood

[122] Perhaps this strikes the reader as counter-intuitive, so let me say a bit more. Consider, for starters, the odd fact that "What is love?" was the most searched query on Google in 2012. (http://www.itv.com/news/2012-12-11/web-users-search-for-meaning-of-love-online-in-2012/) This might suggest that despite the common knowledge that there are many different conceptions of love, many people are genuinely concerned with settling on a particular relation to denote with the word. Second, consider how common claims about "true love" are in our popular culture. Talk of "true love" suggests that many putative cases of love are mere pretenders. This doesn't *entail*, of course, that speakers often fail to co-refer with their use of the word "love." But I think that attention to this usage does suggest that many speakers would reject a agent's claim to be *truly* in love—even if the agent were reflective and informed of the underlying facts—if the agent did not meet the speaker's criteria for being in love.

as a disagreement over some distinct proposition that both might pick out with the sentence "Ron is in love." Rather, the disagreement at hand turns out to be about which kind of relation the speakers shall use the expression "in love" to pick out. Given the power that claims about love have in our culture, and the deference and respect shown to relations that we classify as "love," it is perhaps no surprise that speakers should care that their own use of the term, tailored as it presumably is to each speaker's own way of life, should be the one that should be adopted.

To take one final example, consider disputes concerning various political identities, as when people argue over whether a particular individual is a "real conservative" or a "real feminist." In some such cases, all parties may share relevant beliefs about what would *make* someone such as to be appropriately described by these labels. In many other cases, however, it is evident that conceptions of what is involved in being a conservative or being a feminist differ rather extremely from speaker to speaker. Yet given the social significance attached to terms of this sort, it can be quite sensible for speakers to advocate for their own conception of, say, *conservatism* to be the one picked out by the term.

There is, then, a plausible alternative to viewing paradigmatic moral disputes as the realist does, as involving co-reference to moral properties. Some moral disputes might be best seen as a case of *metalinguistic negotiation*—a sort of tacit bargaining over precisely which standards our moral terms will pick out. Granted, in many cases, speakers will likely co-refer with their moral terms. In such cases, the realist's preferred diagnosis will be the correct one: to say that an action is morally wrong is to say that the action violates the moral standards that all parties pick out with their use of moral terms. In other cases, however, speakers may be best understood as putting pressure on others to revise the reference of their moral terms to pick out

some alternative set of standards.

For a very rough model of how such pressure might work, consider the analogy of our talk about what is *disgusting*. If I know that you and I share similar tastes in food, I might straightforwardly convey information to you about the sort of food served at a restaurant by saying "the food there is disgusting." On the other hand, some uses of the word are not plausibly intended to simply convey information in this way, but function instead to put pressure on others to revise their standards concerning what they are willing to regard as disgusting. If a child gleefully eats a worm and a parent says "that's disgusting," this utterance is presumably not intended simply to inform the child that worm-eating is the sort of thing that meets the child's current standards for *being disgusting*. Rather, the parent can be thought of instead as *insisting* that the child revise her standards for what she regards as disgusting so as to include worm-eating. Such instruction is quite obviously worthwhile, regardless of where we come down on the semantic question of whether the word "disgusting" picks out the same property in the parent's idiolect as in the worm-eating child's. Moral instruction, and advocacy for one's moral perspective even in cases in which one suspects that disagreement might be fundamental, would seem no less practical.

Developing a complete account of precisely how exactly metasemantic pressure is applied (even in non-moral cases) is beyond the scope of this paper. But I hope to have motivated the idea that there is, at least, a promising account of what might be going on in moral disputes, such that these disputes could be quite worthwhile even in the absence of co-reference. Of course, many ordinary speakers might be inclined to reject the metasemantic diagnosis of what it is they are doing in moral disputes. But there are several reasons not to weigh this

rejection too heavily in our theorizing about moral disputes. First, note that in conversations about, say, whether some politician is a *true conservative*, it can be very difficult to determine precisely when we've veered from first-order discussion of whether the individual has certain relevant features to metalinguistic negotiation concerning which features precisely will be taken to constitute *conservatism*. Since most of the people we talk to share, to a large extent, our understanding of the relevant term, we might reasonably hope that the criteria we attach to such terms would converge, at least upon reflection. When this is not the case, agents might nonetheless continue to advocate that their own conception serve as the referent of the term in question. But the exact moment at which we switch from purely intellectual argument to advocating for our own referential assignments will often be far from clear.

This observation also applies to moral disputes. We know that lots of moral disagreement is not fundamental. In such cases, it is indeed plausible that speakers co-refer with their moral terms. Given the importance of our moral concerns to our lives and the role that moral judgments play in regulating behavior, however, it might be very sensible to continue advocating for one's own conception of *what morality requires* even if one begins to suspect that one's interlocutor has a fundamentally different conception. In practice, though, the line between first-order moral discussion and metalinguistic negotiation might well be very difficult to identify. This might explain some of the reluctance of speakers to accept the diagnosis that they are engaged in metalinguistic negotiation.

But there might also be pragmatic reasons that speakers find themselves inclined to reject the diagnosis of metalinguistic negotiation. Since one primary purpose of our conversations about morality is to sway others in their moral views, and since admitting that one's opponent is

speaking truly is a dialectically ineffective way of doing so, we might expect speakers to be reluctant to make such an admission.[123] Of course, the causes of this reluctance might not always be consciously accessible to speakers; it seems likely that many speakers have learned how to engage in effective moral discussion without engaging in deep questions in philosophy of language about what precisely they are doing. But it is nonetheless quite possible that such considerations of dialectical effectiveness are what are driving speaker's intuitions in this case.

The objection that only co-reference can account for the worthwhile nature of moral disputes therefore fails. Given the importance of moral concerns to our lives, and the deep interest we have in coordinating our use of moral language with others, engaging in metalinguistic negotiation concerning the reference of our moral terms could be quite sensible. I don't pretend to have shown that this is *in fact* what we are doing in moral disputes; rather, I have aimed to undermine the contention that co-reference is the "only game in town" when making sense of moral disputes.

Conclusion

Realists think that in paradigmatic moral disputes, all speakers co-refer with their moral terms. I've suggested that the diversity of moral judgments, both at the level of moral verdicts and moral standards, puts pressure on the realist to give an account of how this co-reference might occur. But both internalist and externalist attempts to account for moral co-reference across all such cases run into serious problems. Furthermore, I've tried to show that there is an alternative account of what we are doing in some serious moral disputes, an account which does not require co-reference but nonetheless makes good sense of our intuition that such disputes are

[123] For a similar point, see Plunkett and Sundell (2013), p. 24.

worthwhile. If I'm right about this, semantic considerations give us some reason to abandon invariantism about moral terms, and thereby abandon moral realism.

Bibliography

Alston, William P. (1988). "An Internalist Externalism," *Synthese*, 74: 265–283.

Anscombe, G.E.M. (1957). *Intention*, Oxford: Basil Blackwell.

Ayer, A.J. (1952). *Language, Truth, and Logic.* 2nd Edition. New York: Dover.

Bedke, Matt. (2009). "Intuitive Non-Naturalism Meets Cosmic Coincidence," *Pacific Philosophical Quarterly,* 90: 188-209.

Bentham, Jeremy. (1988). *The Principles of Morals and Legislation*. Amherst, NY: Prometheus Books.

Blackburn, Simon. (1984). *Spreading the Word*, New York: Oxford University Press.

Blackburn, Simon. (1993). *Essays in Quasi-Realism.* Oxford: Oxford University Press.

Blackburn, Simon. (1998). *Ruling Passions.* Oxford: Clarendon Press.

Boyd, Richard. (1988). "How to Be a Moral Realist," in *Essays on Moral Realism*, ed. Geoffrey Sayre-McCord, Ithaca: Cornell University Press.

Brink, David O. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Brink, David. (2001). "Realism, Naturalism, and Moral Semantics." *Social Philosophy and Policy.* 18(2): 154-176.

Brosnan, Kevin. (2011). "Do the Evolutionary Origins of Our Moral Beliefs Undermine Moral Knowledge?" *Biology and Philosophy.* 26: pp. 51-64.

Burge, Tyler. (1979). "Individualism and the Mental," in French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, pp. 73–121.

Cappelen, Herman, and John Hawthorne. (2009). *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Chalmers, David and David Bourget. (2013) "What Do Philosophers Believe?" *Philosophical Studies*: 1-36.

Christensen, David. (2007). "Epistemology of Disagreement: The Good News." *Philosophical Review.* 116(2): 187-217.

Conee, Earl and Feldman, Richard (1998). "The Generality Problem for Reliabilism," *Philosophical Studies*, 89: 1–29.

Copp, David. (2008). "Darwinian Skepticism About Moral Realism." *Philosophical Issues*. 18: 186-206.

Crisp, Roger. (2012). "Reasonable Disagreement," in *The New Intuitionism*, ed. Jill Graper Hernandez. New York: Continuum.

Daniels, Norman. (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy.* 76 : 256–82

Darwin, Charles. (2004). *The Descent of Man.* London: Penguin Classics.

DePaul, Michael. (1993). *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry.* London: Routledge,

De Waal, Frans. (1997). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.

De Waal, Frans. (2006). *Primates and Philosophers: How Morality Evolved*. Princeton: Princeton University Press.

Doris, John and Alexandra Plakias, (2008). "How to Argue about Disagreement," in Walter Sinnott-Armstrong, ed., *Moral Psychology,* Volume 2, Cambridge, MA: Bradford.

Dreier, James. (2004). "Metaethics and the Problem of Creeping Minimalism." *Philosophical Perspectives*. 18: 23-44.

Dworkin, Ronald. (1996). "Objectivity and Truth: You'd Better Believe It." *Philosophy and Public Affairs*. 25(2): 87-139.

Elga, Adam. (2007). "Reflection and Disagreement," *Noûs.* 41(3): 478-502.

Elga, Adam. (2010). "How to Disagree About How to Disagree" in Richard Feldman and Ted Warfield, eds., *Disagreement*, Oxford: Oxford University Press.

Enoch, David. (2010). "The Epistemological Challenge to Metanormative Realism." *Philosophical Studies*. 148 : 413-438.

Enoch, David. (2011a). "Not Just a Truthometer: Taking Oneself Seriously (but not too seriously) in Cases of Peer Disagreement." *Mind* 119: 953-997.

Enoch, David. (2011b). *Taking Morality Seriously.* Oxford: Oxford University Press.

Feldman, Richard (1985). "Reliability and Justification," *Monist*, 68: 159–174.

Feldman, Richard and Ted Warfield, eds. (2010). *Disagreement*. Oxford: Oxford University Press.

Field, Hartry. (1986). "The Deflationary Conception of Truth." in G. MacDonald and C. Wright (eds.), *Fact, Science and Morality*, Oxford: Blackwell.

FitzPatrick, William. (2008) "Morality and Evolutionary Biology", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2008/entries/morality-biology/>

FitzPatrick, William. (2014). "Why There is No Darwinian Dilemma For Ethical Realism," in Michael Bergmann and Patrick Kain, eds., *Challenges to Religious and Moral Belief from Evolution and Disagreement.* Oxford: Oxford University Press.

FitzPatrick, William. (Forthcoming) "Debunking Evolutionary Debunking of Ethical Realism," *Philosophical Studies*.

Fumerton, R. (2010). "You Can't Trust a Philosopher," in Feldman and Warfield, eds. *Disagreement*. Oxford: Oxford University Press.

Goldman, Alvin I. (1979). "What Is Justified Belief?" in G. Pappas (ed.), *Justification and Knowledge*, Dordrecht: Reidel.

Goodman, Nelson. (1983). *Fact, Fiction, and Forecast*. 4th Edition. Cambridge, MA: Harvard University Press.

Haidt, Jonathan. (2012). *The Righteous Mind: Why Good People are Divided By Politics and Religion*. New York: Vintage Books.

Haidt, Jonathan, & Craig Joseph. (2007). "The Moral Mind: How Five Sets of Innate Moral Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules." In *The Innate Mind, Vol. 3,* Ed. Peter Carruthers, Stephen Laurence, and Stephen Stich. Oxford: Oxford University Press.

Hare, R.M. (1991). *The Language of Morals*. Oxford: Oxford University Press.

Hare, R.M. (1986). "A Reductio Ad Absurdum of Descriptivism." in *Philosophy in Britain Today.* ed. S.G. Shankar, Albany: State University of New York Press.

Harman, Gilbert. (1977). *The Nature of Morality.* Oxford: Oxford University Press.

Hauser, Mark. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* New York: Ecco Press.

Horgan, Terence and Mark Timmons. (1991). "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16: 447–65.

Horgan, Terence and Mark Timmons. (1992a) "Troubles for New Wave Moral Semantics: The Open Question Argument Revisited." *Philosophical Papers* 21(3): 153–75.

Horgan, Terence and Mark Timmons. (1992b). "Troubles on Moral Twin Earth: Moral Queerness Revived." Synthese 92 : 221–60.

Horgan, Terence and Mark Timmons. (1996). "From Moral Realism to Moral Relativism in One Easy Step." *Critica* 28: 3–39.

Horgan, Terence and Mark Timmons. (2000). "Copping Out on Moral Twin Earth." *Synthese.* 124: 139–52.

Horgan, Terence and Mark Timmons. (2009.) "Analytical Moral Functionalism Meets Moral Twin Earth." in *Minds, Ethics and Conditionals: Themes from the Philosophy of Frank Jackson*. Ed. Ian Ravenscroft. Oxford: Oxford University Press, pp. 221–36.

Horwich, Paul, (1990). *Truth.* Oxford: Blackwell.

Huemer, Michael. (2005). *Ethical Intuitionism.* New York: Palgrave McMillan,.

Huemer, Michael. (2008). "Revisionary Intuitionism." *Social Philosophy and Policy.* 25: 368-392.

Jackson, Frank. (1998). *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson, Frank and Phillip Pettit. (1995). "Moral Functionalism and Moral Motivation." *Philosophical Quarterly.* 45(178): 20-40.

Joyce, Richard. (2001). *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, Richard. (2007). *The Evolution of Morality*. Cambridge, MA: MIT Press.

Kelly, Thomas. (2005). "The Epistemic Significance of Disagreement," in Tamar Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, Vol. 1. Oxford: Oxford University Press.

Kelly, Thomas. (2010). "Peer Disagreement and Higher-Order Evidence," in Richard Feldman

and Ted Warfield, eds., *Disagreement*, Oxford: Oxford University Press.

Kelly, Thomas and McGrath, Sarah. (2010). “Is Reflective Equilibrium Enough?” *Philosophical Perspectives*. 24 : 325–359.

Kitcher, Philip. (2005). "Biology and Ethics." In *Oxford Handbook of Ethical Theory,* ed. David Copp. New York: Oxford University Press,

Kitcher, Philip. (2006). "Four Ways of "Biologicizing" Ethics." In *Conceptual Issues in Evolutionary Biology*, 3rd Edition. Ed. Elliott Sober. New York: MIT Press,

Kitcher, Philip. (2011). *The Ethical Project*, Cambridge, MA: Harvard University Press.

Kornblith, H. (2010). “Belief in the Face of Controversy,” in in Richard Feldman and Ted Warfield, eds., *Disagreement*, Oxford: Oxford University Press.

Kripke, Saul. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Leiter, Brian. (2014). “Moral Skepticism and Moral Disagreement in Nietzsche,” in *Oxford Studies in Metaethics*, Vol. 9, ed. Russ Shafer-Landau. Oxford: Oxford University Press.

Locke, Dustin. (2014). “Darwinian Normative Skepticism.” in Michael Bergmann and Patrick Kain, eds., *Challenges to Religious and Moral Belief from Evolution and Disagreement.* Oxford: Oxford University Press.

Loeb, Don. (1998). “Moral Realism and the Argument from Disagreement,” *Philosophical Studies*, 90: 281-303.

MacFarlane, John. (In Progress). *Assessment Sensitivity: Relative Truth and its Applications.*

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin Books.

McGrath, Sarah. (2008). “Moral Disagreement and Moral Expertise,” *in Oxford Studies in Metaethics*, Vol. 3, ed. Russ Shafer-Landau, Oxford: Oxford University Press.

McPherson, Tristram. (2013). “Semantic Challenges to Normative Realism.” *Philosophy Compass* 8(2): 126–136

Millikan, Ruth. (1987). *Langauge, Thought, and Other Biological Categories*. Cambridge: MIT Press.

Millikan, Ruth. (1989). “Biosemantics.” *Journal of Philosophy.* 86: 281-297.

Millikan, Ruth. (2010). “On Knowing the Meaning; With a Coda on Swampman.” *Mind*

119: 43-81.

Moore, G.E. (2004) *Principia Ethica*. New York: Dover.

Parfit, Derek. (2011). *On What Matters.* 2 Volumes. Oxford: Oxford University Press.

Platts, Mark. (1997). *Ways of Meaning*. 2nd Edition. Cambridge, MA: MIT Press.

Plunkett, David and Tim Sundell. (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosopher's Imprint*. 13(23): 1-37.

Putnam, Hilary. (1973). "Meaning and Reference." *Journal of Philosophy* 70(19): 699–711.

Putnam, Hilary. (1975). "The Meaning of 'Meaning'." In *Mind, Language and Reality: Philosophical Papers*, *Volume 2*. Cambridge: Cambridge University Press: 215–71.

Rawls, John. (1971). *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Sayre-McCord, Geoffrey. (1986). "The Many Moral Realisms." *Southern Journal of Philosophy,* 24: 1-22.

Sayre-McCord, Geoffrey. (1997). "'Good' on Twin Earth."*Philosophical Issues* 8: 267–92.

Schafer, Karl. (2010). "Evolution and Normative Skepticism." *Australasian Journal of Philosophy* 88(3): 471–488.

Schroeder, Mark. (2013). "Moral Semantics." In *International Encyclopedia of Ethics*, ed. Hugh LaFollette. Hoboken, NJ: Wiley-Blackwell.

Schroeter, Laura and Francois Schroeter. (2013). "Normative Realism: Co-reference Without Convergence?" *Philosophers' Imprint.* 13(13): 1-24.

Shafer-Landau, Russ. (1994). "Ethical Disagreement, Ethical Objectivism, and Moral Indeterminacy." *Philosophy and Phenomenological Research* 54: 331-344.

Shafer-Landau, Russ. (2005). *Moral Realism: A Defence*. Oxford: Oxford University Press.

Shafer-Landau, Russ. (2012). "Evolutionary Debunking, Moral Realism, and Moral Knowledge" *Journal of Ethics and Social Philosophy* 7(1): 1-38.

Sidgwick, Henry. (1981). *Methods of Ethics*. 7th edition. Indianapolis: Hackett Publishing.

Sinnott-Armstrong, Walter. (2010). "Framing Moral Intuitions," in Walter Sinnott-Armstrong, ed., *Moral Psychology,* Volume 2, Cambridge, MA: Bradford.

Skarsaune, Knut Olav. (2011). “Darwin and Moral Realism: Survival of the Iffiest.” *Philosophical Studies.* 152: 229-243.

Smith, Michael. (1994). *The Moral Problem.* Malden, MA: Blackwell.

Sober, Elliott. (1994). "Prospects For an Evolutionary Ethics." In *From A Biological Point of View: Essays in Evolutionary Philosophy.* Cambridge: Cambridge University Press.

Stevenson, C.L. (1937). “The Emotive Meaning of Ethical Terms,” *Mind*, 46: 14–31.

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

Stevenson, C.L. (1963). “The Nature of Ethical Disagreement.” In *Facts and Values: Studies in Ethical Analysis,* New Haven: Yale University Press.

Street, Sharon. (2006). "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127: 109-66.

Street, Sharon. (2008). “Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About.” *Philosophical Issues*. 18: 207-228.

Sturgeon, Nick. (1994). “Moral Disagreement and Moral Relativism.” *Social Philosophy and Policy.* 11(1): 80-115.

Swain, Marshall (1981). *Reasons and Knowledge*, Ithaca, NY: Cornell University Press.

Tersman, Folke. (2006). *Moral Disagreement*. Cambridge: Cambridge University Press.

Thomson, Judith Jarvis. (1996). “Moral Objectivity,” in Gilbert Harman and Judith Jarvis Thomson, *Moral Relativism and Moral Objectivity*, New York: Blackwell.

Vavova, Katia. (Forthcoming) “Debunking Evolutionary Debunking.” In *Oxford Studies in Metaethics*, Volume 10. ed. Russ Shafer-Landau. Oxford: Oxford University Press.

Weatherson, Brian. (2012). “Disagreements, Philosophical and Otherwise,” in *The Epistemology of Disagreement*, eds. David Christensen and Jennifer Lackey, Oxford: Oxford University Press: 54-75.

Wedgwood, Ralph. (2010). “The Moral Evil Demons,” in Richard Feldman and Ted Warfield, eds., *Disagreement*, Oxford: Oxford University Press.

Wielenberg, Eric. (2010). “On the Evolutionary Debunking of Morality.” *Ethics* 120: 441-464.