MORAL DISAGREEMENT AND SHARED MEANING

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By
David Allen Merli, MA

The Ohio State University
2003

Dissertation Committee:
Professor Justin D’Arms, Adviser
Professor Don Hubin
Professor Dan Farrell

Approved by

Adviser
Department of Philosophy
ABSTRACT

In order to have genuine disagreement, interlocutors must share terms, meanings, and concepts. Without this, their dispute is merely verbal; it rests on linguistic confusion. This is true of all conversation, but many philosophers have thought that the univocity of moral discourse poses special problems. Disputes seem intractable, disputants have radically different standards for applying moral terms, and communities lack the sorts of authority and deference relations that are typical in straightforward empirical disagreement. These and other features of our moralizing yield a potent philosophical puzzle: how is it that moral evaluators can share a subject matter while thinking such different things?

Traditionally, the need to ensure the univocity of moral conversation has provided one of the central motivations for noncognitivist theories of moral discourse. If, as noncognitivists suggest, moral utterances serve to evince sentiment or prescribe action, then substantive differences in normative commitment between speakers do not undermine their disagreement. We can clash in attitude despite our different standards. On the other hand, if moral terms express real properties, it’s hard to see how diverse speakers are connected, in the relevant ways, to the same properties.

I examine noncognitivist attempts to make sense of disagreement, and I argue that they fail. The noncognitivist (or expressivist) is obliged to provide an account of the
mental states at work in moral discourse. Either this account fails to identify a distinct species of *moral* evaluation, or it cannot provide for genuine incompatibility between competing moral judgments, or it collapses into circularity. Thus, one of the most important motivations for noncognitivist accounts is undermined.

I show how naturalistic moral realism can be defended from popular arguments that challenge its ability to make sense of univocity. This criticism, most famously associated with R.M. Hare, has been revived in recent work by Terence Horgan and Mark Timmons. I develop three objections to their so-called ‘Moral Twin Earth’ argument and conclude that it has no force against moral realism.

I then show that naturalistic realism faces a *different* problem accounting for univocity. The semantic resources that feature prominently in contemporary realist accounts are unable to save shared meaning. This is because the path of moral inquiry is underdetermined: there is no fact of the matter about the referents of speakers’ terms. I argue that common realistic appeals to the resolution of moral dispute are not sufficient, because they fail to note a distinction between different readings of the convergence claim. The most plausible ways of understanding that claim are of no help to the realist’s semantic requirements.

Finally, I consider a rejoinder suggested by recent work in the philosophy of language. Though not compatible with realism’s moral semantics, this rejoinder suggests that moral and non-moral language are on a par. I offer some reasons for doubting this claim and suggest that moralizing poses unique interpretive challenges.
Dedicated to my mother and the memory of my father
I thank my adviser, Justin D’Arms. I owe him a great intellectual debt for his philosophical acumen and his encouragement. He has provided an exemplary model of how to do philosophy, and this project has improved immeasurably under his guidance. I thank Don Hubin. Since the very beginning of my graduate career, Don has provided philosophical insight, sharp criticism, and good cheer. I thank Dan Farrell. Dan is a superb philosophical interlocutor, and I have benefited from his deep understanding of moral theory and his kindness in offering it. I thank William Taschek for countless discussions of all sorts of philosophical issues, for healthy skepticism, and for good-natured chiding.

I owe a special debt to Nick Sturgeon. Nick’s enormous generosity, both personal and professional, has been a great help over the last few years. I am grateful for this, and for his lucid, insightful comments and discussions.

I am grateful for the friendship and encouragement of many of my fellow graduate students, particularly Andrew Arlig, Jack Arnold, Sondra Bacharach, Emily Beck, Zac Cogley, Julian Cole, Roy Cook, David Eng, Lee Franklin, Rick Groshong, Carol Hay, Eric Heining, John Jakala, Nick Jones, Jose Martin, William Melanson, Ryan Nichols, Brian Stone, and Cathal Woods. I am particularly indebted to Jon Cogburn and Tyler Hower for their intellectual and moral support. I thank Pamela Hieronymi and
Julie Tannenbaum for their kindness and for their friendly disdain toward my philosophical outlook.

I thank Amy Ball, Eben Friedman, Bill and Traci Lepicki, David Levy, Maureen Neary, and Joe Pirone, for their encouragement and understanding.

I am grateful to my early teachers, particularly Bill and Stacey Edgar, Dennis Bradford, and Walt Soffer. I am particularly indebted to George Mickus, who first inspired me to study philosophy.

Finally, I thank Kelly Ball for her support and tolerance. Without her involvement, this project might have been finished earlier, or not at all.
VITA

December 22, 1972………………………Born—Queens, New York

1994……………………………………..BA, Philosophy and English, SUNY-Geneseo

1999………………………………………MA, Philosophy, The Ohio State University

1997-2003………………………………Graduate Teaching Associate,
The Ohio State University

2003-present…………………………….Instructor, Franklin and Marshall College

PUBLICATIONS


FIELDS OF STUDY

Major Field: Philosophy
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>The Failure of Expressivism</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.1 Two arguments involving disagreement</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.2 Ogden and Richards</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2.3 Duncan-Jones</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2.4 Barnes</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2.4.1 Barnes’ second attempt</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2.5 Ayer</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2.6 Stevenson</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>2.7 Hare</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>2.8 Taking Stock</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>2.9 Gibbard</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>2.10 Blackburn</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>2.11 Some Conclusions</td>
<td>128</td>
</tr>
<tr>
<td>3.</td>
<td>Return to Moral Twin Earth</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>3.2 The Moral Twin Earth Argument</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>3.3 Copp’s First Response</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td>3.4 Copp’s Second Response</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>3.5 Disagreement, Here and Elsewhere</td>
<td>147</td>
</tr>
</tbody>
</table>

viii
3.6 The Second Argument: Revising Our Intuitions ..................................................162
3.7 Preserving (Nonmoral) Disagreement ...............................................................170
3.8 Concluding Remarks .......................................................................................182

4. Moral Realism’s Semantic Problem .....................................................................184
   4.1 Introduction .....................................................................................................184
   4.2 How Moral Discourse is Distinct: Three Disanalogies .....................................185
   4.3 The Traditional Response ..............................................................................192
   4.4 The Traditional Response is Not Adequate ....................................................194
   4.5 The Convergence Claim, and How it Helps ....................................................203
   4.6 Agreement on Cases and Theories ...................................................................207
   4.7 Responses to Recalcitrant Theorists ...............................................................211
   4.8 Ambitious Realism: Arguments Inside and Outside Moral Theory ...............222
   4.9 Two Kinds of Convergence .............................................................................230
   4.10 Some Considerations in Favor of Pessimistic Convergence ..........................234

5. Continuity and Innocence by Association .........................................................240
   5.1 Introduction .....................................................................................................240
   5.2 What’s Wrong With Expressivism and Realism? ............................................240
   5.3 A Reaction .......................................................................................................245
   5.4 The Detour .....................................................................................................248
   5.5 The Significance of Underdetermination ..........................................................252
   5.6 A Minimalist Proposal ....................................................................................258
   5.7 Against Continuity ..........................................................................................261
   5.8 Conclusions .....................................................................................................270

Bibliography .............................................................................................................271
Moral agreement and disagreement require a shared subject matter. If we’re to avoid talking past one another, we have to share terms, meanings, and concepts. Without this fulcrum of disagreement, our disputes are spurious, or ‘merely verbal.’ This is true of other sorts of conversations, too, of course, but, for a variety of reasons, moral talk has seemed especially problematic. Even if we could make sense of differences in belief and theory when it comes to non-normative conversation, many philosophers have thought, it’s harder to do so in the moral case. And, of course, if we can’t, our moral discussions will turn out to be confused, as they will rest on linguistic confusion instead of the stable foundation of shared meaning.

This dissertation concerns univocal moral argument. More narrowly, it focuses on two competing meta-ethical views and their attempts to preserve the possibility of genuine disagreement. These are noncognitivism and naturalistic realism. According to

---

1 I take the term ‘fulcrum of disagreement’ from Ronald Dworkin (1986: ch. 1).
the noncognitivist, moral claims do not express beliefs; instead, they express certain noncognitive mental states, such as sentiments, attitudes, or preferences. On this view, our moral talk doesn’t describe—or attempt to describe—some independent moral reality. According to naturalistic realism, on the other hand, our moral sentences work just the way they appear to work. They express our moral beliefs, and they make truth-evaluable claims about moral facts. What’s more, they often get things right.²

Traditionally, worries about univocity have seemed to favor noncognitivism over realism and its predecessors. Here, I argue that this widespread view is false. I also argue, however, that realism’s attempts to preserve shared meaning are also unsuccessful. The conclusion, then, is negative: neither research project has the resources to make good sense of moral argument. This leaves us with a philosophical puzzle, but it also clarifies some important misunderstandings about how univocity might be saved. In doing so, it highlights the importance of several significant features of moral discourse and sheds some light on how we might begin to make sense of them.

Before turning to a more detailed discussion of the arguments for these conclusions, we should look at why the problem of shared meaning is so pressing in the moral case and why noncognitivist proposals seem to have an advantage in solving it. One striking fact about moral conversation is its diversity: speakers have all sorts of views about what actions are right and about what makes them right. Sophisticated moralists invoke claims about utility, practical rationality, virtue, God’s commands, and other considerations, in order to bolster their claims, and speakers without explicit

² This qualification is important since we might think, as the error-theorist does, that our moral claims systematically misdescribe the world. According to this line of thought, our sentences function in a straightforward way, but the world simply fails to live up to our presuppositions. Mackie (1977) is the most famous defense of this view.
commitments to normative theories offer all sorts of competing justifications for their moral claims. When we argue about moral issues, it looks as though you’re asserting one fact while I’m denying another. Equivocation threatens.

Suppose, though, as the noncognitivist suggests, that, in calling this action right, I’m endorsing it, or saying hurray for it, or commanding you to do it, while, in calling it wrong, you’re condemning its performance, saying boo for it, or commanding that it not be done. If so, then it becomes less important that we’ve got different standards at work behind our moral judgments, since the incompatibility of our claims can be understood in the clash of sentiments or commands. Our different theoretical commitments don’t stand in the way of real disagreement.

Likewise, we can see how naturalistic realism, which claims that moral terms express (natural) properties, will have a harder time making sense of this phenomenon. Given that we use such different standards of evaluation, it’s tempting to think that, say, “right” in your mouth and in mine simply expresses different properties. This leaves us unable to vindicate our intuition that we have genuine disagreement. The realist, in response, must show that our disagreements are compatible with shared concepts, which, given the substantive difference between our views, looks like a daunting task.

This is just an initial sketch of the problem, but it’s enough to show why the noncognitivist tradition is thought to have an edge when it comes to the preservation of univocity. Since the problem arises from substantive differences between various moral speakers, the noncognitivist has, in effect, sidestepped the issue by locating the fulcrum of disagreement not in belief but in noncognitive attitude.
In Chapter Two I take a closer look at this line of argument. Noncognitivism’s success depends on two conclusions: that it can preserve disagreement, and that its opponents cannot. The first of these claims is on unstable ground. Noncognitivist proposals are unable to give an adequate account of disagreement in the context of a plausible view of moral discourse. I argue for this in part through an examination of historical examples: I discuss the early proposals of the first generation of noncognitivists and show that they face serious problems. They have difficulty preserving real incompatibility between speakers (and sentences), they don’t succeed in identifying moral evaluation as one sort of normative judgment among many, and they threaten to rely illicitly on the judgments they seek to explain.

These historical investigations are significant for two reasons. First, they reveal a legacy of noble failures: early noncognitivists were aware of many of the significant problems facing their views and made efforts to solve them using the resources available to them. These thinkers are often dismissed too quickly, and close reading reveals that such insouciance is unjustified. Second, the same problems that faced early theorists still haunt the contemporary versions of noncognitivism. By seeing the origins of these problems, we’re better able to appreciate their force.

After this partly historical discussion, I turn to the work of Allan Gibbard and Simon Blackburn, who have revitalized the noncognitivist tradition. Gibbard is interesting for our purposes because of his reliance on particular sentiments—guilt and anger—in developing his account of moral judgment. This attempts to solve the ‘specification problem’ (that is, the problem of saying just which mental states serve as the foundation of moral discourse) but threatens his account with circularity. Blackburn,
on the other hand, tries to avoid the problem altogether. This has all the benefits of theft over honest toil, and I argue it’s a more radical break from the noncognitivist tradition than Blackburn allows. It’s also unsuccessful in its own right. Thus, I conclude, the initial promise of the noncognitivist approach to univocity results in disappointment.

In Chapter Three, I turn toward an argument against realism’s ability to make sense of univocity. The argument I discuss has deep roots in the noncognitivist tradition, going back at least to Hare and Stevenson, and it has been revived by Terence Horgan and Mark Timmons in the form of the so-called ‘Moral Twin Earth’ argument. I offer three independent arguments against their criticisms and conclude that the realist faces no threat from Moral Twin Earth. The best-known argument against realism’s ability to account for univocity fails. (One interesting result of this discussion is an unusual but, I think, highly plausible view about the connections between moral judgment and other forms of evaluation. This view allows the realist to synthesize the most plausible parts of the noncognitivist tradition while defending his original thesis about moral judgment and language.)

Realism is not yet off the hook, however. Chapter Four suggests that, contrary to appearances, the realist is in arrears on the issue of disagreement. Realism’s appeal to popular resources in the philosophy of language seems to solve the problem by providing a way to link diverse speakers to a shared set of moral properties despite their differences in theory. This sort of response fails to make sense of various important and distinctive features of our moralizing. In particular, it sits uneasily beside the contestability of moral judgments and the absence of authority and deference relations in moral discourse. These differences are significant, because they leave the realist’s semantic claims without
sufficient justification. That is, the realist needs to argue, not assert, that speakers should be interpreted as sharing reference to common properties, and various features of our moral discourse make that argument impossible.

I suggest that, in making this argument, the realist needs to rely on the convergence claim, the thought that idealized moral inquiry would result in the resolution of moral dispute. This appeal is made more difficult by a distinction between what I call optimistic and pessimistic readings of that claim. The former says that we converge because our commitments linked us to shared properties all along, and continued inquiry simply reveals this fact, while the latter suggests that there’s no fact of the matter about where we’ll converge, at the end of the theoretical day. The plausibility of pessimistic convergence is fatal to the realist’s semantic project.

Finally, in Chapter Five, I consider some issues in the philosophy of language that might serve as the basis for a rejoinder from the realist, or, at least, as a way of saving some of the realist’s central commitments. Recently, several philosophers have argued that something like pessimistic convergence is true of all inquiry, or at least of many straightforward cases of empirical inquiry. That is, our conceptual commitments underdetermine the path of future investigation. If so, we might think that realism’s semantic story is wrong, though it’s right to be committed to the idea that moral language (or normative language generally) functions just as other discourses do. I argue that this rejoinder fails because of various differences between moral and empirical discourse. Our moralizing simply places too great a strain on univocal interpretation.
2.1 Two arguments involving disagreement

Moral disagreement figures in two distinct types of arguments for noncognitivist construals of moral discourse. The first, which we might call the *explanatory* argument, appeals to the (alleged) existence of intractable moral dispute and claims that some variety of noncognitivism is part of the best explanation of this phenomenon. The second is what we might call the *semantic* argument. This begins with the platitude that, in order to preserve meaningful conversation, interlocutors must share meanings; if we’re not talking about the same thing, we’re talking past each other, and so we’re not engaged in real dispute at all. The semantic argument then goes on to claim that various forms of moral realism cannot make sense of shared meaning in arguments between disputants with strikingly different normative commitments.\(^3\) As a positive proposal, this argument suggests that the best way to preserve these disputes is by appeal to something like

---

\(^3\) The semantic argument is also tied to the resolution of dispute, because *resolvability* would provide evidence of semantic continuity and thus shared meaning. Thanks to Nick Sturgeon for pointing this out.
clashing attitudes or competing endorsements. If we understand speakers as manifesting some sort of noncognitive attitude toward, say, a proposed action, the argument continues, we’ll be able to make sense of real incompatibility where the realist sees only equivocation.

The semantic argument thus makes both a negative claim about the anti-noncognitivist’s prospects and a positive claim about the preservation of dispute. It’s this second claim I’ll take issue with here. The noncognitivist has made no dialectical headway if she’s unable to come through with the promised account of how genuine dispute is possible. And, as we’ll see, it’s quite difficult to provide a satisfying story about moral disagreement using only the meager resources allotted to the theory.

So the point of this discussion is negative: it seeks to cast doubt on the noncognitivist’s breezy confidence in having at hand a compelling account of disagreement, and, in doing so, it attempts to turn the tables by showing that a traditional strength of noncognitivist proposals is in fact a weakness. I begin the argument for this conclusion with a survey of the first generation of noncognitivist theorists. This is partly of historical benefit: many meta-ethical discussions are hampered by a misunderstanding of early noncognitivist views, and they fail to appreciate the extent to which these thinkers raised and at least attempted to answer puzzles that are of interest to contemporary philosophers. This look backward will also help us to understand how these puzzles arose and why they resist easy solutions. In addition it will help us to see what requirements a successful account must meet.

The second part of this chapter is an examination of the two leading contemporary noncognitivists (or expressivists, as they prefer to be known), Allan Gibbard and Simon
Blackburn. Despite important similarities in their views, they approach the semantic argument quite differently. I’ll argue that neither one is successful in providing a satisfying account of moral disagreement.

2.2 Ogden and Richards

One of the earliest statements of an identifiably emotivist view comes from C.K. Ogden and I.A. Richards’ *The Meaning of Meaning*. The authors don’t give us a thoroughly worked out view of evaluative language, but they do suggest several ideas that will play an important role for early noncognitivists. In a passage later quoted on the frontispiece of Charles Stevenson’s *Ethics and Language*, they write:

This peculiar [ethical] use of ‘good’ is, we suggest, a purely emotive use. When so used, the word stands for nothing whatever, and has no symbolic function....‘is good’...serves only as an emotive sign expressing our attitude to this, and perhaps evoking similar attitudes in other persons, or inciting them to action of one kind or another. (Ogden and Richards 1927: 125)

Later, they elaborate on this by comparing the ethical use of language to other functions.

The emotive use of words is a more simple matter, it is the use of words to express or excite feelings or attitudes. It is probably more primitive...in many cases, moreover, emotive language is used by the speaker not because he already has an emotion which he desires to express, but solely because he is seeking a word which will evoke an emotion which he desires to have; nor, of course, is it necessary for the speaker himself to experience the emotion which he attempts to invoke. (Ogden and Richards 1927: 149)

Finally, they offer a guide to identifying the emotive use of language.

The best test of whether our use of words is essentially symbolic or emotive is the question—‘Is this true or false in the ordinary strict scientific sense?’ If this question is relevant then the use is symbolic, if it is clearly irrelevant then we have an emotive utterance. (Ogden and Richards 1927: 150).
These passages are interesting because they prefigure later, and more fully developed, attempts to offer expressivist or emotivist construals of normative language.\footnote{It’s also interesting to note that Ogden and Richards are a bit ahead of their time when it comes to the truth predicate. They recognize that we sometimes call emotive utterances “true,” and so they see a danger in their proposed test. The solution, they say, is to distinguish between two senses of “true”: one ‘strict’ sense (true\textsubscript{s}) is applied to symbolic sentences, while the other (true\textsubscript{e}) applies to emotive utterances. This latter sense is “a purely evocative use of True—[it is used] to excite attitudes of acceptance or admiration” (p. 151), which suggests how the noncognitivist might go about preserving the use of the truth predicate in moral discourse. It’s important to see how this suggestion differs from a minimalist notion of truth that appeals to later noncognitivists. Ogden and Richards’ use of two different senses of “true” fails to preserve the univocity of the predicate across (symbolic and emotive) contexts, just as it must if the dividing line between symbolic and emotive meaning is drawn by invoking the (strict) notion of truth. (That is, if “true” were used in the same sense across all contexts, then the application of “true” couldn’t be the test of whether a sentence had emotive or symbolic meaning). The point of more recent attempts to secure the truth predicate in ethical contexts is that the predicate remains univocal across different discourses, though it doesn’t carry with it the metaphysical presumptions that would trouble the ethical anti-realist. See Wright (1992) for a discussion of minimalism and pluralism about truth. Charles Stevenson also makes use of the idea of “true” as a term of endorsement. See (1944: 169-174).}

There are several important ideas here. First, Ogden and Richards think that bits of normative language (such as “good” in its ethical sense) aren’t used to express properties or refer to various entities. Calling something good is different, they claim, from calling it red or heavy. Second, they invoke the notion of truth (in its narrow sense) as a way of distinguishing emotive use from more straightforward referring applications of language. These first two points are connected, at least in the following rough and ready way: if, in using moral language, we’re not describing how things are, then we can’t ask if our description is an accurate one. Third, Ogden and Richards offer a positive, albeit broad and inconclusive, proposal concerning what is going on in our moral evaluations. When we use emotive language, we’re expressing, evoking, or inciting feelings, attitudes, emotions or actions. There are important differences between these options, of course. Later in the chapter I’ll be particularly interested in how attitudes, emotions, and feelings work, and how these different states might help us...
preserve meaningful disagreement. For now, though, I want only to get the basic ideas on the table.

2.3 Duncan Jones

According to C.D. Broad, Austin E. Duncan Jones held views similar to those suggested by Ogden and Richards’ discussion (see Broad 1933: 250- 253). He suggests that the grammatical form of sentences such as “this is good” might be misleading—“good” there \textit{appears} to stand for the name of a characteristic or property, but in fact it serves some other function. Duncan Jones’ view identifies three alternative roles that the utterance of such a sentence might play. First, it might serve to \textit{express} the speaker’s emotion; second, it might be an instrument to \textit{evoke} certain emotions in its audience; third, it might be used to “command or forbid certain actions in the hearer” (Broad 1933: 251).

These, we should note, are importantly different options. The first two have close ties with particular emotions, and so it’s fair to ask just which mental states are being expressed or evoked with evaluative utterances. We would also need to see just how this form of expression worked, since, clearly, a moral claim isn’t quite like a cry of pain. And if normative terms are (merely) tools to evoke sentiments, rather than referring terms that also happen to evoke sentiments, we’d have to give an account of why, say, moral discourse has the surface grammar it does, despite this fact. We’d also have to explain how some uses of moral terms don’t play the evocative role this view sees as central.\footnote{Familiar Frege-Geach worries loom: what’s evoked by “I wonder if Smith thinks it’s right to X?” And condemnatory attitudes seem to be brought about by inflammatory claims like “Smith thinks infanticide is good.” This isn’t a careful argument, of course, but we can see how it would be difficult to account for all the various uses of “good” (even in its moral sense) by saying that “good” is a tool for evoking one
The third suggestion doesn’t take on this commitment, of course, but it does leave us the task of explaining how different kinds of commands (say, those related to different kinds of normative assessment) are distinguished from one another. Working out how to fit these diverse proposals together is a project for another time. What’s intriguing here is that Duncan Jones’ three possibilities cover some of the most important noncognitivist alternatives: the expression and evocation of emotion are the central ideas of Stevenson’s (developed) view, while imperatives play a crucial role in Hare’s prescriptivism. And both of these are also suggested in Ayer’s rather polemical critique of ethical discourse in *Language, Truth, and Logic*.

2.4 Barnes

Another early statement of similar ideas occurs in Winston H.F. Barnes’ brief article “A suggestion about value.” Barnes’ discussion is worth examining because it raises two important noncognitivist themes: it attempts to account for moral disagreement by appeal to clashing attitudes, and it recognizes the need to distinguish between various kinds of evaluation. Looking carefully at his work will raise problems that will play an important role in our discussion of contemporary proponents of expressivist views.

Barnes claims that value judgments are not, strictly speaking, *judgments* at all, by which he seems to mean that they don’t express propositions and aren’t truth-evaluable. Instead, they—or, properly speaking, their linguistic expressions—are “exclamations expressive of approval” (Barnes 1933: 45). Like Ogden and Richards, Barnes thinks that evaluative claims don’t describe the way things are, and he follows them in appealing to particular sentiment, as though uttering the word were like pulling an emotionally efficacious lever in one’s interlocutor.
(noncognitive) attitudes to explain what these claims are doing. So far, this is nothing new. But he goes on to add something interesting to the discussion: an attempt to account for the special features of certain evaluative disagreements.

The problem he recognizes is simple. In disputes about, say, ‘matters of taste,’ we’re likely to end the conversation by noting that we just feel differently—and once this move is made, there’s little more to say, and little temptation to continue. Other disputes about value, particularly in the moral case, don’t end this way, because we’re often not content to stop once we see that we’ve got different attitudes. But why this disparity, if all value-judgments are expressions of attitudes of approval and disapproval? To put the puzzle another way, Barnes’ suggestion leaves us with an explanatory debt. We need to say why some, but not all, differences in approval lead to protracted dispute.

This, in turn, requires solving a more fundamental puzzle. Just how can differences in attitude serve as the basis of real disagreement in the first place? We can see how contrary doxastic attitudes (belief and disbelief, say) might serve as the basis for continued argument, because both beliefs can’t be true. But attitudes of approval and disapproval, as Barnes understands them, aren’t representational, so they can’t conflict in just the same way. In response, we’ll need to provide some account of the clash between noncognitive attitudes.

Barnes’ proposed solution to these puzzles draws on the notion of incompatible actions:

The reason we are not content [to end the conversation by noting a difference in feeling and leaving the matter at that] seems to lie in the fact that the action of one man dictated by his approval of something is frequently incompatible with the action of another man dictated by his approval of something. (Barnes 1933: 46)
Actions are incompatible, in a sense yet to be developed, and it’s this incompatibility that grounds the tension between the attitudes leading to these actions. This tension, in turn, explains protracted disputes. The resulting arguments are attempts to change others’ attitudes so that they are in harmony with one’s own:

It is this opposition between the approval of one man and that of others which lies at the bottom of controversies. If I maintain ‘A is good’ against the contention ‘A is bad,’ my attempt to prove the truth of my statement is not really what it pretends to be. I point out the details in A which are the object of my approval. By doing so I hope that my opponent, when he becomes aware of these, will approve A: and so be ready to say ‘A is good.’ But what I have done is not really to gain his assent to a proposition but to change his attitude from one of disapproval to one of approval towards A. All attempts to persuade others of the truth of value judgment are thus really attempts to make others approve the things we approve. (Barnes 1933: 46)

So Barnes’ answer to our first puzzle—why moral disputes function differently from differences in taste—has something to do with the idea that some differences in attitudes, and not others, lead to incompatible actions, and so must be resolved. This resolution requires argument, which, properly understood, is simply an attempt to bring about changes in our interlocutors’ attitudes.

These responses are incomplete. The notion of incompatible actions isn’t perfectly clear, and so we don’t have a developed account of just how we’re at odds when we offer competing evaluative judgments. Furthermore, we don’t know how to distinguish moral evaluations from other sorts, a distinction we need to make if we’re to explain fully the asymmetries between different kinds of evaluative disagreement.

---

6 It’s interesting to note that incompatible actions seem to stem from agreement in attitude, for example when we both pursue the same unshareable goal out of a conviction that it’s good. This appearance disappears if we think that my attitude is, for example, an attitude of approval directed to a state of affairs in which I get the last cookie, and your attitude is an attitude of approval of a different state of affairs, namely, the one in which you get it. The difference in the attitudes’ objects is conflated in the common description “we both want the last cookie.” Thanks to Don Hubin for pointing this out to me.
Since our interest here concerns disagreement, the first of these worries is the most relevant. How can different attitudes of (dis)approval serve as the basis for real disagreement? Most disagreements about how to act just don’t seem to involve incompatibility in a narrow sense, because my attitudes most directly concern my actions: if I disapprove of stealing, for example, I can perfectly well refrain from shoplifting while you pursue your life of crime. Hence your approval of stealing leads to an action that is compatible with my disapproval of stealing and subsequent abstention; but it surely seems that we have a disagreement here, if anywhere, and this isn’t the kind of discussion we could plausibly resolve by saying that we just happen to feel differently about stealing. This point undermines the attempt to make sense of the differences between moral attitudes and differences in taste, because just the same point can be made about the latter sort of case. My attitudes toward ice cream dictate what kind of cone I’ll order, and your different attitudes merely prompt you to order something else. Since we’re looking for a distinction between these kinds of evaluation, Barnes’ suggestion doesn’t appear to be enough. What’s needed, in response, is a more developed notion of attitudes and their links to behavior.

---

7 This is just a particular instance of a general problem: whatever (noncognitive) attitude is said to be the root of the explanation of evaluative judgments, that attitude has to fit neatly with the judgments it allegedly explains. Here, for example, the problem is that approval might not have the ‘outward-reaching force’ of moral judgments, but there are other worries as well. For example, George Pitcher has argued that to be the object of approval something must be under human control, and so it makes no sense to approve of constellations or weather patterns (see Pitcher 1958: 198 ff.). If so, then approval can’t be behind all judgments of goodness, because we can judge these things to be good. If a candidate attitude doesn’t apply to enough of the right things, then we can’t plausibly claim that this attitude provides a successful account of the predicate in question.
2.4.1 Barnes’ second attempt

Barnes takes up some of these issues in a later article, “Ethics without propositions” (Barnes 1948). His starting point is Duncan-Jones’ suggestion that moral claims are a kind of *command*. In particular, they are commands that meet the following conditions: (a) they are “fortified” with reasons; (b) the reasons are not desires on the part of the person addressed; and (c) the reasons are “of a certain number which can be listed” (Barnes 1948: 18-19). What’s interesting about this approach is that it prefigures R.M. Hare’s prescriptivism, by relying on the analogy between moral claims and imperatives, while hinting, in its third condition, at *substantive* constraints on what counts as a moral judgment—constraints that Hare will take pains to avoid (see Hare 1963: 16 ff). (In fairness, it’s not completely clear what sorts of reasons Barnes has in mind here, and they could turn out to be additional ‘formal’ features that don’t put any limit on the content of moral judgment. But his remarks do suggest a limit on what sort of consideration counts as a moral reason, and this is plausibly understood to pick out the kinds of restrictions that Hare avoided.)

The ‘command theory,’ as Barnes calls it, has a breezy plausibility when applied to, say, second-person moral claims in the present tense. It needs to be broadened, though, to capture other sorts of moral utterance, and Barnes’ attempt to extend the suggestion relies again on the notion of attitudes. These are distinct from mere likings or dislikings, he thinks, because they involve broader and more complex commitments. I might like helping the needy, yet dislike others doing the same thing, but if I *approve* of helping the needy then I endorse and encourage others’ doing so, and I feel pleased if they do (Barnes 1948:19). Likewise, disapproving of my own cruelty, unlike merely
disliking it, requires that I have a similar “anti-emotion” toward anyone else’s cruelty (Barnes 1948: 20). Hence a claim such as “you ought not to be cruel” might “have the force of Don’t be cruel; don’t enjoy cruelty in others; I don’t” (Barnes 1948: 20).

This broader account of attitudes will help us to preserve real conflict in cases where, say, I approve of something and you disapprove of it. Part of what I’m saying in expressing my approval, according to Barnes’ extended model, is do this! and part of what you’re saying in rejecting my claim is don’t! In the clash of imperatives, we might find a way to preserve a real incompatibility and disagreement. I can’t sincerely assent to contradictory indicative claims, because this evinces incompatible beliefs; likewise, I can’t assent to contradictory imperatives, because doing so commits me to incompatible actions.8

(In passing, it’s worth noting a point about the relationship between moral judgment and all-in endorsement or, to put it another way, between judgments of what’s right or obligatory and judgments of what to do. The role of imperatives in Barnes’ view might suggest that, in making moral judgments, we’re commanding actions, and that this, in turn, means that these judgments settle the question of how to act. As we’ll see, this sort of connection is problematic. But Barnes doesn’t have to accept it, as long as his condition (c) is understood in a substantive way. Suppose we understand the constraint as identifying a certain set of considerations as defining what counts as moral reasons or the moral point of view. You and I might be fairly lax in our commitment to doing what these reasons require or to seeing things from the moral perspective. We can then

8 One might object that our notion of contradiction is too bound up with the notion of truth. It seems fairly clear, though, that our ordinary talk allows for commands to contradict one another, and this expansive notion of contradiction is familiar and natural enough to relieve worries about covert reliance on the truth-predicate.
imagine a case where we agree that the thing to do is to do what’s best for us prudentially [and we agree about what action that is], yet we disagree about what to do in the hypothetical cases in which there are no contravening non-moral reasons, or in which we’re both fully committed moralists. That is, we might have a disagreement about what action is demanded by the moral reasons, though neither one of us is committed to performing that action. For example, you might think that, morally speaking, we ought to heed considerations of desert, while I think that the moral point of view demands that we act for the greater good—though in the here and now we both think that prudential considerations override either of these reasons. The disagreement is, in some sense, about what to do, though it doesn’t require that we’re committed to incompatible actions in the actual world. By delineating a substantive set of moral reasons [or by identifying the moral point of view in some non-formal way], Barnes allows us to construct the relevant subjunctive conditional, viz. if these moral reasons were all that mattered, then.... Without this specification, it’s hard to see how we could preserve disagreement by appeal to clashing prescriptions without also making the further, and I think implausible, claim that moral commitments settled the broader question of what to do.)

One last feature of Barnes’ discussion deserves comment: his claim that particular emotions mark out moral judgment. Barnes thinks that nonmoral attitudes of approval or disapproval can have the same structure as moral attitudes. He notes, for example, that a man might dislike smoking, dislike others’ smoking, and discourage them from smoking.

---

9 Thanks to Justin D’Arms for pointing this out to me. Of course, it’s not at all easy to see just what restrictions on content could be ‘built in’ without controversy, since most interesting proposals would be controversial.
while not having a *moral* attitude. This sort of disapproval is not yet a moral commitment. Hence, Barnes thinks,

> The differentia of the moral attitude must be sought elsewhere. I think that we must admit that if there is to be a moral attitude, it must contain at least one peculiarly ethical component, viz., the pro- or anti- emotion which I feel when I contemplate actions of the kind in question. (Barnes 1948: 22)

He also remarks that “a peculiarly moral feeling is involved in our moral attitudes” (Barnes 1948: 23). This is, admittedly, a rather odd way of distinguishing moral and non-moral attitudes: we might have thought that the difference amounted to more than the occurrence of a particular emotion or ‘feel’ when contemplating the action at issue. As we’ll see, Charles Stevenson also relies on the notion of a distinct moral emotion, for similar reasons.

The salient points of this discussion, for our purposes, are these. First, it’s not immediately obvious just how attitudes of approval or disapproval form the basis of real conflict and disagreement. Second, Barnes’ puzzles about different sorts of disagreement, and different sorts of evaluation, are serious, even if his solutions are not entirely satisfying. As we look at more developed noncognitivist accounts, we should keep in mind the broader issues raised by these puzzles, namely, questions about different sorts of evaluations and their relationship to endorsement. Third, the appeal to *imperatives* provides a way to make sense of genuine disagreement, though it’s not sufficient, by itself, to make sense of different sorts of evaluation. Finally, Barnes’ appeal to distinct moral emotions is in line with Charles Stevenson’s contemporary view and foreshadows Gibbard’s (1990) reliance on distinct sentiments in his own account of moral discourse.
2.5 Ayer

A.J. Ayer’s argument for an emotivist view of moral language has often been viewed as a prototypical example of early noncognitivist thought. This is in some ways unfortunate, because Ayer, unlike Stevenson, Hare, and perhaps Barnes, is motivated more by general epistemological and semantic concerns than by features of our moral practice. Chapter Six of *Language, Truth, and Logic* is, after all, a critique of ethics and theology, and Ayer considers moral and theological discourse because they provide an objection to positivism’s view of cognitive significance. Ethical claims appear to be meaningful, yet they aren’t either analytic or empirically verifiable—so they pose a threat to the view of meaningfulness that Ayer defends. Positivism of this sort is less popular than it once was, and we might wonder if Ayer’s view is less interesting because one of its principal motivations has been undermined. Nonetheless, his remarks on ethics are

---

10 Austin Duncan Jones’ motivations also seem to be more attuned to features of moral language than to broader concerns like those that drive Ayer. From Broad’s account, it appears that Duncan Jones takes Moorean considerations as his starting point. The basic idea is that expresivism offers a powerful explanation of why (a) analyses of ethical terms containing only non-normative terms seem to be missing something, as Moore’s open questions point out, and (b) analyses that do rely on some normative notion don’t feel inadequate in quite the same way (Ayer 1984: 22 makes this latter point as well). Moore’s questions, after all, seem to do more than just lead us to reject particular analyses; they seem particularly good at making us pessimistic about the very project of looking for more. See Darwall, Gibbard, and Railton for a related discussion of the diagnosis of the open question argument’s appeal.

11 Ayer begins his discussion in this way:

There is still one objection to be met before we can claim to have justified our view that all synthetic propositions are empirical hypotheses...It will be said that ’statements of value’ are genuine synthetic propositions, but that they cannot with any show of justice be represented as hypotheses...and, accordingly, that the existence of ethics and aesthetics as branches of speculative knowledge presents an insuperable objection to our radical empiricist thesis...In the face of this objection, it is our business to give an account of ’judgments of value’ which is both satisfactory in itself and consistent with our general empiricist principles. (Ayer 1952: 102)

This isn’t an account that arises from reflection on the nature of our ethical thought, and, of course, Ayer will go on to say that our thought is radically confused. Notice that Stevenson, whose view is quite similar to Ayer’s in many ways, in its content, is more sympathetic to the idea that our meta-ethical analysis must pay more respect to the workings of our discourse. When considering the idea of clarifying the question “what is good?,” he proposes that any successful reformulation must be relevant in the sense that “those who have understood the definition must be able to say all that they then want to say by using the term in the defined way” (Stevenson 1937: 15). So it’s more likely that, for Stevenson, a reconstruction’s incompatibility with our practice is a mark against the reconstruction, not our practice.
worth considering, despite this troubled provenance, not least because he gives explicit
attention to the issue of disagreement. I’ll suggest here that his treatment of the issue
faces serious difficulties, but examining just how he goes wrong will help us to see what
noncognitivism must do in order to give a satisfying solution to the problem.

Ayer’s first statement of his view suggests that he thinks of ethical terms as
instruments we use to evince particular sentiments:

Thus if I say to someone, ‘You acted wrongly in stealing that money,’ I am not
stating anything more than if I had simply said, ‘You stole that money.’ In adding
that this action is wrong I am not making any further statement about it. I am
simply evincing my moral disapproval of it. It as if I had said, ‘You stole that
money,’ in a peculiar tone of horror, or written it with the addition of some
special exclamation marks. The tone, or the exclamation marks, adds nothing to
the literal meaning of the sentence. It merely serves to show that the expression
of it is attended by certain feelings in the speaker. (Ayer 1952: 107)

Immediately after this, he says that moral claims (e.g., of the form “stealing is wrong”)
“[express] no proposition which can be either true or false”—instead, it’s as though, by a
certain linguistic convention, the presence of the moral term shows “that a special sort of
moral disapproval is the feeling which is being expressed” (Ayer 1952: 107). Later, he
claims that “the function of the relevant ethical word is purely ‘emotive’” and that “it is
used to express feeling about certain objects, but not to make any assertion about them”

So far, this is more or less the view of Ogden and Richards or the early Barnes,
with the important difference that it’s feelings or sentiments that are being expressed,

---

12 Later, Ayer would retreat from this view:
To say, as I once did, that this moral judgments are merely expressive of certain feelings, feelings
of approval or disapproval, is an oversimplification. The fact is rather than what may be described
as moral attitudes consist in certain patterns of behavior, and that the expression of a moral
judgment is an element in the pattern. The moral judgment expresses the attitude in the sense that
it contributes to defining it. (Ayer 1949: 238)
rather than *attitudes*. This alteration is not a step forward, though, because feelings—and, to a lesser extent, sentiments—seem to be insufficiently stable to serve as a foundation for moral judgment. It’s plausible to think that attitudes are dispositional: one can have an attitude toward something without being conscious of it from moment to moment, and without any constant occurrent state. This point seems less clearly true of sentiments and emotions. This is problematic, because moral judgments can be made in a flat affective state, and so they don’t require any particular occurrent feeling or emotion. Hence moral judgments and feelings are modally distinct, something Ayer’s account must deny.

But Ayer offers other suggestions, which are importantly different and which parallel the suggestions made by Duncan-Jones and the later Barnes. His second proposal concerns the effects of ethical utterance. Moral claims are “calculated also to arouse feeling” and may be defined “in terms both of the different feelings they are ordinarily taken to express, and also the different responses which they are calculated to provoke” (Ayer 1952: 108). (This, as we’ll see, is quite close to Stevenson’s view of emotive meaning.) The third view is that moral claims function as *commands*. A sentence such as “it is your duty to tell the truth” “may be regarded both as the expression of a certain sort of ethical feeling about truthfulness and as the expression of the command ‘tell the truth’” (Ayer 1952: 108). In total, there are three distinct positions suggested in his discussion, roughly the same as those suggested by Duncan Jones: moral

---

13 Whatever kind of definition this is, it can’t be one that gives moral claims *content* or implies that they express propositions, because this contradicts the claim that moral utterances don’t express propositions and can’t have truth values. This suggests that we’d have less an account of meaning than a view about the *uses* of the term.
terms are used to express feelings or sentiments; they arouse these states; and they work as imperatives.\textsuperscript{14}

As we’ve seen, these three suggestions are different in important ways. The first two are committed to an intimate tie between moral discourse and some set of noncognitive mental states. (Ayer speaks most often of feelings, less frequently of sentiments, and he doesn’t make reference to broader categories such as emotion or attitude). The nature of these states is important because, if this kind of view is to be at all plausible, they must be intimately connected with moral judgment.\textsuperscript{15} That is, we must be able to offer a convincing explanation of our moral practice by appealing to these (as yet unspecified) mental states. This view thus requires that there are sentiments, attitudes, feelings, or other appropriate candidates that are attached, with sufficient accuracy, to our moral judgments.\textsuperscript{16} If this requirement can be met, though, we can distinguish moral evaluation from other sorts of judgments by appealing to the particular mental states at work, as Barnes did by appealing to distinct moral emotions. Moral claims are marked out by their ties to these particular sentiments or feelings.

\textsuperscript{14} Toward the end of his career, Ayer had developed the suggestion that moral claims express commands in a way that synthesized it with his suggestions about emotive expression: Suppose that we render ‘This is good’ as ‘This is to be approved of,’ where ‘is to be’ is construed in a purely prescriptive fashion....This device...has the advantage of bringing out the prescriptive element in our use of ethical terms. (Ayer 1984: 30)

If we understand the claim “it’s right to []” as “approve of []!” then we get something related to Allan Gibbard’s norm-expressivism, insofar as our moral statements as prescribing particular attitudes rather than action.

\textsuperscript{15} Ayer seems to be in arrears here. As noted above, feelings and emotions seem to be modally distinct from moral judgments. On the other hand, we do sometimes talk as though emotions are at least partly dispositional: John’s still angry about what you did, even though right now he’s not actively fuming because he’s got other things on his mind.

\textsuperscript{16} In order to be at all plausible, in other words, this sort of account must first show that some noncognitive mental state accompanies most instances of moral judgment; without this, the further claim, that the state explains or constitutes the judgment, is untenable.
The third, ‘imperatival’ view is, on its surface, less intimately connected to any particular sentiments or feelings. On this suggestion, my moral judgment isn’t an expression or an evincing of any particular mental state; it’s a command for you to act in a certain way, which isn’t quite the same thing. We’ve looked quickly at this idea in our discussion of Barnes, but it’s worth a more careful examination. To understand how this sort of view might work, we’ll need to look more closely at imperatives and their ties to mental states and to action. Suppose a person makes a moral judgment and (sincerely) expresses it by saying something like “it’s wrong to A” to her interlocutor. What’s just happened? Surely her utterance isn’t exactly like issuing the command “don’t A” to a particular person, because moral judgments don’t function as singular imperatives. The most general reason for thinking so is that singular imperatives involve fewer commitments on the part of the speakers who issue them than do moral judgments.

One of these additional commitments involves what one says about similar cases. Moral claims are more like universal than singular imperatives, because their scope extends beyond the particular occasion of utterance. Suppose James is entertaining several guests, including Tom, who’s about to light a cigarette. “You ought not smoke in other people’s houses,” says James. It’s not too farfetched to think that, whatever else James has done, he’s ordered Tom to refrain from smoking. But he’s also done more than this: he’s leveled the same injunction at his other guests. There’s something odd about Chris asking, “but, James, do you mind if I smoke?” because it’s clear that the question has been answered by James’s earlier statement. This suggests that his

---

17 In a later essay, Ayer (1984) suggests that a moral claim is a prescription to have certain attitudes (e.g., approval or disapproval) toward something, not simply a command to action.
statement is more than a one-off command to Tom, and that, in making this claim, James has committed himself to commanding others (just who hasn’t been specified) not to smoke. James has also taken on commitments having to do with his own actions. If, over dinner at Tom’s, James takes out his cigarettes, Tom is entitled to wonder what’s going on, and he is entitled to complain—didn’t James say just last week that it’s wrong to do what he’s doing? The point here is that moral claims commit the speaker to make similar assessments of identical cases, including ones in which his own actions are under the scope of the imperative. Consistency pressures clearly apply here, as they don’t with singular imperatives.

As we saw in discussing Barnes’ ‘command theory,’ the use of imperatives raises a question about the relationship between moral judgment and all-in endorsement. Suppose moral claims are some sort of generalized imperative. This suggests that assenting to them settles the question of what to do. That is, there’s something suspicious

---

18 Identical with respect to morally relevant features, of course, but just what these are needn’t be spelled out here. For our purposes, what matters is that moral claims, unlike singular imperatives, bring with them a cluster of consistency pressures.

Another way of seeing this is to think about sincerity. We can imagine what it is to make insincere moral judgments—in fact, this seems all too common. But it’s much less clear what an insincere singular imperative would be like. To make a moral claim insincerely seems to involve a refusal to make all the commitments that the judgment brings with it: I say it, but I don’t really mean it (as you see when you notice I do what I say oughtn’t be done). Since singular commands don’t involve the same kind of commitment, it’s not clear what it would be like to be insincere in uttering one.

Universal prescriptions (commands applied to a type of action and person, rather than individual instances) might allow for something like insincerity because one could utter the command without really taking on the commitment to act as the command demands when one is in the relevant situation. Just because uttering the universal imperative involves appearing to take on commitments one might not in fact take on, we might be able to make sense of insincere claims.

19 These pressures, in turn, might involve various mental states, though not the sentiments or feelings mentioned in conjunction with Ayer’s first two suggestions. For example, since making a moral judgment seems to commit me to evaluate similar cases in the same way, I seem to need an intention to make similar commands in relevant circumstances (all other things being equal). If I’m committed to making this command in cases where I’m the one tempted to do the thing I’ve judged wrong, I might need to have the relevant intention or commitment not to , since I’ve just commanded myself not to do this. These issues will be discussed in greater detail in our examination of Hare’s view, which attempts to develop the link between moral judgment and imperatives more thoroughly.
about agreeing to the claim while refusing to act as it demands. Ayer’s third suggestion seems to entail that moral claims are simply answers to the question of what to do, so there’s no room for any gap between what appear to be two distinct judgments. Unlike Barnes, Ayer doesn’t provide a way of retreating from this conclusion, so he has no room to allow that someone might judge that moral considerations are overridden by other sorts of reasons when deciding how to act.

This raises another, related, problem. Just as we’d like to be able to distinguish moral judgments from all-in endorsement, we should also be able to differentiate between various types of evaluation and command. Not every declarative sentence that functions as an imperative is a moral claim, as Barnes’ example, “full academic dress will be worn,” demonstrates. If we’re to develop Ayer’s suggestion in more detail, we’ll need some further means of dividing moral and nonmoral claims.

Let’s turn to Ayer’s discussion of disagreement. Early in his presentation, Ayer makes clear that, by his lights, moral claims are not properly called true or false. Thus, and strictly speaking, a person with different views about the permissibility of stealing cannot contradict me, because we aren’t making factual statements—we’re “merely” expressing different sentiments (Ayer 1952: 107).²⁰ A bit later, Ayer considers Moore’s famous argument to the effect that subjectivism must be false because, according to that view, interlocutors engaged in what appear to be real disagreements are not, in fact, defending genuinely incompatible views. “If Moore’s argument really refutes the

---

²⁰ It’s important to notice that this is a special and rather idiosyncratic sense of ‘contradiction.’ We would normally think that two people have contradicted each other if they issued incompatible imperatives, even though we all understand that imperatives don’t have truth values. Think, for example, of Kant’s cases illustrating the first formulation of the Categorical Imperative in the *Groundwork*. He appeals to contradictions in practical reasoning—the agent in the ‘mutual aid’ case contradicts himself by willing incompatible prescriptions, for example—and this seems, to most readers, to be uncontroversial.
ordinary subjectivist theory,” Ayer says, “it also refutes ours.” But, he continues, Moore’s argument fails in both cases, because “one really never does dispute about questions of value” (Ayer 1952: 110). So Ayer overtly rejects the idea of (fundamental) moral disagreement.21

We can see why Ayer drew this conclusion if he thought that (a) moral claims aren’t truth-evaluable and (b) only truth-evaluable claims can stand in the sort of relationship required for real disagreement, contradiction, and argument. Our moral judgments might be incompatible in some other way; for example, it might be psychologically or conceptually impossible for me to retain my attitude while adopting yours. But if our moral claims merely reveal our sentiments, and if these sentiments are understood in a certain way, then our claims don’t clash with one another. I feel this way, you feel that way, but there’s not yet a robust conflict. Barnes, faced with this puzzle, struggled to preserve some kind of clash or real incompatibility, while Ayer appears to capitulate, concluding that there simply is no incompatibility. This, however, is hard to square with his earlier appeal to imperatives, since it seems that incompatible commands, like incompatible assertions, fall within our ordinary notion of contradiction or dispute. That is, Ayer’s third proposal seems to make sense of the very phenomenon he’s just denied.

21 On the next page, he does say that “argument is possible on moral questions only if some system of value is presupposed” (Ayer 1952: 110). The kind of argument he has in mind, though, concerns questions of fact that are taken to be relevant to questions of value. For example, we might agree that acts that maximize the good are obligatory, and yet disagree as to whether X is right because we have different beliefs about its effects. Thus, our argument about whether X is right or not is genuine. On Ayer’s view, we can’t really argue about general moral principles—“we merely praise or condemn them in the light of our own feelings” (Ayer 1952: 112).
This is puzzling. One way to make Ayer’s position more coherent is to read him as glossing over a distinction between genuine incompatibility (that is, real, rather than equivocal, disagreement, whatever it consists in) and disputes permitting reasonable discussion. He’s mistakenly rejected both sorts of disputes, while he means to reject only the latter. Notice just what it is he’s denying: we don’t dispute about questions of value; we can’t bring forth arguments to convince our opponent to come around to our moral view (Ayer 1952: 111); we can’t produce an argument about a question of value that doesn’t reduce to a question of logic or a matter of empirical fact (Ayer 1952: 112). Ayer isn’t claiming that the appearance of incompatible moral views is illusory. Instead, he’s suggesting that there aren’t neutral reasons or fair arguments in favor of one position over another. When people know all the empirical facts, make no mistakes in logic, yet differ in their fundamental moral commitments, there’s nothing more to be said. This makes sense, if moral commitments aren’t representational and moral claims don’t express propositions: since there’s nothing here to be true or false, then how could we show that our opponents have gone wrong? More accurately, in what would this going wrong consist? Even if we were committed to issuing contradictory commands, we would have a kind of contradiction that’s not rooted in some kind of error in either party. Once we’ve reached that point, there’s little more to say.

This way of understanding Ayer’s view gets some support from his discussion of what’s really happening when we argue about moral issues. He denies that we ever dispute about questions of value, yet he’s aware that we often seem to do just this:

---

22 If this is the right way to think about Ayer’s view, he’s claiming that ethics lacks what Crispin Wright has called “cognitive command.” See Wright (1992).
But, in all such cases, we find, if we consider the matter closely, that the dispute is not really about a question of value but about a question of fact. When someone disagrees with us about the moral value of a certain action or type of action, we do admittedly resort to argument in order to win him over to our way of thinking. But we do not attempt to show by our arguments that he has the ‘wrong’ ethical feeling towards a situation whose nature he has correctly apprehended...if our opponent happens to have undergone a different process of moral ‘conditioning’ from ourselves, so that, even when he acknowledges all the facts, he still disagrees with us about the moral value of the actions under discussion, then we abandon the attempt to convince him by argument. We say that it is impossible to argue with him because he has a distorted or undeveloped moral sense...It is because argument fails us when we come to deal with pure questions of value, as distinct from questions of fact, that we finally resort to mere abuse. (Ayer 1952: 110-111)

If we think of arguments as attempts to get at the truth, and there’s no truth to be had in ethical commitment, then no real arguments are possible here.

This interpretation is also consistent with some of Ayer’s later remarks on moral discourse. In his paper “The analysis of moral judgments” (originally published in 1949), Ayer wonders about the relationship between descriptive features and the moral evaluations that are said to follow from or be supported by these facts. The relationship can’t be logical, he thinks, because two people could agree about all of the circumstances of a particular case and offer different evaluations without either one contradicting himself (Ayer 1949: 236). But the descriptive features offered as reasons for moral assessment don’t support those assessments in what Ayer calls a “scientific sense,” which, presumably, means that they don’t provide inductive support for the moral conclusion. If this were the right kind of evidential relation, Ayer thinks, then the goodness or badness of the situation, the rightness or wrongness of the action, would have to be something apart from the situation, something independently verifiable, for which the facts adduced as the reasons for the moral judgment were evidence. But in the moral cases the two coincide. There is no procedure of examining the value of the facts, as distinct from examining the facts themselves. We may say that we have evidence for our moral judgments, but we
cannot distinguish between pointing to the evidence itself and pointing to that for which it is supposed to be evidence. Which means that in the scientific sense it is not evidence at all. (Ayer 1949: 236-137)

The point here is that natural features aren’t probabilistic indicators of some further moral facts; they seem to constitute the moral facts. But Ayer thinks that this possibility has been discounted by his point that two people could disagree only in evaluation without any self-contradiction. Whatever kinds of clash arise from these different assessments, the disagreement is not over some further fact about the situation, or about the way things are. Thus, something else must be going on.

This reading makes sense of Ayer’s puzzling remarks on moral disagreement. What he’s denying, I suggest, is some fair means of adjudicating genuine conflicts or disputes, rather than the existence of any kind of clash between interlocutors. His position follows naturally from the idea that fundamental moral commitments aren’t representational or truth-evaluable: since they don’t actually represent the world, what evidence or argument could show them to be mistaken? Despite this, it’s still possible for interlocutors to have genuine disagreements or to be genuinely at odds with one another.

As we’ve seen, his suggestion that moral utterances function as imperatives leaves room for this sort of incompatibility between interlocutors, without providing any additional hope for reasonable resolution. (This reading still leaves Ayer in an awkward position, because it forces him to abandon his rejection of Moore’s argument against subjectivism. That argument rests on the presence of genuine incompatibility and criticizes the subjectivist for being unable to make sense of it. Since, on my proposed reading, Ayer’s view does preserve one sense of disagreement—i.e., incompatibility—while denying another, he can no longer say that Moore’s argument relies on a mistake.)
Let’s look once again at some of Ayer’s early views of moral disagreement, which raise serious questions for the noncognitivist project. The Ayer of *Language, Truth, and Logic* thinks that, in cases where we appear to be disputing about questions of value, we either agree on fundamental moral principles and disagree about relevant matters of fact, or we’re engaged in ‘mere abuse.’ This view raises two puzzles. One is why we bother with mere abuse at all—what motivates it if we just happen to have different feelings? What grounds for insult are there in the fact that we don’t share a particular (noncognitive) mental state?

The second puzzle is related. Notice that Ayer’s discussion invokes standards of correctness governing our attitudes: our opponent’s moral sense is *distorted* or *undeveloped*, and our own attitudes are better. Just as the idea of abuse suggests a standard of assessment—I abuse people for being deficient in some way—the claim that someone’s moral sense might be defective relies on the idea that some moral attitudes are better than others, and that my interlocutor is missing something. It’s not only that he’s different; he’s worse off, as an evaluator, than I am. This isn’t going so far as to say that our moral sentiments are truth-apt in some way, but it does suggest that the norms governing them are richer than Ayer recognizes. This point is one that might help a more sophisticated noncognitivist. If there really are moral emotions or attitudes, and if they

---

23 The same point can be made about Ayer’s earlier description of moral dispute. After saying that moral claims are neither true nor false, Ayer notes that

*Another man may disagree with me about the wrongness of stealing, in the sense that he may not have the same feelings as I have, and he may quarrel with me on account of my moral sentiments. But he cannot, strictly speaking, contradict me.* (Ayer 1952: 107)

What remains mysterious is why this quarrel should seem so natural, given that we have only a case of different sentiments. What Ayer *can’t* say is that there’s a right way to feel; he can’t invoke standards of correctness for sentiments—unless, of course, he offers an expressivist account of these correctness-judgments. Notice, though, that this would require a distinction between the *correct* way to feel and the *morally commendable* way to feel. It’s the former notion that Ayer needs, so it’s not enough to invoke his *moral* analysis to explain judgments about warranted sentiment.
are richer than the moral feelings Ayer invokes, then these will provide valuable resources for the noncognitivist’s attempt to account for the nuances of our moral practice.

Ayer’s early account raises a number of interesting issues for the aspiring noncognitivist. His suggestion concerning imperatives, like Barnes’ later discussion of the same sort of view, promises to allow for the requisite incompatibility between clashing moral claims. And he raises, but does not develop, the promising idea that particular sentiments might give us the resources we need in order to account for disagreement. In order to make good on this suggestion, we’ll need to clarify just what mental states are involved in moral judgment, and explain how they work. As we’ll see, several of these issues are taken up by Charles Stevenson, who attempts to make good on these debts.

2.6 Stevenson

Though the first examples of Charles Stevenson’s work in ethics appeared shortly after the publication of *Language, Truth, and Logic*, his discussions of moral discourse show a sensitivity to its nuances that’s for the most part lacking in the earlier noncognitivists. While Ogden and Richards, Duncan Jones, Barnes, and Ayer offer identifiably noncognitivist views of moral judgment, Stevenson is arguably the first to attempt to motivate this position in a compelling way. J.O. Urmson credits Stevenson, rightly, I think, as being “the first emotivist of all to take ethics seriously, for its own sake, in print” (Urmson 1968: 22). Though I’ll argue that Stevenson’s view is not as plausible as other noncognitivist alternatives, it deserves attention; like Ayer’s work, Stevenson’s efforts are instructive even in their failures. More importantly, Stevenson
gives an insightful articulation of some early motivations for the noncognitivist approach. In this section I’ll outline some of the key features of Stevenson’s work and discuss some difficulties arising from his treatment of disagreement and the attitudes it involves.

In his early paper “The emotive meaning of ethical terms,” Stevenson describes his task as the search for a *clarification* of the question “is X good?” by way of substituting another question for it (Stevenson 1937: 14). Any plausible candidate for this substitution must be, as Stevenson puts it, *relevant*. This requires that “those who have understood the definition must be able to say all that they then want to say by using the term in the defined way” (Stevenson 1937: 15). The idea is that any reformulation of “this is good” must allow us to do the important part of the work that we wanted to do with our original (and problematic) terminology.\(^2\) This hints at an important noncognitivist theme, namely, the idea that we should explicate moral terms and concepts with an eye toward their *functions* or, roughly, what tasks they perform for us.\(^3\) Because so much of the work of moral discourse is *practical*—guiding conduct and shaping character, for example—this line of thinking, implicit in the relevance criterion, suggests that our account of moralizing should preserve its action-guiding role.\(^4\) This, in turn, has encouraged both early and contemporary noncognitivists to develop views grounded on attitude rather than belief.

---

\(^2\) We might not be able to save everything. It could turn out that our ordinary moral discourse is so riddled with metaphysical confusions that some of our commonsense thought must be abandoned.

\(^3\) As we’ll see, both Blackburn and Gibbard invoke roughly this idea in motivating their expressivistic accounts. The appeal has become more explicit in Blackburn’s later work; see for example (Blackburn 1998) and (Blackburn 1999).

\(^4\) When introducing his idea that moral claims have interesting similarities to imperatives, Stevenson appeals to practical functions:

> both imperative and ethical sentences are used more for encouraging, altering, or redirecting people’s aims and conduct than for simply describing them...And in arguments that involve disagreement in attitude, it is obvious that imperatives, like ethical judgments, have an important place. (Stevenson 1944: 21)
In addition to the relevance criterion, Stevenson gives us three other requirements for any plausible proposal. First, it must allow us to disagree sensibly about whether or not something is good. This condition, inspired by Moore’s argument against subjectivism, is thought to rule out views that understand “this is good” as a claim about a speaker’s desires or preferences. Were we simply reporting our attitudes, then our apparently contradictory moral claims would both be true, and our disagreement would disappear. (Another way of putting this is to say that our claims seem to be about the same thing, e.g. the action in dispute, but, according to the subjectivist line, we’re talking about different things). Later, I’ll argue that Stevenson has more difficulty meeting this requirement than is usually supposed, because it’s not clear that he captures the right sort of disagreement.

Second, (judgments of) goodness must have “a magnetism”—that is, they must carry with them some sort of motivational weight. A person judging that something is good “must ipso facto acquire a stronger tendency to act in its favor than he otherwise would have had” (Stevenson 1937: 16). This, like the relevance criterion, is an attempt to ensure that any proposed analysis leaves ethical terms able to play their characteristic role. It isn’t too much of a stretch to see this condition as foreshadowing Hare’s

---

27 It’s interesting that this condition seems to exactly contradict Ayer’s view that we don’t really disagree about questions of value after all. I think this apparent tension is at least partly illusory, because it seems to rest on different sense of disagreement or dispute.

Michael Smith (1986) points out that the subjectivist may not be as badly off as Moore and Stevenson suppose. Two interlocutors with apparently contrary moral views are, on this interpretation, reporting their divergent attitudes, but, in cases of sincere utterance and reasonable self-awareness, it follows that they have the divergent attitudes they report. If so, they stand in a certain kind of conflict, even if their claims do not, at least in a straightforward sense. To defend Moore’s argument, we’d have to say that our intuitions about disagreement concern the semantic workings of the sentences uttered, rather than the looser notion of disagreement between the speakers involved. And one might think that our linguistic intuitions aren’t quite as fine-grained as that.
argument that descriptivist accounts of normative language cannot capture their essential endorsing function.

Third, Stevenson requires that a proposed account of “goodness” must not allow for questions of goodness to be “verifiable solely by the scientific method” (Stevenson 1937: 16). A proposed reformulation of “what is good?” fails this test if the reconstruction allows us to settle the question by means of empirical research alone. Consider Hume’s view as Stevenson understands it: to say that something is good is to say that most people approve of it. If Stevenson’s Hume were right, we could settle the question by taking a vote, and “this use of the empirical method, at any rate, seems highly remote from what we usually accept as a proof, and reflects on the complete relevance of the definitions which imply it” (Stevenson 1937: 17). That is, the allegedly Humean suggestion leads us to the conclusion that our reasoning about evaluative questions is seriously confused. One consideration in favor of the third requirement, then, is that it preserves an important feature of our methods of normative inquiry.

It’s worth pausing for a moment to examine this requirement, and Stevenson’s support for it, a bit more carefully. Consider the role of empirical considerations in evaluating normative claims, particularly those about goodness. Stevenson offers us the example of a person who knows that a majority approves of X, yet protests that, while this fact might be evidence that X is good, it doesn’t settle the question of X’s goodness. One can imagine an observer with an aristocratic bent simply disregarding facts about mass appeal as another indication of “the low state of people’s interests” (Stevenson 1937):

---

Notice, just for fun, that these three requirements parallel the three problems set out for the realist in Chapter 3. The disagreement requirement is the univocity problem; the magnetism requirement is the practical role problem; and the ban on purely empirical research is, more or less, the moral inquiry problem.
1937: 17). That is, one could acknowledge the empirical fact and reject the evaluation that’s said to follow, and this would not constitute a conceptual error or linguistic confusion. Empirical facts alone don’t seem to force us to particular evaluative conclusions. To use a formulation suggested by Stevenson’s example, one is always free to deny the relevance of the ostensibly decisive facts. What Stevenson doesn’t say, but implies, is that someone who denies this relevance is still making a judgment about what’s good. A speaker might be within the realm of conceptual competence despite having eccentric views about good-making features. Refusal to accept particular descriptive conditions for the application of “good” doesn’t rule that speaker out of the conversation. Stevenson’s requirement looks like one of the earliest articulations of the idea that evaluative concepts are essentially contestable. Very roughly, to say that a concept is essentially contestable is to say that there’s nothing about its rules for application that settle hard cases, even cases taken by some parties to be paradigmatic.

So, if moral concepts are essentially contestable, they permit all sorts of (substantive) latitude in their applications; that we’re using moral concepts doesn’t restrict us from any particular evaluative standard. This is not the place to examine the idea carefully, but it is

---

29 Immediately after making this point, Stevenson appeals to the popularity of the open question argument as further evidence for his third requirement. “No matter what scientifically knowable properties a thing may have,” Stevenson says, giving a gloss of Moore, “you will find, on close inspection, that it is an open question to ask whether anything having these properties is good” (Stevenson 1937: 18). The conclusion in both cases is that empirical facts by themselves don’t close normative questions.

30 Another hint at the idea comes from Stevenson’s later essay “The nature of ethical disagreement” (1949). There, he claims that it’s never certain that agreement in belief will lead to agreement in attitude: It is logically possible, at least, that two men should continue to disagree in attitude even though they have all their beliefs in common, and even though neither had made any logical or inductive error or omitted any relevant evidence. Differences in temperament, or in early training, or in social status, might make the men retain different attitudes even though both were possessed of the complete scientific truth. (Stevenson 1949: 591)

30 See also Foot (1967).

31 I’m indebted to Justin D’Arms for this way of formulating the idea.
worth noting that it is lurking in the background of the first generation of noncognitivism. It’s also interesting that the notion of essential contestability might be important to preserving the distinct role or function of moral predicates—if moral concepts weren’t permissive in this way, they wouldn’t be able to serve as a device for adjudicating certain disputes, perhaps because particular challenges couldn’t be brought up within moral discourse.\(^{32}\)

Having considered his requirements for a tenable account, let’s turn to Stevenson’s positive proposal. The central idea is that the purpose of moral claims is not “to indicate facts, but to create an influence” (Stevenson 1937: 18). Moral judgments are concerned with recommending something for approval or disapproval; and this involves something more than a disinterested description, or a cool debate about whether it is already approved, or when it spontaneously will be. (Stevenson 1944: 13)

If moral claims were just descriptions of our attitudes, they wouldn’t allow for real disagreement, as we’ve seen, and they couldn’t play their characteristic role in guiding action. Stevenson’s view is motivated by the need to account for these phenomena, and his approach to the problem relies on the idea that moral discourse is concerned with not the description but the shaping of interests and attitudes.

Now the task is to move from these remarks about what moral judgments do (or what function they perform, how they are used...) to a view about their meaning. Start with the sensible idea that some uses of language communicate beliefs, while others

---

\(^{32}\) A rough argument for this might go as follows. Suppose moral concepts did rule out certain evaluative standards as a conceptual matter. Then someone couldn’t coherently suggest that those standards should be taken up for making rightness-judgments; her proposal would be ruled out simply by the nature of the concepts involved. But these radical challenges should be possible within our moralizing, because one of the functions of these concepts is to provide a way of addressing these disputes. Hence, denying essential contestability leaves us with a view of moral concepts on which they’re unable to do some important work. See Gibbard (1993) for an interesting discussion of a point much like this one.
“give vent to our feelings...create moods, or...incite people to actions or attitudes” (Stevenson 1944: 21). Call this second class dynamic use—language is used dynamically when the speaker’s purpose is to express or influence attitudes other than beliefs. Of course, just about any sentence might be used dynamically, given the right context and tone of voice. Given the seemingly endless variations in dynamic use, it won’t be plausible to identify this as a term’s meaning. The solution, for Stevenson, is to abstract from particular uses by offering an account of meaning in terms of a word’s dispositional tendencies to bring about various mental states:

Instead of identifying meaning with all the psychological causes and effects that attend a word’s utterance, we must identify it with those that it has a tendency (causal property, dispositional property) to be connected with. This tendency must be of a particular kind, moreover. It must exist for all who speak the language; it must be persistent; and must be realizable more or less independently of determinate circumstances attending the word’s utterance...Moreover, we must include, under the psychological responses which the words tend to produce, not only immediately introspectible experiences, but dispositions to react in a given way with appropriate stimuli...Suffice it to say that ‘meaning’ may be thus defined in a way to include propositional meaning as an important kind. (Stevenson 1944: 22)

It’s important to stress that this causal view of meaning is meant to include both descriptive and emotive meaning: when the word’s causal tendencies have to do with various beliefs, then it has descriptive meaning; when it’s connected to other kinds of attitudes, it has emotive meaning, in roughly Ogden and Richards’ sense. In Ethics and Language Stevenson characterizes his view of meaning in this way:

---

33 It’s probably a mistake to insist that this tendency exists in all speakers of a language, because this rules out, say, affectively disabled members of the speech community. It’s more plausible to think that there’s some way of accounting for (individual cases of) attitudinal (or cognitive) deviance by making them parasitic on the mainstream practice. The details of this will be tricky, as we’ll see in discussing Simon Blackburn’s proposal later in this chapter.

34 In fact, Stevenson cites the passages I’ve quoted from The Meaning of Meaning as the source of the view he presents in this paper.
The meaning of a sign, in the psychological sense required, is not some specific psychological process that attends the sign at any one time. It is rather a dispositional property of the sign, where the response, varying with varying attendant circumstances, consists of psychological processes in a hearer, and where the stimulus is his hearing the sign. (Stevenson 1944: 54)

Emotive meaning is “a meaning in which the response (from the hearer’s point of view) or the stimulus (from the speaker’s point of view) is a range of emotions” (Stevenson 1944: 59). Thus emotive meaning will resemble a kind of generalized dynamic use.

With this view of meaning on the table, we can begin to look for a definition of (or an account of the meaning of) sentences such as “X is good.” As Stevenson points out, an exact equivalent can’t be provided, but we can get a rough approximation by considering the dynamic use of “We like X.” The descriptive meaning of this sentence is simply a psychological report, but the sentence in question can be used to do more than convey our attitudes. Claims of this sort often suggest that, if we don’t already, we had better start liking X, as when a child is told that we like to keep our rooms clean. So it has a dynamic use as well, because it’s used to influence attitudes as well as to report on them. Now we’ll need a way of accounting for this influence in terms of the meaning of these sentences.

A natural place to start is with imperatives, since these have a practical role like that of ethical claims. Both of these, Stevenson thinks, “are used more for encouraging, altering, or redirecting people’s aims and conduct than for simply describing them” (Stevenson 1944: 21). This similarity suggests the following rough “working model” for analysis: the sentence “X is good” is equivalent to “I like X; do so as well” (Stevenson

It’s also important to keep in mind that Stevenson’s view of attitudes understands them to be dispositional, not merely occurrent, states.

35 What to make of the disjunction here? That is, “express or evoke” isn’t completely unproblematic, since the two can come apart.
1937: 25; Stevenson 1944: 21). The first part of this gives us a gloss on the descriptive meaning of “X is good,” and the second, imperatival component is a tool for “changing or intensifying the attitudes of the hearer” (Stevenson 1944: 22). The use of an imperative here distinguishes moral claims from those of science and helps to preserve “the hortatory aspects of ethical judgments” (Stevenson 1944: 32). The action-guiding or commending function of ethical terminology can be captured, in rough terms, by thinking of ethical claims as commands to have particular attitudes.36

This, Stevenson thinks, gets something right, but it’s too crude. The imperative is “too blunt an instrument to perform its expected task” because if a person is explicitly commanded to have a certain attitude, he becomes so self-conscious that he cannot obey. Command a man’s approval and you will elicit only superficial symptoms of it. But the judgment, ‘this is good,’ has no trace of this stultifying effect; so the judgment’s force in encouraging approval has been poorly approximated. (Stevenson 1944: 32)

Invoking an imperative is helpful because this emphasizes the action-guiding or dynamic character of ethical claims. But Stevenson objects to this initial proposal on the plausible grounds that it’s not nuanced enough to capture what’s going on in moral discourse. Instead, ethical terms “must be explained with reference to a characteristic and subtle kind of emotive meaning,” where this, again, is “the power that the word acquires, on account of its history in emotional situations, to evoke or directly express attitudes, as distinct from describing or designating them” (p. 33). Emotive meaning is meant to do roughly the same work as the appeal to imperatives, but in a less explicit way:

36 One interesting feature of Stevenson’s view is that the imperatives are directed toward attitudes rather than actions, as they are for Ayer and for Hare. According to the first approximation of the first pattern of analysis, to call something good is to command others to approve of it. As we’ll see, this is quite like Allan Gibbard’s norm-expressivism, which also appeals to the idea of imperatives directed toward mental states.
In virtue of this kind of meaning, ethical judgments alter attitudes, not by an appeal to self-conscious efforts (as in the case with imperatives) but by the more flexible mechanism of *suggestion*. Emotive terms present the subject of which they are predicated in a bright or dim light, so to speak, and thereby *lead* people, rather than command them, to alter their attitudes. (Stevenson 1944: 33)

The notion of emotive meaning is thus meant as a way of saving what’s insightful about the appeal to imperatives while adding the missing subtlety. Emotive meaning allows speakers to change others’ attitudes by suggestion, rather than command, and it directs the audience’s attention to the object of evaluation, rather than to the speaker’s attitudes towards that object. This alteration of the first pattern of analysis sees moral claims as exerting a kind of causal pressure, via suggestion, on their audience, where the first suggestion saw them as ordering compliance with the speaker’s attitudes.

Later, I’ll suggest that the imperatival model has an advantage over the refined account when it comes to preserving an interesting feature of moral disagreement. Thus it’s important to see these as different—though related—proposals. This reading gets some support from the way that Stevenson discusses the role of emotive meaning in his account: he notes that the imperative “do so as well” is “much too crude,” and begins his chapter on the first pattern of analysis by wondering “how we can dispense with the overt imperative that was used in the working models, recognizing a more appropriate emotive meaning in its place” (Stevenson 1944: 81). Ethical terms “have a quasi-imperative function which, poorly expressed by the working models, must be explained with careful attention to emotive meaning” (Stevenson 1944: 36). It looks as though the developed first pattern of analysis is meant to supersede the working model introduced for the sake of illustration.
The difference between the working model and its successor has to do with just how attitudes enter into the account. On the working model, interlocutors are commanded to share the speaker’s attitudes by means of an imperative, while in the fully developed account the attitudes of the speaker are spread by suggestion and the evocative power of emotive meaning. The attitudes themselves, though, are a fixed point in the transition from one linguistic view to the other, and so they are the basis of the points of agreement between the two proposals. One of these concerns the identification of moral language by appeal to particular mental states. Stevenson thinks that the ethical use of “good” can be marked out by appeal to distinct sentiments or feelings:

When a person likes something, he is pleased when it prospers, and disappointed when it doesn’t. When a person morally approves of something, he experiences a rich feeling of security when it prospers, and is indignant, or ‘shocked’ when it doesn’t. (Stevenson 1937: 26)

There’s a similar appeal in *Ethics and Language*:

The peculiarly moral attitudes, associated with the moral senses of the ethical terms, are not easily described, but can roughly be marked off in this fashion: It will be recalled that an attitude is a disposition to act in certain ways and to experience certain feelings, rather than itself a simple action or feeling. If we wish to distinguish one sort of attitude from another, then, we can proceed by specifying the different sorts of response that attend typical stimuli...Suppose that a man morally disapproves of a certain kind of conduct. If he observes that conduct in others, he may then feel indignant, mortified, or shocked; and if he finds himself given to it, he may feel guilty or conscience-stricken. But suppose that he dislikes this conduct, as distinct from morally disapproving of it. He may then be simply displeased when he observes it in others, and simply annoyed with himself when he finds that he is given to it....These differences in response given similar stimuli, help to distinguish the attitudes which are moral from those which are not. The full distinction, should occasion require it, could be made by supplementing these remarks by others of a similar kind. (Stevenson 1944: 90).

---

37 Stevenson’s imperatives, unlike Hare’s, are directed to attitudes instead of actions, and so he can mark out moral utterances by appeal to the sentiments they command. As we’ll see, this foreshadows some of Gibbard’s suggestions.
Stevenson, like Barnes, thinks that the distinction between moral and non-moral evaluations can be drawn using the resources of sentiment: the difference between dislike and moral disapproval is found in the different responses they involve. An attitude, imperative, or judgment qualifies as *moral* in virtue of its ties to particular sentiments.

This passage also highlights another significant feature of Stevenson’s proposal, the relationship between moral judgments and actions. Recall that an attitude of the relevant sort is partly constituted by various dispositions to act and feel in certain ways. To influence your attitudes is to affect the way you act in the relevant circumstances. If ethical terms are “instruments used in the complicated interplay and readjustment of human interests” (Stevenson 1937: 20), then they are tied to action by the practical nature of attitudes, which include dispositions to act in particular ways. But attitudes and actions are also tied together in more complex ways. As we saw in our discussion of Barnes’ (1933) early suggestion, it’s not clear how a simple construal of attitudes would allow them to conflict in ways that are robust enough to ground real disagreement. Stevenson’s notion of attitudes as complex clusters of dispositions helps to flesh out the noncognitivist’s story about disagreement, because it offers a more complete account of the relationship between, say, your actions and my attitudes. My attitudes are constituted not only by dispositions to act in certain ways (say, when I find myself in your circumstances) but by dispositions to feel various sentiments (e.g. indignation, shock) in *response* to your actions. Presumably these sentiments are also tied to the other ways I respond to you, either by action or by forming additional attitudes. This more complex

---

38 It’s important to be clear that the imperatives of the first approximation are directed not to actions but to attitudes. As I’ve noted, though, these two are connected by the action-guiding nature of attitudes.
picture suggests ways that the noncognitivist might capture a more subtle notion of
disagreement by appeal to a richer conception of the attitudes that undergird moral
discourse.

With that thought in mind, let’s turn to Stevenson’s discussion of disagreement to see if this promise can be made good. Just how will Stevenson’s idea of emotive
meaning allow him to account for disagreement, as his first criterion demands? Plausibly
enough, we begin with a distinction between disagreement in belief and disagreement in interest (or, in his later work, attitude). Disagreement in belief is straightforward enough: it occurs when, say, A believes that p and B disbelieves it (though Stevenson seems to be
at odds with himself on this, as I note below). Disagreement in interest or attitude occurs
when A has a favorable interest in X, while B has an unfavorable interest in it. Or, as a
slightly later Stevenson explains the distinction,

The two kinds of disagreement differ mainly in this respect: the former is concerned with how matters are truthfully to be described and explained; the latter is concerned with how they are to be favored or disfavored, and hence with how they are to be shaped by human efforts. (Stevenson 1944: 4)

It’s noteworthy that having diametrically opposed attitudes isn’t sufficient for
disagreement in attitude, as Stevenson understands it. Disagreement requires something
more: there must be some interest in resolving this opposition. Speakers don’t have
disagreement unless “neither is content to let the other’s interest remain unchanged”
(Stevenson 1937: 27). The same idea occurs in Ethics and Language: “Two men will
be said to disagree in attitude when they have opposed attitudes to the same object...and

---

30 The requirement of a motivation to resolve the opposition isn’t accidental to (1937). For example: “Two men will be said to disagree in attitude when they have opposed attitudes to the same object...and when at least one of them has a motive for altering or calling into question the attitude of the other” (Stevenson 1944: 3).
when at least one of them has a motive for altering or calling into question the attitude of
the other” (Stevenson 1944: 3). Opposed attitudes are not enough: people may neither
agree nor disagree, as when “they simply ‘differ,’ having divergent beliefs or attitudes
without a sufficient motive for making them alike” (Stevenson 1944: 5).\(^4\)

This is a strange claim: why aren’t the attitudes in themselves sufficient for
disagreement? Why is this extra element, the motive to come to consensus, required for
disagreement in attitude but not disagreement in belief? Looking at these questions will
help us to see some interesting problems with Stevenson’s account. In this section I’ll
take up two broad points about the nature and location of disagreement in belief and in
attitude.

The first point takes up the question of why we’d need something \textit{apart} from our
moral attitudes in order to have disagreement. Let’s start by thinking about the case of
conflicting belief. Suppose you believe that p and I believe that -p.\(^4\) Here, we appear to
have a disagreement \textit{regardless} of any additional desire to sort things out. The motive to
come to agreement is superfluous: you might think I’m too obstinate, maybe it’s not

\(^4\) There’s a small but exciting ambiguity in Stevenson’s treatment of this issue. In (1937) he says that
disagreement in belief occurs simply “when A believes p and B disbelieves it,” while disagreement in
attitude requires, furthermore, that the parties be interested in coming to share attitudes. Yet in (1944) he
says that speakers may \textit{differ} rather than \textit{disagree} when they have attitudes or beliefs that diverge while
failing to have the motivation to change this. The question is whether or not the motive to converge is
required in the case of belief.

Why does this matter? If the convergence motivation isn’t required for disagreement in belief, yet
is required for disagreement in attitude, there’s a troubling disanalogy. Speakers with contrary beliefs
disagree even if they don’t care to talk about it, yet, by Stevenson’s lights, this isn’t so in the case of
attitude—which is counterintuitive. If, on the other hand, the motivation to resolve things is needed in both
cases, then we might think that Stevenson has given us an account of when \textit{people} disagree, though he’s
failed to give us a story about how their ‘judgments’ disagree or conflict (I mean ‘judgment’ there in a
sense that’s neutral between beliefs and attitudes).

\(^4\) Assume that p is some straightforwardly truth-apt claim that doesn’t involve vagueness.
worth the effort—or you might have strategic reasons to want me to have false beliefs.\textsuperscript{42}

We might agree to disagree, as the saying goes, which would seem to indicate that we recognize the incompatibility in our positions while endorsing a commitment to leave it be. But we still conflict in some robust way—at least, our beliefs do—because we have different answers to the question of whether or not it’s the case that $p$, and we can’t both be right.\textsuperscript{43} This certainly seems to constitute legitimate disagreement.

What’s more, just the same points can be made about moral disagreements: there, as in straightforward disagreement in belief, contrary views are sufficient for robust disagreement, and there’s no need for any additional commitment to convergence. If so, then Stevenson has a problem, because he’s relying on some other mental state to account for properties that moral judgments have in and of themselves. On this account, real disagreement between people who assent to “it’s wrong to $\Box$” and “it’s obligatory to $\Box$” is contingent on the presence of some further desire or motivation. Strictly speaking, according to Stevenson, the speakers conflict only because of this attitude that’s external to moral judgment—and, likewise, their utterances aren’t in tension with one another.\textsuperscript{44}

\textsuperscript{42} There’s a potential problem here, because we shouldn’t assume that the relevant motivation is overriding. It’s harder to argue that there needn’t be some pro tanto motivation to reach agreement.

Note, too, that Stevenson himself mentions some of these features in explaining how people might differ in their attitudes without having the motive to resolve things. See (1949: 111).

\textsuperscript{43} When all else fails, appeal to ordinary language: “Steve and I discovered that we disagree about...” sounds perfectly natural. The locution suggests that disagreement exists prior to discovery, and hence any motivation to come to resolution. This argument is lame insofar as one might think that along with the belief that $p$ I have the tacit commitment to convince others that $p$ on the occasion of discovering that they think otherwise. The rejoinder is feeble, though, because clearly Stevenson thinks of the motivation to settle things as being only contingently attached to beliefs (and attitudes). That is, it isn’t simply part of what it is to believe that $p$ to have a commitment of this sort. If so, then it’s possible to have conflicting beliefs without really caring, and it’s just as permissible to say that we disagree (and that we discovered this) in a case like this.

Granted, people sometimes refer to their incompatible commitments as differences rather than disagreements. This seems like mere politeness.

\textsuperscript{44} Note that this is true even if the utterance straightforwardly expresses the attitude involved in moral commitment. The utterance, presumably, would express that attitude, not the apparently independent one.
The best response seems to be a slight revision in Stevenson’s conception of an attitude. There’s no need to make the implausible claim that disagreement requires this additional, apparently superfluous motivation if the attitudes involved in moral judgment and commitment are sufficient. Attitudes of approval, after all, require more than my own preference for, say, performing an action; they involve thinking that other people should share this commitment, too.⁴⁵ If I approve of A-ing in circumstances C, I’m committed, all other things being equal, to A when I find myself in C, but I also have something like a desire that others A when in C. If A-ing’s not for you (either you have a preference to refrain or a con-attitude toward it), then you and I have incompatible attitudes, because my approval of A-ing involves a desire that you act a certain way, and you’re committed to acting in some other way. When considered this way, our contrary attitudes ensure that we can’t both be satisfied, as Stevenson puts it, apart from any further motive to come to consensus.⁴⁶ This at least gives us the result that there’s some kind of conflict (at least the conflict of mutually unsatisfiable desires) in cases of moral

⁴⁵ See for example Pitcher’s (1958) account of approval. Pitcher thinks that approval involves judging that it’s right to favor the object of one’s approval, but I don’t need such a contentious analysis as this. My point is simply that approval is a richer notion than, say, preference, in that it has some bearing on other people’s attitudes as well.
⁴⁶ Mark van Roojen (1996) notes that expressivist attempts to come to terms with the Frege-Geach problem fail because they require a notion of inconsistency that goes beyond logical inconsistency, and this expansive notion is bound to produce inconsistency where there isn’t any. The issue of the nature of our disagreement raises similar questions, because, to come to terms with this issue, we’ll have to say just what kind of disagreement occurs in cases like this.
disagreement, and the welcome conclusion that this conflict exists solely because of the moral attitudes involved.\(^47\)

The conflict that exists between us when we have contradictory moral views is, by Stevenson’s lights, the conflict of mutually unsatisfiable desires or preferences: your disapproval of A-ing involves preferring that I refrain from \(\neg\)A-ing, and my approval involves having at least a *pro tanto* desire (or preference, or...) to A. It’s interesting that this kind of conflict shares several features with the prototypical case of disagreement, in which we have incompatible beliefs. Inconsistent beliefs cannot all be true; inconsistent desires cannot all be satisfied. (To complicate needlessly: there is no possible world where all of the propositions picked out by the beliefs in the inconsistent set are true, and there is no possible world where all of the inconsistent desires are satisfied). In the case of belief, I couldn’t take on your belief while maintaining my own unless I abandoned theoretical rationality by believing p&\(\neg\)p. Doing so would ensure that I have at least one false belief, and, since the *point* of belief is to get things right, or to present things as they are, assenting to the contradiction would go against the purpose or role of having beliefs. In the case of attitudes, to take on your preferences in addition to my own would ensure that I have preferences that cannot be satisfied; the same state of affairs that satisfied one would frustrate the other. Just as my contradictory beliefs left me unable to get things right, my contradictory attitudes leave me sure to get less than what I want, whatever

\(^{47}\) Some of Stevenson’s examples seem artificial because they involve, say, two people deciding where to have dinner, or deciding whether to see a film or a concert. This might suggest that he’s illegitimately constraining us to cases of collective activity, where, say, the disputants are committed to doing something together and then need to figure out just what. But construing attitudes in the way I’ve suggested makes other sorts of cases unproblematic.
happens. When these desires are directed toward my own actions, I’ve got an analogous case of practical irrationality: I can’t be guided by desires that pull in opposite directions, so I’ve managed to make the role or function of having desires unmanageable.

There are also differences between the two cases. Beliefs, unlike desires or attitudes, can be measured against the world they represent, and so there’s an independent standard we can use to assess them. This might suggest that the points about rational incompatibility noted above are secondary to the primary conflict between contradictory beliefs—namely, that the propositions at issue cannot both be true. Either p or -p must be false. And it’s because of this that a rational agent cannot be committed to both claims. Attitudes have no such requirement, since they’re not the kinds of things that can be true or false, and so the incompatibility of competing attitudes can’t rely on this standard.

This raises an interesting wrinkle in cases of conflict, where an agent finds herself with contradictory beliefs or mutually unsatisfiable desires. In both cases, it seems, revision is needed. But there are different constraints on which revision is best. Beliefs will of course be revised in light of an agent’s wider doxastic commitments, because it makes sense to reject either p or -p depending on which one coheres less well with our other beliefs. There’s the additional possibility that the coherence-maximizing revision is not the truth-conducive one, if one’s beliefs are sufficiently out of line—that is, the agent could still be wrong in believing that p even if p coheres better than -p with her other beliefs. Desires (or other noncognitive mental states) don’t face the possibility of this additional breakdown, because there’s nothing more than coherence to motivate change

48 More optimistic readers may take some consolation in the thought that contradictory preferences ensure that I get at least some of what I want.
in commitment. That is, there are various strategic considerations that come into play when we choose which desire to (try to) reject, but these seem to settle the issue and leave no room for additional error.

This is just another way of saying that beliefs and not desires can be false, but it helps us to see a problem for Stevensonian accounts of moral discourse. The worry is that this sort of account seems ill-suited to explain our attachment to our own evaluative positions and our subsequent opposition to the conflicting attitudes of others. Thinking about our evaluative commitments in terms of noncognitive attitudes shows us how we have an incentive to come to agreement.\footnote{Strictly speaking, we want attitudes that mesh together in the right ways, rather than agreement in the sense of attitudes shared by all. See Gibbard’s (1990) criticisms of Adam Smith.} Things go better for us, collectively and individually, when our attitudes fit with those around us, and when they do we’ve made progress toward coordinating our actions in useful and beneficial ways. We might think, though, that there are many equally efficient ways to solve these problems, because what really matters is that we agree on some strategy rather than any one in particular. What matters most is agreement, rather than the content of that agreement (given appropriate restrictions). Once attitudes are meshed in the right ways, there’s no additional standard, such as truth, to which they can be held accountable. Yet, if this is right, and if moral disagreement, at bottom, is grounded in conflicting attitudes, it seems we’re left with an explanatory debt. Moral agents have an attachment to their evaluative commitments that often leaves these commitments recalcitrant in the face of various sorts of pressure. Our
reluctance to change our attitudes when faced with others with incompatible commitments is merely obstinate if there’s nothing for these attitudes to get right.\textsuperscript{50}

The second point concerns the location of the disagreement. In a simple case where you believe p and I believe -p, we’re in conflict because the propositions we believe are incompatible; they can’t both be true. The incompatibility of our beliefs rests on the incompatibility of the propositions that are their objects. By extension, the same point can be made about the sentences that express these propositions. One way of putting this is to say that the disagreement is semantic: the sentences we utter in expressing our beliefs are in tension with one another (or contradict one another?), and this is true in virtue of what they mean, independent of any further desire to resolve or ignore that incompatibility. It’s not just the speakers who are in conflict; it’s what they say. If we think that moral language exhibits this feature (that is, if we think that sentences used to express conflicting moral views are themselves in conflict), then there’s an additional requirement on Stevenson’s account. If he’s to respect our semantic intuitions, he must not only show us how attitudes can conflict, in and of themselves, without some additional motive toward consensus; he must also show that his account of language is able to preserve this conflict in the appropriate way.

This is potentially problematic because there are cases in which speakers disagree, and their disagreement is manifested in what they say, though their sentences are, strictly speaking, compatible with one another. Consider a case of disagreement in ‘speaker meaning’ without a corresponding incompatibility in ‘semantic meaning.’ The fastidious

\textsuperscript{50}Of course, there are a variety of psychological explanations one might give to make sense of this commitment. If those explanations are successful, there’s room for a satisfying Stevensonian reply.
department chair says, imperiously, to his slovenly junior colleague, “full academic dress will be worn,” to which the hapless assistant professor replies, “it’s going to be very hot at commencement.” These sentences aren’t literally in conflict, since they’re probably both true, but there’s a clear sense in which the speakers involved in this conversation disagree about what ought to be worn, and they say what they do in order to influence the other’s choice of costume. The point of this example is just to show that speakers with a real disagreement (say, about how the untenured should dress) can manifest this without uttering sentences that are, as a matter of meaning, in tension with one another. The upshot for Stevenson is roughly this: it’s not enough that his account of disagreement make sense of how there’s a form of conflict between moralizers with (apparently) opposing views. He must also provide us with an account of how the sentences that express this disagreement are in conflict.

Ironically, it may be that Stevenson has to admit that this is a problem if he wants to hold on to his criticism of subjectivism. His complaint against that view, of course, is that is fails to allow for disagreement. His accusation is right in one sense, because, according to the view, the sentence “stem cell research is wicked” in your mouth and “stem cell research is morally required” in mine can both be true, since—by the subjectivist’s lights—we’re making claims about our own attitudes toward the research. But in most cases the truth of these sentences, construed as the subjectivist suggests, will indicate or reveal a conflict: you report a con-attitude, while I report a pro-attitude. If we’re sincere and reasonably authoritative about our own attitudes, we’ll be correct: we

---

51 This sentence, which looks indicative but performs an imperatival function, comes from Winston H.F. Barnes.
really will have clashing attitudes. In order to maintain that subjectivism’s account of disagreement is faulty, Stevenson must require more than conflicts in attitude, which subjectivism can supply. He must also demand that the disagreement occur at the level of sentences or utterances.  

This thought might look suspiciously question-begging, because it might be thought to rule out any account of moral language that sees its claims as failing to express propositions. If the only way to meet this requirement is to show that competing moral claims (e.g., “it’s morally wrong to []” and “[[]] - ing is morally permitted”) are in logical conflict, then expressivism is out in the cold, because expressions of sentiment don’t have truth-values and so aren’t the kinds of things that can meet this demand. But the objection isn’t quite right, for two reasons. The first is that the argument rests on independent linguistic data; if expressivism can’t account for them, so much the worse for the theory. Second, we haven’t assumed that moral sentences must express propositions in order to meet this requirement, because we’ve granted that conflicting imperatives might be contradictory in the required sense. Were Stevensonian moral utterances to have the logical form of imperatives—that is, if it were more like Hare’s view—this puzzle would be solved easily: if I command you to [], and you command me to refrain from []-ing, our sentences are in the right sort of tension (of course, this is true in the case where what’s commanded is an attitude or sentiment rather than an action). But, as we saw, Stevenson rejects this working model in favor of something more sophisticated. The question now is what the fully developed “first pattern of analysis”

---

52 See Smith (1986).
53 See van Roojen (1996).
54 It’s not even clear that imperatives express states of mind or attitudes.
says about the conflict between *sentences* that express competing moral attitudes. I’ve suggested that it’s not enough for the account to show a conflict between *speakers*; it must also preserve a certain relationship between (some of) the sentences that are used to manifest this conflict. We now need to determine whether or not Stevenson’s developed view meets this challenge.\(^{55}\)

Whether it does or not is a delicate question. If speakers, in making moral claims, are *expressing* attitudes that are in some form of robust conflict, we might think that the utterances themselves are in conflict. After all, speakers who have contradictory beliefs express them with utterances, declarative sentences, that meet the test we’ve been discussing. But this is too quick, because these same beliefs can be expressed by sentences that *aren’t* in literal conflict: the disagreement between us that might be voiced with claims such as “p” and “¬p” can also take a different conversational form, for example, sentences like “I think that ¬p;” “well, I heard that p.” Since *this* exchange manifests a disagreement between speakers via a pair of claims that are consistent, it’s not sufficient to say that evincing or expressing conflicting attitudes gives rise to conflicting sentences. That argument fails in the case of belief, and there’s no reason to think it will be more successful when it comes to noncognitive attitudes.

Complicating matters still further, Stevenson gives us a number of different ways to think about the relationship between a moral utterance and its underlying attitudes. He

\(^{55}\) It’s interesting to note, in passing, that some of Ayer’s later views invoke a similar strategy. For example, in (1949) he considers the relationship between the descriptive features of a situation and the moral claims that are said to follow from, or to be justified by, those natural facts. He thinks that what are accounted reasons for our moral judgments are reasons only in the sense that they determine attitudes. One attempts to influence another person morally by calling his attention to certain natural features of the situation, which are such as will be likely to evoke from him the desired response. (Ayer 1949: 237)
says that emotive words are “suitable for ‘venting’ the emotions” (Stevenson 1949: 38), as, for example, a laugh or cry might be, with the difference that their suitability for the task depends on conventions of use. Emotive words “are fitted both to express the feelings of the speaker and to evoke the feelings of the hearer” (Stevenson 1949: 41); emotive meaning is a meaning where “the stimulus (from the speakers’s point of view) is a range of emotions” (Stevenson 1949: 59). Part of the meaning of a sign (in this case, an evaluative term) is a disposition to be used in conjunction with certain psychological processes (presumably, the relevant attitudes), and this disposition is grounded in a correlation between those processes and the use of the sign (Stevenson 1949: 57). The problem is that it’s not clear how these relationships will ground the kind of conflict we need to preserve.

We can see at least grounds for suspicion about meaning-disagreement in Stevenson’s account of moral conversation. In uttering competing moral claims, he thinks, we’re invoking words with causal powers to change attitudes, and thus we’re putting pressures on our disputants’ attitudes (and, by extension, their commitments to action and emotion in various circumstances). Our interlocutors, in turn, apply the same sorts of pressure to us. And these utterances usually stem or arise from roughly the same attitudes in the speaker (they vent the attitude, which is the utterance’s stimulus). But there’s nothing here that resembles incompatible content. The pressures that exist are between speakers, and the incompatibility isn’t in the right place, namely, the sentences of moral conversation. The terms involved (or, more accurately, the sentences) have causal dispositions that pull in opposite directions, but this isn’t quite the same thing.
Suppose that Stevenson did accept the challenge, thus admitting that he needs to provide a view of meaning according to which particular moral sentences are in some sort of conflict. One response is to rest on the idea that moral claims express incompatible attitudes, and thus are in the right sort of relationship when they express those attitudes (or are tied to them in some other appropriate way), as a matter of meaning. In an analogy with the case of belief, I denied that sentences expressing incompatible mental states were always incompatible—rightly so, if we take a fairly broad view of what it is to express a mental state. Some declarative sentences suggest that \( p \) even if they don’t, strictly speaking, say that \( p \). If we count that suggestion as expressing the belief that \( p \), then my earlier point is right. But in the case of emotive meaning, and in direct expressions of moral attitude (i.e., “\( p \)” rather than “I think that \( p \)”), the conflicting attitudes are tied to the terms that express them by the bonds of meaning. Sentences may indicate conflict in virtue of implicature, and so may fail to be genuinely incompatible. But if they indicate this conflict in virtue of their meaning, then the sentences should be taken to be in semantic conflict—and this is true for emotive as well as descriptive meaning.

Another response, of course, involves denying the challenge. Stevenson can say, with some plausibility, that our linguistic intuitions aren’t so fine-grained as to discriminate between different locations for disagreement, moral or otherwise. As long as he’s ensured that moral attitudes conflict, he’s done enough to capture disagreement. This may require giving up his argument against subjectivism, as we noted. In addition, it rests on a suspicious claim about our linguistic intuitions. We commonly note the distinction between what’s directly said and what’s merely hinted at, implied (in the
ordinary-language, not logical, sense), suggested, and so on.\(^5\) This, in conjunction with our mastery of negation, is enough to suggest that our ordinary linguistic understanding contains the resources needed to get a notion of semantic disagreement off the ground. Stevenson’s view, then, is not adequate.

For our purposes here, the lessons to be learned aren’t directly about Stevenson, but about the prospects for a satisfying noncognitivism. If we take the ‘location of disagreement’ argument seriously, we’ll need to provide a way for moral utterances, as well as moral speakers, to stand in the right sorts of conflict. Attitudes may serve to explain conflict between speakers, but, as Stevenson’s example shows, it isn’t obvious just how to use them to account for the relationship between (apparently) contradictory sentences. The failure of Stevensonian emotive meaning, and the larger failure of the causal account of meaning generally, is a mark in favor of views that rest on imperatives, which, as we’ve seen, can stand in the right sorts of relationships with (at least some) other sentences. We’ll turn next to Hare’s prescriptivism, which capitalizes on this advantage. Our discussion of problems for this view will reveal ways in which attitude-driven accounts provide useful resources for the noncognitivist. As we’ll see, these competing motivations are synthesized in Allan Gibbard’s norm-expressivist analysis of moral judgment.

2.7 Hare

Hare’s work is interesting for two reasons. First, his criticisms of both noncognitivism and descriptivism will help us to see problems that must be solved if

---

\(^5\) E.g., when children are at the rule-fetishist stage of linguistic development, they invoke this feature to avoid the spirit, if not the letter, of poorly-formulated prescriptions.
either of these views can be made more palatable. Like Stevenson, Hare is useful in
focussing our attention on important features of moral discourse that must be respected
by any plausible meta-ethical account. Second, his own view offers a powerful way of
making sense of these phenomena while heading off the objections that plague earlier,
less sophisticated suggestions.

Let’s begin by looking at his critical remarks on the expressivist views we’ve
been considering. Hare offers a few different criticisms of ‘attitudinal’ accounts of moral
discourse in the early sections of *The Language of Morals*. The first comments address
views that attempt to “‘reduce’ imperatives to indicatives” in the manner of the
subjectivist view that evaluative claims are in fact statements about the mental states of
the speaker (1952 p. 3). These criticisms seem to offer a fairly conventional
objection—they don’t allow for disagreement or contradiction, and they seem to change
the subject:

But there remains the difficulty that the sentence ‘shut the door’ seems to be about
shutting the door, and not about the speaker’s frame of mind, just as instructions
for cooking omelets (‘Take four eggs, &c.’) are instructions about eggs, not
introspective analyses of the psyche of Mrs. Beeton. (Hare 1952: 6)

So far, this isn’t much beyond Moore’s criticism. But there’s another, more subtle
problem contained in this critique, one that will affect views that are more plausible than
crude subjectivism. This has to do with the mental states that are either being expressed,
or being referred to, by evaluative utterances. Hare’s point is that these mental states
can’t be part of an analysis of moral language because the moral concepts they purport to
explain are conceptually prior to the states invoked in the analysis. We get the first
suggestion of this in *LM* 1.3:
'I approve of A' is merely a more complicated and circumlocutory way of saying 'A is right.' It is not a statement, verifiable by observation, that I have a recognizable feeling or recurrent frame of mind; it is a value-judgement; if I ask 'Do I approve of A?' my answer is a moral decision, not an observation of introspectible fact. 'I approve of A' would be unintelligible to someone who did not understand 'A is right,' and the explanation is a case of obscurum per obscurius. (Hare 1952: 6-7)

A clearer statement of the same objection appears slightly later:

The only safe way of characterizing the attitude which is expressed by a universal imperative is to say ‘The attitude that one should not (or should) do so and so’; and the only safe way of characterizing the attitude which is expressed by a moral judgment is to say ‘The attitude that it is wrong (or right) to do so and so.’ To maintain an attitude of ‘moral approval’ towards a certain practice is to have a disposition to think, on the appropriate occasions, that it is right; or, if ‘think’ itself is a dispositional word, it is simply to think that it is right...But there is in all of this nothing to explain just what one thinks when one thinks that a certain sort of act is right....Sentences containing the word ‘approve’ are so difficult of analysis that it seems perverse to use this notion to explain the meaning of moral judgments which we learn to make years before we learn the word ‘approve’... (Hare 1952: 11-12)

The charge here is that the attitudes invoked in these analyses are constituted partly by the judgment they seek to explain. If we’re attempting to account for moral meaning in independent (non-moral) terms, this won’t do; a circular analysis is unilluminating.

Hare’s point is well taken: if the expressivist seeks to explain moral discourse in non-moral and naturalistic terms, then she will need to appeal to resources that are independent of the phenomena in need of explanation. At the same time, we might think that Stevenson’s remarks on attitudes at least make some progress toward explicating, say, what it is to have an attitude of moral approval, without appealing to any of the forbidden judgments brought to mind by Hare’s objection. The task for noncognitivism is twofold. First, we’ll need to specify mental states that track our moral judgments closely enough to be plausible analysans of them. Second, we’ll need accounts of those
states that don’t invoke the judgments these states are meant to explain. It’s crucial to the noncognitivist’s project that there are plausible candidates meeting both these requirements. Our concerns at the end of the discussion of Stevenson’s views are relevant here as well, because those worries concern the complexities of the mental states at issue in his account and raise the possibility that his attempts to ward off a challenge like Hare’s may not be completely successful.

Let’s turn now to Hare’s criticisms of naturalist and descriptivist views. It’s helpful to understand Hare’s general complaint through the lens of Stevenson’s ‘relevance’ criterion. Stevenson, of course, thought that any plausible reconstruction of ‘good’ would have to allow us to do most of what we wanted to do with the original term. Hare’s arguments against the descriptivist revolve around the same point. If the descriptivist were right, he contends, we couldn’t use a term, say, “good,” to do what we’d like to do with it, namely, to commend things. Suppose we agreed that, in order for a strawberry to be a good one, it must be sweet, juicy, and large. Having arrived at a consensus as to the good-making features of strawberries, we might be tempted to say that what it means to call a strawberry good is simply to say that it has these features. This won’t do, Hare thinks, because if this is just what “good strawberry” meant, then we couldn’t commend a strawberry for having those features. Saying that a strawberry is good because it has these features would be a trivial claim, rather than a substantive and informative one. And this just isn’t what we mean to be doing in calling the strawberry good. So, Hare concludes, “the proposed definition would prevent our saying something that we do succeed in saying meaningfully in our ordinary talk” (Hare 1952: 85-86). This, Hare thinks, is the point Moore was getting at with his ‘open question’ argument;
it’s also a point that Duncan-Jones suggests in his attempt to motivate expressivism. The descriptivist has missed something essential to our use of normative terminology.\footnote{Ayer (1984) seems to suggest that this kind of point can be made about any proposed set of synonyms, and so it’s nothing more than the paradox of analysis. At the same time, he acknowledges that this problem seems especially pressing in the case of normative terminology, perhaps because of the role or function of this sort of language.}

This failure has important ramifications. If calling something “a good X” were equivalent to saying that a thing is an X and had some further features, it would be a bit mysterious how to keep the meaning of “good” constant across different sortals. There are good chronometers and good people, but the features in virtue of which these things are good are quite different, and, were “good” to be understood descriptively, we’d have a puzzle as to how it could be a single predicate. The phrase “good chronometer” has descriptive meaning, of course, and this is different from the descriptive meaning of “good man,” but these are secondary to the evaluative or prescriptive meaning that these uses share, and which allows us to identify them as uses of the same term. Descriptive meaning alone is unable to account for this commonality (see Hare 1952: 118-119).

More importantly, descriptive meaning is unable to preserve the univocity of “good” (or related evaluative terms) across different communities. Suppose we have two different speech communities with quite different moral standards who both use “good” as their most general adjective of commendation. Despite their evaluative differences, we’d nonetheless think that, were they to encounter one another, they would be able to engage in meaningful dispute about what makes for a good person. Suppose that “good” is to be understood descriptively. If the descriptive meaning of “good person” is determined by the evaluative standards at work in these communities, as Hare thinks it is, then the two groups’ terms have different descriptive meaning, and so the interlocutors
are simply talking past one another (Hare 1952: 148-9). It’s only because of the evaluative meaning of terms such as “good,” Hare thinks, that we’re able to say that these people have a real, and not merely verbal, disagreement. This point is clearest in this famous articulation of the argument. Hare asks us to imagine a Christian missionary who lands on an island inhabited by cannibals, who, coincidentally, speak a language in which the most general adjective of commendation is “good.” If the meaning of this term is understood in terms of its descriptive criteria for application, he says,

then when the missionary said that people who collected no scalps were good (English), and the cannibals said that people who collected a lot of scalps were good (cannibal), they would not be disagreeing, because in English (at any rate missionary English), ‘good’ would mean among other things ‘doing no murder,’ whereas in the cannibals’ language ‘good’ would mean something quite different, among other things ‘productive of maximum scalps.’ It is because in its primary evaluative meaning ‘good’ means neither of these things, but is in both languages the most general adjective of commendation, that the missionary can use it to teach the cannibals Christian morals. (Hare 1952: 148-149)

Disagreement here rests on the evaluative meaning of “good;” if it were a descriptive term, speakers would be talking past one another.

These arguments are best understood as extensions of Hare’s rather Stevensonian point about the need to preserve the practical function of normative discourse. The problem is that a certain view of moral language leaves it unable to play its characteristic role. The preservation of disagreement across disputes is crucial because one of the chief functions of moral language is to provide a way of discussing competing answers to questions about what to do. Having such language is valuable precisely because it allows interlocutors with substantive differences to hammer out consensus. If moral terms were not shared among people with different moral standards (or, in Hare’s language, people who attached different descriptive meanings to their moral terms), then its role would be
undermined; it would be unable to perform one of its central functions. Any account that leaves us with this result should be rejected, and so wholly descriptive accounts of normative language are bound to failure.

What’s needed is an account of the meaning of moral terms that ensures they are capable of doing their characteristic work. For Hare, the purpose of moral language is to guide choice and action (see Hare 1952 ch. 8, especially sections 1-2). In order to meet the requirement, then, we’ll need an account of meaning that ties moral language to this purpose. Hare’s answer, implicit in his criticisms of descriptivism, comes in the form of an account of evaluative or prescriptive meaning. The primary meaning of “good” (and, by extension, of related moral terminology, as we’ll see) is bound to its role in guiding choices; to call something good is to commend it as choiceworthy relative to a certain comparison class. To commend something is to offer it as an answer to the question, “which shall I choose?”

Consider, again, the idea that there are good instances of, among other things, automobiles, chronometers, and strawberries. Imagine that we’re teaching our language to a foreign philosopher. We might teach him the meaning of “automobile” or “chronometer” by showing him various examples from which he could extrapolate the various features of a thing that make it an automobile or chronometer. But we couldn’t teach him the term “good” in quite the same way, because, in this case, we’ve got one predicate the application of which involves different descriptive criteria in different cases. In applying it to cars, we look for one set of features; in watches, another set, and so on. What makes something a good strawberry makes it a terrible chronometer. Yet we’ve got one predicate, not many. To teach the evaluative meaning of “good,” Hare suggests,
centrally involves the notion of choice: what our English student would choose (in the
category of strawberries or the fictional shmakums) reveals what he thinks best. And in
teaching him this, we don’t need to know what a shmakum is; we need only to know
what shmakum he would choose in order to tell him which one he thinks is best. The
meaning of the term is, in this case, independent of the criteria by which it is applied.

Judgments about what’s good, and, by extension, judgments about what we ought
to do, are closely connected to action, recommendation, and choice. In this way
evaluative judgments are more like imperatives, which also answer questions about what
to do, than they are like indicative sentences, which are relevant to these questions only in
conjunction with some principle moving from descriptive claims to conclusions about
action (e.g., “if p, then I’ll A”). But evaluative claims are more like universal
imperatives, because they invoke general standards and consistency pressures. One
cannot say, for example, that a painting is good, yet another, which is like it in all
respects, is not good. Another way of putting this is to say that a thing is good because of
some other features; its goodness can’t vary independently, as its color can. The reason
for this, Hare thinks, is that universal imperatives and evaluative claims are guides to
action, and this purpose would be defeated if our evaluative language didn’t have to
respect this constraint. If we could call one thing good and its twin bad, we’d be setting
no standard at all for choice and action. Calling a picture good is committing oneself to
some standard for evaluation, and this involves commitments to make similar evaluations
in other cases.

58 In choosing a particular shmakum, of course, he’s committing himself to a particular descriptive standard
for evaluating them, since there must be some descriptive features in virtue of which a shmakum is good.
But this descriptive meaning is not the primary meaning of “good shmakum” because one can know the
meaning of that phrase (in an important sense) while having no idea what a shamkum is.
Let’s consider the analogy to imperatives more carefully. Hare thinks that moral
claims (including ‘ought’ claims), in their primary evaluative use, entail imperatives (see
Hare 1952: 11.2). Saying, for example, “One ought not to A” is like saying “don’t A!”
both to oneself and others. Sincerely assenting to an imperative, in turn, requires the
commitment to acting accordingly in the right circumstances:

It is a tautology to say that we cannot sincerely assent to a second-person
command addressed to ourselves, and at the same time not perform it, if now is
the occasion for performing it and it is in our (physical and psychological) power
to do so. (Hare 1952: 2.2)

Therefore, assenting to a moral claim commits us to acting accordingly, if we can.
Furthermore, moral claims are a special class of imperatives, distinguished by two
additional features: they are also universalizable and overriding. To say that moral
claims are universalizable is to say that, in making them, one commits oneself to making
the same assessments of all relevantly similar cases. That is, one cannot say, on pain of
contradiction, that an act is right though another, which is just like it in all other respects,
is not right (see Hare 1952: 11.5; Hare 1963: 1.3, 3.2, 8.2, and elsewhere).\footnote{The rationale for this differs from The Language of Morals to Freedom and Reason, though the difference is largely one of emphasis. In LM the justification has to do with the purposes of the practice—to allow one’s moral evaluations to vary independently would be to impart no standard for acting, and so would defeat the purpose of moral language. In Freedom and Reason the issue has to do with the descriptive meaning of any particular use. Notice, though, that universalizability will be a feature of any universal prescription or evaluative claim. If you think that this couch is ugly, you’re committed to thinking that all couches that are identical in relevant respects (color, shape, etc.) are also ugly, and you’re bound to come up with some difference between this couch and some other sofa that you judge to be nice. And so for other evaluations, for the same reasons Hare gives: it defeats the purpose of the practice to allow evaluations to vary without descriptive difference.} An
imperative, or a prescriptive claim, is overriding when it trumps other competing
prescriptions. Consider this claim about moral judgments:

There is a sense of the word ‘moral’ (perhaps the most important sense) in which
it is characteristic of moral principles that they cannot be overridden in this way
[as an aesthetic principle is overridden in order to keep one’s partner happy], but only altered or qualified to admit of some exception. This characteristic of theirs is connected with the fact that moral principles are, in a way that needs elucidation, superior to or more authoritative than any other kind of principle. A man’s moral principles, in this sense, are those which, in the end, he attempts to guide his life by, even if this involves breaches of subordinate principles such as those of aesthetics or etiquette. (Hare 1963: 9.4)

Commitments to aesthetic principles and the demands of etiquette are, like moral claims, action-guiding and universalizable; they are distinct from moral claims because only commitments to the latter are overriding. 60

Moral claims, then, are identified by their formal features, not by some accompanying moral sentiment or feeling, and not by any fixed content that distinguishes them from other evaluations. This allows Hare to preserve disagreement in cases where parties appear to be using quite different substantive standards for evaluating courses of action, because the disagreement is located in contradictory commendations. Suppose you say that it’s wrong to A, and I think A-ing is obligatory. Furthermore, we have very different reasons for holding these views: we appeal to different features of A in order to justify our positions. According to Hare, our claims differ in descriptive meaning, because we think that different (descriptive) features of the act are those in virtue of which it’s properly thought to be right or obligatory. If moral language had only descriptive meaning, we’d be forced to admit that many apparently important disagreements are illusory. But we still clash in some real sense, because your claim commits you to the imperative “don’t A!,” while I’m committed to “A!” Since these are imperatives, assent requires a commitment to act appropriately in the specified circumstances. Because they demand incompatible actions, I can’t assent to your claim.

60 See Frankena (1988) for a discussion of Hare’s attempt to identify moral claims in these terms.
while still holding on to mine. Our claims conflict in at least this sense. There’s also a more robust conflict. My command applies to myself and to all others similarly situated—including you. My moral claim tells you what to do, in effect, and your incompatible judgment offers an incompatible answer to the same question. Hence we have something very much like real disagreement.

Hare’s view seems to get at something true and important. Why is it, after all, that we take people with radically different evaluative standards to be talking about the same thing? They’re really at odds with one another, we might think, because they have competing answers to the question of what to do. It doesn’t really matter what standards they use in answering this question—that is, we’re not tempted to interpret them as talking past one another because of the diversity of their moral views—as long as these standards give directions to action. Issuing incompatible (universal) imperatives is enough for some form of genuine incompatibility, regardless of what descriptive standards were used to arrive at the clashing commands. Hare thus ensures that we’ll have real conflict, and not merely verbal disagreement, in any case where speakers make the right sorts of contradictory prescriptions.

This advantage comes at a cost, because it’s available only once Hare identifies moral judgments as judgments of what to do, all things considered. The disagreement is about action; it’s the prescriptive element of moral meaning that preserves real dispute in cases of diverging descriptive meaning. In order to stand in the right sorts of conflict, moral judgments must settle the question of what to do. The moral ought and the last ought before action are no longer distinct. This leads to some familiar objections against Hare’s view, all of which take issue, implicitly or explicitly, with the idea that making a
moral judgment commits one (conceptually) to acting accordingly if doing so is possible. Darwall, Gibbard, and Railton, for example, think that “one consequence [of Hare’s view] is that a person can never want wrong things done—things he judges wrong, that is” (Darwall, Gibbard, and Railton: 18). And Hare’s view has been criticized for its supposed inability to give a satisfying account of backsliding or weakness of the will, which occurs just when agents make sincere moral judgments and then fail to act accordingly. Since it makes sense to wonder whether or not a moral decision settles the broader question of what to do—in fact, there are times when the answer seems to be that it doesn’t—Hare’s position seems implausible, since it threatens to conflate two distinct judgments. The disagreement that he preserves may not be of the right kind.

Let’s consider an example that’s meant to illustrate the problem for prescriptivism. Consider a committed aesthete whose life is governed by his judgments of beauty and who makes decisions about how to act by appealing to considerations we’d label as aesthetic. “Is it elegant and suave to act this way?” he might wonder, “or would one look better doing something else?” It’s these considerations that carry the day when thinking about what to do. This aesthete, we might imagine, sometimes makes (what we’re tempted to call) moral judgments about what’s right and wrong by appealing to considerations about rights, desert, human flourishing, and the like. These are more than

---

61 It’s misleading to put this in terms of wants because this suggests an irrelevant psychological issue. If we understand “want” here as synonymous with “desire,” then Hare has no problem admitting that a person might want to do the wrong thing or want wrong things done. As long as this desire is overridden, there’s no problem. All that this person cannot do, on Hare’s view, is prescribe doing what he’s judged to be wrong.

Think of the Kantian distinction between inclinations and what the agent wills. What matters isn’t what the agent yearns for or desires—it’s what she rationally wills. So too for Hare, in an important sense. The agent can have visceral urges toward doing the wrong thing; this presents no problem. What is impossible, according to Hare, is for the agent to sincerely judge something wrong and then prescribe it or decide that it’s to be done in spite of this.
just ‘inverted commas’ judgments, because he thinks critically about his moral judgments, and, if aesthetic considerations are neutral on a given occasion, he’ll take his moral judgments to be decisive. But judgments of beauty are always overriding, and, what’s more, he thinks others should act and think in just the same way. What should we say about such a character?

One option is to interpret the case as I’ve presented it. According to this reading, the aesthete makes genuine moral judgments when he thinks about rights and duties, and he feels some motivation to act accordingly. Though these judgments are often overridden by aesthetic considerations, they are, nonetheless, moral. A second, less natural interpretation sees him as making overriding moral judgments that just happen to be about beauty and elegance, where our moral assessments look to questions about, say, fairness and desert. We might think that he’s elevated his aesthetic commitments to moral principles, but this is most plausibly understood as claiming that aesthetic considerations play a role in his decision-making that’s analogous to the role that moral thinking plays in our own. In short, the most natural interpretation of this character is the one that Hare’s view denies us: for this particular agent, aesthetic, not moral, considerations are overriding when it comes to deciding what to do. Because Hare identifies moral judgments and claims by their formal role in practical reasoning (specifically, by appeal to universalizability, prescriptivity, and overridingness), he can’t say what we’d like to say about our committed aesthete.

The underlying problem is that our intuitions seem to give some support to the idea that moral claims are identified (at least in part) by their content. The example seizing on this and exploits it to make trouble for Hare. To the extent that it has force, it’s
because we have a rough-and-ready idea about what moral judgments are about, what kinds of reasons are relevant to supporting them, and so on, and these features sometimes come into conflict with the formal role that Hare demands of moral commitment. The aesthete simply dramatizes the point made earlier: there seems to be some room between moral judgment and a decision about what to do, all things considered. One might make sincere moral claims while still thinking that the broader question of practical reason remained unanswered.

Think again of Hare’s discussion of the relationship between judgments of aesthetics or etiquette and moral judgments. Judgments of aesthetics or etiquette share the features of prescriptivity and universalizability, though, of course, they’re not overriding. We might wonder, then, how they’re to be identified and distinguished from one another, given that formal features alone don’t seem to be up to the task. The natural way to do this is by appeal to their content: judgments of etiquette are what they are because they’re about particular kinds of actions and distinct sorts of reasons in favor of those actions. If so, then why not make a similar appeal in the case of moral judgment as well? We’re tempted to invoke various non-formal features to distinguish between different species of evaluations, and this suggests that moral discourse might also be identified by this same kind of appeal. This is particularly true if we think, contra Hare, that moral judgments, like those of etiquette and aesthetics, are relevant but not decisive when it comes to deciding what to do. Once we assent to this, overridingness can’t be used to identify moral judgments.

I don’t mean to be offering a completely developed and decisive argument against Hare’s view. For now, our needs are more modest: we need to see some of the costs that
come with the advantages of Hare’s prescriptivism. This view, more than any of the earlier noncognitivist proposals we’ve been considering, preserves the genuine incompatibility of clashing moral judgments. It does so by appealing to contradictory imperatives and incompatible courses of action. This places a strain on our ordinary understanding of moral discourse, insofar as we no longer have any conceptual room between moral judgments and judgments of what to do, all-in. Thus, prescriptivism gets something wrong in its explication of our moral practice. In its own way, it fails the test that Stevenson and Hare himself set out, because it renders incoherent a question that we take to be substantive and important. The aspiring noncognitivist needs a more nuanced way of marking the boundaries of moralizing. This leaves us with a puzzle: how are we to preserve the advantage of Hare’s account while offering a narrower, more accurately drawn view of the boundaries of moral judgment.

2.8 Taking stock

Reflecting on the early history of noncognitivism allows us to see some recurring themes in the attempt to develop a philosophically satisfying view of moral discourse. Here, I’ll identify some of these themes and discuss their importance in helping us to see the challenges facing the noncognitivist. Two issues have played a central role in our discussion. The first concerns \textit{disagreement}: when two speakers take opposing views on some moral issue, how exactly does their dispute work, on the noncognitivist view, and can it be legitimized? The question is whether and how we’re at odds with one another when we’re involved in moral disagreement. As we’ve seen, the first generation of noncognitivists were tempted to appeal to the notion of clashing attitudes or commitments to incompatible imperatives, but these replies raise as many questions as
they answer. If we’re to make good on these suggestions, we’ll have to say more about how these attitudes and imperatives work, how they come into conflict, and how they’re tied to moral judgments.

The second main issue has to do with the identification of moral discourse as something distinct from evaluation generally. How are moral judgments distinguished from those of aesthetics, etiquette, prudence, and so on? Barnes, Stevenson, and perhaps Ayer answer this question by appeal to characteristically moral sentiments, attitudes, or feelings, and so they claim, more or less, that a moral judgment is what it is in virtue of its link to some mental state of this sort. In addition, Barnes hints that moral judgments might be identified by the kinds of reasons used in their justifications, because he seems to think that formal considerations alone are insufficiently fine-grained to mark out moral judgment. Hare, on the other hand, eschews this strategy in favor of an appeal to the formal features of universalizability, prescriptivity, and overridingness, and identifies moral commitments or principles as any that meet these criteria. This, as we’ve seen, has its disadvantages, but it rescues Hare from a serious problem that, we might think, has been insufficiently appreciated by other early noncognitivists, namely, the need to identify just what mental state is involved, whether by being evinced, expressed, evoked, or in some other way, in moral utterance and judgment.

These two issues are related, of course. Hare’s identification of moral discourse by appeal to a certain sort of prescription tells us how to distinguish moral judgments

---

62 Recall that he’s worried about attitudes that involve the right kinds of general features (they involve commitments to act in certain ways and appropriate responses to others’ failure to adhere to the same norms, and so on) but which nonetheless fail to qualify as moral. Thus we might attribute to him a worry like the one we raised against Hare—the point that formal features alone will give an inaccurate delineation of different sorts of evaluation.
from other kinds of evaluations while also explaining just how incompatible judgments collide: they involve conflicting imperatives. As we saw, Hare’s views led him to implausible conclusions concerning the practical role of moral judgments. If moral judgments are overriding prescriptions, they settle the question of what to do, and so Hare’s view doesn’t allow us to ask if moral considerations are sometimes trumped by other kinds of reasons. We might wonder if appeals to particular sentiments, emotions, or attitudes would do a better job providing us with an accurate delineation of the boundaries of moral judgment. This is promising because these mental states might serve as a means of fixing the subject matter of conversation and thus of holding off the threat of equivocation without appealing to any particular descriptive standards ‘built in’ to moral judgment. And disagreements about moral questions might be possible without Hare’s broader commitment to disagreement about what to do. To pursue this strategy, we’ll need to take on the challenge of showing how the resources that these states provide can be used in an explanation of the incompatibility of (seemingly) inconsistent moral views.

There’s another important connection between the two problems we’ve been discussing. Suppose we thought that Hare’s view didn’t capture what’s distinctive about our moral judgments, and we were inspired to do better. To improve on prescriptivism, we’d have to make sense of the idea that there are several varieties of evaluative judgment. Moral attitudes just don’t function in the way that generic pro- or con-stances do, and we can distinguish kinds of evaluations: to call something ugly is not to call it
disgusting, wrong, rude, shameful or shocking.\textsuperscript{63} The expressivist’s account of moral discourse must be able to distinguish its judgments from those of aesthetics, etiquette, prudence, and other non-moral, though still \textit{evaluative}, perspectives. Its identification of moral judgment must also account for the fact that different kinds of pro- and con-attitudes don’t always conflict—that is, it must provide disagreement at just the right places. You claim that what Steven did was in poor form; I claim he acted within his rights. So far, we don’t have a dispute, because we’re offering two different kinds of assessment, and these can be orthogonal to one another.

Prescriptivism ran into difficulty here just because it’s intuitively plausible to distinguish all of these judgments from a view about the last ought before action, and because it’s hard to see how a purely formal specification will capture the distinctions between them. Hare recognized, at least to some extent, that aesthetics and etiquette are distinct from each other despite their formal similarities, but he didn’t provide a satisfying way of keeping them separated.\textsuperscript{64} If expressivism is to do better, it must give us a better way of carving out a recognizably \textit{moral} discourse. The mental states it specifies must be linked not to evaluation generally, but to moral judgments in particular.

It’s important to be clear about the requirements such a strategy takes on, because noncognitivism has been lax about meeting these crucial demands for too long. It’s far

\textsuperscript{63} See D’Arms and Jacobson (2000).

\textsuperscript{64} When discussing overridingness, Hare gives an example along these lines. Suppose your spouse gives you some cushions for the sofa as a present. The problem is that they clash with the sofa’s covering, and so displaying them would violate one of your aesthetic principles. This principle is overridden by one’s duty to be respectful of one’s spouse, say—it remains true that the clashing colors look bad together, but it’s more important to do the right thing. (Notice that this is different from having an aesthetic principle that said these colors look bad together unless one’s spouse doesn’t think so, in which case they look good.) So both aesthetics and etiquette are \textit{prescriptive} and \textit{universalizable}, but they aren’t overriding. Hence they share formal features, yet Hare discusses them as distinct kinds of evaluation. He doesn’t seem to notice the problem this raises for his position.
from adequate to say, as Ayer does, that moral claims evince “a certain moral sentiment” and leave the matter at that. This insouciance is unsustainable in light of the fact that one of noncognitivism’s purported advantages is its metaphysical economy. It’s supposed to allow us to make good naturalistic sense of our evaluative discourse by explaining how our talk of value can go on without invoking any problematic or mysterious entities.  

But this makes for meager ontological savings, if the mental states at the theory’s core are left unspecified. We cannot simply assume that there are the right kinds of noncognitive mental states with sufficiently close ties to our practice of moral judgment. We need reassurance that the noncognitivist’s candidates are an independently secure part of our best view of the mind. In the absence of an obvious and distinct moral phenomenology, this is a potentially serious challenge.

The problem becomes more troubling when considering an additional constraint on these mental states. Noncognitivism aspires to a reductive analysis of moral judgment; it promises to give us an account in purely non-moral terms. The account strives to avoid circularity, so the states to which it appeals must be independent of, and prior to, the judgments they’re used to explain. The challenge is simple: the noncognitivist’s mental states of choice must mirror our moral judgments in a fairly accurate way—that is, there can’t be too many cases in which the two seem to diverge—but it must not contain or require that judgment, on pain of circularity. So, to use a simple example, if the attitude of approval is partly constituted by the judgment that something is good, then we can’t explain what it is to judge something good by saying

---

65 The force of this advantage will depend in part on how mysterious are the metaphysical commitments of rival approaches.
66 This requirement is discussed in D’Arms and Jacobson (1994).
that this is merely the evincing of an approving attitude. (This is, of course, simply a restatement of Hare’s early criticism of expressivist views.) It’s hard to see what state might be plausibly thought to track moral judgment with sufficient accuracy while remaining conceptually independent.67

2.9 Gibbard

We’ll see these problems more clearly by considering Allan Gibbard’s ‘norm-expressivism.’ This view, which shows the influence of both Hare and Stevenson, attempts to pay noncognitivism’s outstanding debts by providing answers to some of the problems we’ve been considering. In particular, it tries to understand moral judgment by appealing to particular sentiments. By looking at how Gibbard’s account struggles to make sense of univocal disagreement, we’ll be able to understand how the requirements we’ve been developing and discussing affect the aspiring noncognitivist.

Gibbard’s view about moral judgment is embedded in a broader theory about judgments of rationality. By his lights, there’s no property we’re ascribing to something (such as an action, a decision, a feeling, or whatever) when we call it rational. Instead, we’re endorsing that thing. Any purely descriptive account of rationality will miss this essential part of what it is to judge something rational—just as, according to Hare, descriptive accounts of “good” will miss the crucial commending function of the term (see Gibbard 1990 ch. 1). For any descriptive specification of what it is to be rational, we

67 See D’Arms and Jacobson (1994) for a discussion of this issue in connection with Gibbard’s view.

John McDowell and David Wiggins have independently stressed the importance of this point in their arguments against a Blackburn-style projectivism. If projectivism were right, McDowell says, “we ought to be able to focus our thoughts on the response without needing to exploit the concept of the apparent feature that is supposed to result from projecting that response” (McDowell 1997: 218). But we can’t do this, claims Wiggins, because “there will often be no saying exactly what reaction a thing with the associated property will provoke without direct or indirect allusion to the property itself” (Wiggins 1997: 232). Hence the projectivist has failed to secure the needed explanatory priority.
can sensibly wonder whether or not we should endorse something that meets the specified criteria, and so descriptive accounts of rationality fail to capture something essential to the notion. Rationality judgments, on Gibbard’s view, are just ‘bland, flavorless’ endorsements meant to settle the question of what to do, or to feel, or to think.  

(Granted, this is a somewhat stipulative claim, since in ordinary language we sometimes do wonder whether we should do what’s rational. “Stop being so reasonable—loosen up and do something crazy,” we tell the Vulcans in our midst. Gibbard is aware of this, of course, and admits that his use is a bit of a stipulation; he also proposes the locution “makes sense to” as another way of getting at the notion of endorsement he’s proposing. But, he thinks, this notion of all-in endorsement captures something crucial to at least one important use of the term “rational.”)

Let’s leave these worries about linguistic propriety aside. The important point is that Gibbard is an expressivist about these judgments, whether or not they’re properly called judgments of rationality. But what are we expressing in making them? In calling something rational, he says, we’re expressing acceptance of norms that permit the thing in question. A norm is “a possible rule or prescription, expressible by an imperative” (Gibbard 1990: 46). To accept a norm is to be in a fairly complex psychological state that involves dispositions to act in accord with the norm, to defend it in conversation, and

---

68 Judgments of rationality aren’t meant to settle the question of what to do if one happens to go in for doing what’s rational. They’re meant to settle the question of what to do, all things considered. Gibbard’s view of the rational ‘ought’ identifies it with ‘the last ought before action.’ Thus, Gibbard’s rationality judgments and Hare’s moral judgments are quite similar.

69 Pace David Byrne, who enjoins us to stop making sense, thus endorsing what it doesn’t make sense to do. One wonders how interesting is the issue of whether or not we have a linguistic expression that consistently signals all-in endorsement.
so on. This view thus has ties to both expressivism (insofar as what appears to be a truth-evaluable judgment is the expression of some noncognitive mental state, in this case norm-acceptance) and to Hare’s prescriptivism (because of the all-in imperative that’s part of what’s expressed).

On Gibbard’s view, judgments of morality are not assessments of the rationality of particular actions. Instead, they are judgments about the rationality of the moral sentiments, guilt and (impartial) anger. To say that an act is wrong is to say that it’s rational (or that it makes sense) to feel guilt because one performed it, or to feel anger at its performance by someone else.⁷⁰ (We might think that this gets something right about the ties between reason and the moral sentiments, while resisting expressivism about rationality judgments.⁷¹) When conjoined with the Gibbardian view about rationality, we have this proposal: to call an act wrong is to express one’s acceptance of norms that permit guilt (for the agent) and impartial anger (for the observer) at its performance.⁷² A person’s moral views are revealed by the episodes of guilt and anger that he endorses.

This view has several immediate attractions. First, it makes moral judgments appropriately independent of our actual emotions, sentiments, or feelings, because what matters is not whether we feel guilt or anger but whether we endorse our responses. The

---

⁷⁰ There are some complex and important issues here concerning the precise nature of the endorsement. Gibbard says that to judge an act wrong is to judge that guilt and anger are warranted, but he never says very much about just what notion of warrant this is. One problem is this: there are sometimes strategic, prudential, perhaps moral reasons to endorse sentiments that don’t ‘fit’ the situation—as when it’s dangerous to get angry at a superior’s outrageous action. We might all-in endorse refraining from anger here, but it’s clear that anger is warranted (in some sense) because, after all, what the boss did was really awful. What Gibbard needs is a special flavor of endorsement, but he can’t have that, because then judgments of warrant aren’t (bland and flavorless) judgments of rationality. This problem is discussed in detail in D’Arms and Jacobson 2000a.

⁷¹ We could also reject expressivism about morality and think that Gibbard gets things right about the last ought before action. Whether this is the same as the rational ‘ought’ is another question.

⁷² Gibbard’s use of the language of permission, rather than requirement, is puzzling: we might think that it’s better to say that guilt or anger is required by moral norms.
idea that moral judgment is nothing more than the expression of sentiment seems wrongheaded, because it’s fairly clear that moral judgments are modally distinct from any occurrent emotion. We can make moral judgments without experiencing emotion, as in case of fatigue, disaffection, or depression. Conversely, we can experience an emotion without the associated judgment, as when I feel guilty while thinking I did nothing wrong. This is commonplace—but it’s at least less clear that I could endorse guilt (in the appropriate sense) while thinking I did nothing wrong.73

Second, guilt and anger, the mental states at the theory’s core, are as familiar as Ayer’s ‘certain moral sentiment’ is mysterious. These emotions seem to be a familiar component of our everyday psychology, rather than ad hoc sentiments conjured up to meet noncognitivism’s needs.

Third, Gibbard’s construal of moral judgments as endorsements of particular sentiments allows conceptual room for us to think that moral considerations don’t settle the question of what to do. Unlike Hare’s prescriptivism, norm-expressivism makes sense of the thought that we might have reason to act contrary to our moral judgments. By the norm-expressivist’s lights, the question of whether one has reason to act immorally is equivalent to the question of whether it ever makes sense to perform an action about which it makes sense to feel guilty. It’s at least conceptually possible to endorse an action and endorse guilt at its performance. Does the fact that an act warrants guilt mean that it’s not rational to do it? This is an open question, just as Gibbard wants

73 It’s interesting that guilt and anger are sentiments that commonly disobey our own norms for them. Consider survivor guilt, as when I feel guilty about being the only one to make it out alive, even though I didn’t do anything wicked to ensure my survival. There’s also what we might call survivor anger, enshrined in our best pop-psychological theories as one of the stages of grieving, during which one is angry at, say, a deceased relative because she died—even though she did nothing to bring this about. It makes sense that we feel these sentiments, but it doesn’t make sense to feel them.
it to be. Since people do seem to think, on occasion, that it makes sense to act immorally, this is an advantage.⁷⁴

Fourth, the specification of guilt and anger seems to get something right insofar as these sentiments seem connected, in a deep way, to our moral practice. At first glance, at least, we might think that these are especially promising resources to use in giving a more accurate account of moral judgment. We often feel guilt in response to wrongdoing, and we talk each other out of guilt with reassurances that no wrong has been done.⁷⁵ These mental states certainly seem to be better candidates for a close fit with moral judgments than are Hare’s overriding prescriptions.

(One worry, of course, is that they map too well onto the contours of our moral practice. Gibbard is committed to a non-circular analysis, and so he has to argue that guilt doesn’t include the judgment of wrongdoing. If guilt is partly constituted by the judgment Gibbard seeks to explain, then his account fails to meet his own adequacy conditions.⁷⁶ He’s thus committed to claiming that the best views of guilt don’t involve the troubling judgment—though we might find his arguments a bit quick here. For example, he relies heavily on a point about the modal distinctness of guilt and judgments of wrongdoing: we might feel the sentiment without the allegedly constitutive judgment. As we’ll see later, this argument ignores the possibility that feeling guilty involves

⁷⁴ Devout moralists, on the other hand, will have a further commitment: they accept a norm that says that it doesn’t make sense to do anything that warrants guilt. The question about guilt settles the question of what to do, for people who accept this moralistic norm, though this is a substantive commitment, not a conceptual one.
⁷⁵ It’s also hard to ignore that being guilty is a morally infused notion having to do with responsibility and desert. This, as we’ll see, raises the threat of circularity. In addition, it’s interesting that guilt seems to fit our moral judgments more accurately than anger does, since anger makes sense at various nonmoral infractions. Whether Gibbard’s specification of impartial anger is enough to avoid this worry is an interesting question; it’s discussed in D’Arms and Jacobson (1994).
⁷⁶ See D’Arms and Jacobson (1994) for a discussion of the circularity problem.
construing one’s actions as wrong without judging them to be wrong, and so gives short
shrift to a moderately attractive view of the emotions. For now, though, I want only to
note the problem and leave it to one side, partly because we’ve got different concerns
here, and partly because the problem has been discussed insightfully elsewhere.)

The sentiments are an attractive resource for noncognitivism because they seem to
be metaphysically unproblematic, they at least appear to be distinct from judgments, and
they have ties with motivation and action. What’s more, invoking the sentiments seems
to be an appealing strategy for a noncognitivist who, unlike, say, Ayer, wants to preserve
and legitimate important features of our moral discourse. Emotions are subject to
criticism: it’s not at all odd to think that there’s a right way to feel, or that some
emotional reactions are warranted and others are not. That is, they are the subject of
normative reflection, discussion, and evaluation. These commonplace features of our
emotional lives are exciting because of what they offer to the aspiring noncognitivist.
Early expressivist views were criticized for failing to make sense of moral reasoning, and
we’ve seen that there’s much to invite this charge in the work of, say, Ayer and
Stevenson. But if, like Gibbard, we appealed to actual emotions, and gave careful
attention to the ways these worked, we might hold out hope of doing better. Since
emotions are not judgments, and yet invite critical scrutiny, we might think that they
could provide a way of preserving some of the nuances of moral discourse without
committing us to the idea that moral claims express truth-evaluable propositions.

77 There’s an additional puzzle here about how the moral sentiments are related to our choices and actions. Suppose I endorse guilt for A-ing. There’s a crude view about how this might give us reason to refrain from performing A: guilt feels bad, and this pain tips the scales against acting in a way that will produce guilty feelings. This seems to get the deliberations wrong—it’s just not the way we think about our emotions in our practical reasoning. (This issue is connected to the circularity worry discussed below).
With these motivations on the table, let’s turn to the issue of disagreement. How will the norm expressivist preserve the right kind of conflict between speakers with conflicting moral views? Though Gibbard doesn’t say much about this explicitly, we can construct an answer on his behalf. For Hare, competing moral claims are at odds with one another because they direct us to act in incompatible ways. Gibbard, on the other hand, sees moral commitments as norms directly governing guilt and anger, not action, but he can rely on roughly the same answer to the question of disagreement. By saying that \( A \)-ing is wrong, you’re expressing your acceptance of norms (i.e. rules or imperatives) that permit guilt or anger at its performance; when I say it’s all right to A, I’m expressing my own acceptance of norms that don’t permit this. The imperatives “feel guilt at A-ing!” and “don’t feel guilty when you A!” are incompatible in just the same way as the imperatives “A!” and “don’t A!” So we’re giving contradictory directions, though about sentiments rather than actions.\(^7\)

We might think that Gibbard has another, more subtle view of disagreement available, though, on inspection, this thought cannot be sustained. This alternative involves the notion that emotional episodes are ‘evaluative presentations’ of things: to feel a certain way about what you did, for example, is to see it in a certain light. This is promising insofar as the incompatibility between different ways of seeing a particular action (whatever this comes to) might serve as the fulcrum of disagreement. To see why this avenue isn’t available to the expressivist, we’ll have to take a short detour through judgmentalist and quasijudgmentalist views of emotions. Though our task isn’t to assess

\(^7\)This is similar to the view suggested in Ayer (1984).
these views directly, we can get enough on the table to see their relevance to our current concerns.

Start with the observation that there’s an interesting connection between certain emotions and corresponding thoughts. For example, being afraid of something is closely tied to thinking that it’s a threat. We might say that fear presents its object as dangerous, or that feeling fear involves seeing the object as threatening. One way of explicating the connection between the emotion and the proposition is by way of belief: to feel fear is, in part, to believe that the object of one’s fear is dangerous. This is problematic insofar as there seem to be cases of fear without the relevant belief. Phobics acknowledge that their fear is irrational because they don’t believe that its object is dangerous. Yet they feel fear nonetheless. Thus recalcitrant emotion presents a problem for judgmentalism, because particular judgments can’t be constituents of emotions if occurrences of those emotions are possible without the relevant judgments.

The puzzle, then, is how to hold on to the plausible idea that emotions are linked to certain contents while making sense of recalcitrant emotions. One solution involves giving up the idea that feeling an emotion involves believing that content in favor of a weaker propositional attitude. You don’t have to believe that the dog is dangerous in order to be afraid of it, but you do have to construe the dog as dangerous, or see it as dangerous, or make a quasi-judgment that it’s dangerous. This attitude does not require belief, but it’s still an attitude toward (or involving) the same propositional content.79

The construal theorist or quasi-judgmentalist agrees with the judgmentalist that various

---

79 It’s unclear just what this attitude is, but the intuitive idea is that we’re somehow entertaining the thought without endorsing it in the way we would be if our attitude were one of belief.
emotions involve propositional attitudes toward particular contents but disagrees about just which propositional attitude is involved.

This way of thinking about the emotions is appealing for several reasons.\(^8\) It allows for a close connection between, say, an episode of fear and the thought that something is dangerous, while also making sense of the idea that these two can (and commonly do) diverge. It also explains why there’s something troubling about this divergence: someone with this collection of attitudes is in an unstable or self-confuting position of *seeing* something as dangerous while *believing* that it is not.\(^9\) It’s plausible that these attitudes pull in different directions, which would also give us a story about why changes in belief often *are* effective in changing emotions and why the evidence for belief ("look how friendly the feared dog is with others!") influences the emotion as well.

As my liberal reliance on the notion of propositional content has suggested, this way of thinking about the emotions is no friend to the expressivist. In fact, some of the attractions of this view emphasize important restrictions on what expressivism can’t say about the mental states involved in moral discourse. The problem here has to do with the notion of *representation*. According to broadly realistic strains of thought, moral claims tell us about (some aspect of) the world. For the expressivist, on the other hand, these claims are best understood as something other than responses to interactions with some distinct ethical reality. They don’t describe; instead, they project attitudes on the world,

\(^8\) For an interesting criticism of quasi-judgmentalism, see D’Arms and Jacobson (2003).
\(^9\) More will have to be said about the nature of this tension. Our earlier discussion of varieties of endorsement (or condemnation) shows that not all of these are in genuine conflict or agreement with one another. So in order to develop this account of why such a combination of attitudes is unstable, we’d need a fully fleshed-out story about why these attitudes are unhappily compatible. Optimism is warranted, though, because *seeing something as true* and *endorsing as true* are closely related attitudes, and so it’s likely that a more developed account of "seeing as" will yield the desired result.
or make commands, or express sentiments, or express the acceptance of norms. To use Anscombe’s old metaphor, moral claims and straightforward descriptive claims have a different ‘direction of fit.’

What the expressivist will find so discomforting about the construal theory of emotion is that it views emotional episodes as representing things as being a certain way—they give us evaluative presentations of their objects. While judgment or belief isn’t involved directly, the relevant propositional content is part of the emotional experience, and, while the experience of an emotion can’t be true or false, it can be criticized as misrepresenting things. This is a plausible account of what’s gone wrong in cases where emotion and belief come apart: I know that the dog is harmless, yet I fear it. I see my fear as inappropriate or unwarranted just because I believe the dog to be harmless, and so I think that my fear is getting something wrong. Since, on this view, emotions have representational content, they aren’t available to the expressivist. The problem isn’t with the particular doxastic attitude taken toward a particular content, it’s that there’s a problematic content there to begin with. The broader lesson here has to do

---

82 Does the endorsement (in the relevant sense) of an emotional occurrence have a tie to judgment or belief? We endorse (in some quasi-epistemic sense) an evaluative presentation of the way things are, and this seems quite close to thinking that things are as the emotion portrays them.

83 Note too that this precludes one way of accounting for disagreement by appeal to different emotions or different norms for them. If we thought that guilt represented things a certain way (it presents my action as wrong, for example), we might say that our disagreement consists in our contradictory representations. My claim that [X]-ing is wrong endorses guilt at [X]-ing, while your claim that it’s permissible rejects that sentiment. So we disagree about whether guilt gets something right or not. It’s just this kind of incompatible representation that isn’t what’s going on in moral disagreement, according to the expressivist.

Another, related problem has to do with a common-sense way of understanding our moral utterances. One might object that Gibbard’s view makes our moral claims into claims about our sentiments, rather than about the actions that appear to be their objects. One reason this seems to be an unappealing objection is that our notion of guilt is not only object-directed, but presents those objects in a particular way; hence talking about guilt and anger (or other psychological states) isn’t really a change of subject. But, of course, the intuitive appeal of this thought does Gibbard no good if he has to construe the sentiments in ways that preclude this sort of response. In evaluating Gibbard’s proposal, we need to be careful not to rely, even implicitly, on the resources that are beyond the reach of the scrupulous noncognitivist.
with the threat of circularity: Gibbard must avoid any account of guilt that relies on the moral concepts he seeks to explain. Because quasi-judgmentalism sees episodes of guilt as involving the thought that the agent has done wrong, it runs afoul of this restriction.

What’s more, the scope of the circularity worry extends beyond views that see individual occurrences of guilt (for example) as being constituted by problematic propositional attitudes. As an example, consider the view of guilt advanced by the social psychologists John Sabini and Maury Silver. Roughly, they think that there is no distinct emotion of guilt—rather, guilt names a cluster of feelings associates with being guilty, in the way that ‘feeling like a king’ names a cluster of feelings associated with being a king (Sabini and Silver 1997: 83). To say that one feels guilty is to say that one feels as a guilty person usually feels or is thought to feel, namely shameful, regretful, and so on. ‘Feeling guilty,’ then, is “an umbrella that collects under itself a broad range of feelings having to do with transgressions” (Sabini and Silver 1997: 84).

What’s interesting about this view is that, if it’s right, guilt is inaccessible to the expressivist even though experiences of guilt need not involve the problematic propositional content we’ve discussed. The individual components of the guilt-cluster, such as shame, might be suitable for Gibbard’s use when taken singly, but their assembly into guilt will be off-limits, because this proceeds through moral notions. On the Sabini-Silver view, the notion of the guilty agent is conceptually prior to the emotion of guilt; a description such as “the way guilty people feel” is used to fix the cluster of feelings and emotions picked out by “guilt.” Since we need to have a handle on the idea of a

84 Notice that we could take on Sabini and Silver’s view and still say that guilt is an emotion, if we wanted to count distinct “cognitive sharpenings” as emotions. Guilt would then be the emotion of shame/etc.-at-transgressions-of-a-particular-sort.
(morally) guilty agent before we can identify the feelings such a person typically experiences, we can’t identify the cluster without already having the idea of moral wrongdoing. The circularity problem thus has power even if episodes of guilt are free from problematic moral concepts.

These considerations are offered here just to emphasize the point and scope of the circularity worry. Expressivism’s invocation of sentiment comes at a cost, because it’s burdened with the task of accounting for these states without appealing to various propositional attitudes. This point also affects Gibbard’s response to the problem of disagreement, since it looks as though he must rely on clashing imperatives to preserve real conflict.

Once these conceptual resources are off the table, Gibbard’s problems get worse, because concerns about circularity add force to another set of problems. This issue involves what we might call the scope of moral discourse and, by extension, moral disagreement. Is our moralizing universal, or is it limited to particular agents and cultures? And who could engage in the discourse? These questions put pressure on the norm-expressivist to make sense of the boundaries of moralizing. Keeping the dangers of circularity in mind will show this to be a difficult task, because the most plausible answers to these questions suggest a need for the sorts of conceptual resources that are off-limits to the noncognitivist.

On Gibbard’s view, moral norms are identified by their objects, guilt and impartial anger. This matters because it makes morality as parochial as the sentiments at

---

85 It’s interesting, too, that shame, which on Sabini and Silver is the principal component of guilt, is a response to flaws of some sort. Episodes of shame that constitute guilt are directed at moral flaws, such as defects in character—but we need moral norms in place before we can identify traits as moral failings. Without these, the episodes of shame don’t make sense.
its core. If moral commitments are a matter of accepting norms for guilt and anger, people who don’t accept any norms for regulating these sentiments don’t have moral commitments (in the narrow sense). Hence, people without guilt and anger in their emotional repertoires don’t engage in moral discourse.\(^8^6\) So we cannot have moral disagreements with them, though we may still label their actions right or wrong, and we may have broader disagreements about what to do and how to live. Gibbard is aware of this point and discusses it clearly:

> Are there cultures that get along without morality? Here again, much hinges on what is meant by the term. Some brand morality a peculiarly Western phenomenon, with its roots in the Judeo-Christian tradition; others think morality a cultural universal. I do not claim to know what in general is at issue in these claims, but the norm-expressivistic analysis does direct us to a significant question: are norms for guilt and resentment universal?...The question is rather whether the emotions I have been calling “specifically moral” are culturally universal, and whether the existence of norms for these emotions is a cultural universal. (Gibbard 1990: 54)

And later:

> We ourselves experience guilt and shame, anger and disdain as distinct emotions, and morality in the narrow sense is ours. It consists, we can still say, in norms governing emotions that we ourselves experience and name. If others lack concepts of guilt and anger, that simply means they lack morality on this narrow construal. (Gibbard 1990: 141)

To have moral concepts, commitments, and norms requires guilt and anger; to enter into moral discussion with us requires that one’s emotional repertoire contain these sentiments. Gibbard is thus committed to views about the scope of moral disagreement in virtue of his reliance on norms for guilt and anger.

\(^8^6\) Here I’m assuming that having norms governing guilt and anger requires having guilt and anger. I’ll return to this assumption below.
This commitment is a potential source of problems because there’s a current debate, among some psychologists and anthropologists, about guilt’s universality. Some of these investigators argue that guilt is culturally parochial, that is, it’s not found in all societies and cultures. Suppose, just for the sake of argument, that something like this is right: there are guilt cultures and shame cultures, and members of the latter don’t experience guilt. Participants in a non-guilt culture can’t have norms for guilt, let’s assume, and so they can’t express acceptance of norms for guilt. The norm-expressivist story tells us that they’re not making moral judgments, though, we might imagine, they engage in practices that look an awful lot like morality. They have a discourse that’s used to assess and regulate conduct and character, has an important role in practical reasoning, and perhaps has important ties to various sentiments (shame, rather than guilt, say, or perhaps an emotion for which we have no identifying term). Suppose that their putatively moral judgments appear to come into conflict with our own. Here, we’d like to say, we have a case of genuine moral disagreement. Yet that’s just what Gibbard’s view precludes us from saying, because these people aren’t making claims about what acts are guiltworthy, and they’re not commanding us to feel guilt in particular situations.

There’s a way to sharpen this point. Suppose, plausibly enough, that the norm-expressivist understands our shame-culture participants as expressing their acceptance of norms for shame when they make their quasi-moral judgments. The Gibbardian will then say our apparently conflicting judgments don’t meet each other head-on, because evaluators from the shame culture endorse shame at a particular action, while we don’t endorse guilt. These two attitudes aren’t necessarily at odds with one another. Earlier, I pointed out that different varieties of pro- and con-stances are sometimes compatible.
We’re not always disagreeing with one another when you have an attitude of condemnation toward something while I commend it, because these attitudes may represent different species of evaluation. You might condemn a work of art, say, as shocking, while I commend it as beautiful, and we might still fail to disagree, since these attitudes can happily coexist. In fact, we might endorse this conjunction of attitudes, because we’re sensitive to values that are sometimes at odds with one another, and it’s important that neither one blind us to the other. The point is that endorsements and condemnations of any sort are not enough to ensure real disagreement. We need something more fine-grained—guilt as opposed to shame, for example—but when we have it, we risk failing to capture cases that look like real dispute.

There are two problems here. The first is the most straightforward: our (mostly) hypothetical case contains the makings of a problem for Gibbard’s view, because it offers an example of dispute that his view doesn’t seem to capture. In order to think that this is a serious challenge to Gibbard’s view, we’d have to think that these cases of non-conflicting emotions (or, more accurately, endorsements of emotions) are, contrary to norm-expressivism, real cases of moral disagreement. And this is a live question, since interlocutors might have genuine disagreement (say, about what to do) without disagreeing about the demands of morality. If we think that these shame-culture participants have a form of evaluation that’s like morality in many ways without being morality, we won’t consider our disagreements with them to be moral disagreements, narrowly construed.

The second point is more serious. The challenge for Gibbard is to show that his view can make sense of the boundaries of moral discourse and disagreement using only
the resources at his disposal. Gibbard is not simply committed to the possible
corporalities of both moral discourse and guilt-norms. He is committed to the additional
claim that guilt and moralizing share the same boundaries. It’s not enough for both to be
corporal, in other words, because they must also be corporal in just the same ways.
This is an obvious point, but it’s significant in light of our discussion of the requirements
on Gibbard’s account of guilt. He must give us reason to think that norms for guilt give
us an accurate mapping of the limits of moralizing, and he must do this without invoking
the sorts of propositional content that, say, judgmentalist views build into emotions. The
problem, as we’re about to see, is that it’s quite tempting to violate this stricture in
addressing puzzles having to do with the boundaries of moralizing.

Suppose, for the sake of illustration, that we think morality is corporal. Gibbard
couches this possibility in cultural terms: is moralizing peculiar to the West, or is it
shared by all people? If norm-expressivism is true, it has to turn out that the answer to
this question matches our answer to the parallel question about guilt. Furthermore, the
explanation of why this is so cannot rely on problematic moral concepts. This is a
serious issue, because there’s a looming explanatory gap: why would guilt be culturally
limited in this sense, and limited in just the same ways, unless it’s somehow dependent
on concepts that Gibbard can’t invoke? Or, to put this another way, the corporality of
guilt suggests that it’s a conceptually rich emotion, and this raises worries about just what
concepts are involved and whether Gibbard can help himself to them or not. As we’ve
seen from our discussion of Sabini and Silver, there are problems here even if Gibbard

87 Cultures haven’t evolved different emotions, so guilt’s corporality rules out the idea that it’s some sort
of hardwired response.
avoids the threat posed by quasi-judgmentalism. So far, this is more of a challenge than an objection: if moralizing is parochial, Gibbard must say how it is that guilt matches its boundaries while remaining conceptually independent.88

The next, related challenge to Gibbard’s view involves a point about the modal distinctness of moral concepts and guilt-norms. First, I’ll argue, on intuitive grounds, for a seemingly innocuous point about moral concepts and their acquisition. Then I’ll use this point to put more pressure on Gibbard’s account of guilt.

Our earlier argument about the extent and location of moral disagreement between, say, members of a guilt culture and a shame culture was, in one important way, misleading, because it directed our attention to actual moral dispute, and, in turn, the location of actual moral concepts. An issue that’s just as important, it seems to me, is this question’s modal variant: where could we find moral concepts? It might turn out that moral norms are parochial; certain cultures have them and other cultures don’t. Suppose this is right. Is it possible that a person from an amoral culture (that is, one without moral concepts) could adopt moral discourse as her own?

Take as an example either a person from this distant culture or that favorite character from philosopher’s fiction, the intelligent alien. Suppose we think, for reasons already canvassed, that this person doesn’t participate in moral discourse, doesn’t have moral norms, and so on. It seems enormously plausible to say that this person might acquire moral concepts through the right kind of exposure to them. That is, she could

---

88 It might look as though the answer is obvious: morality and guilt go together because guilt explains (and is the basis for) morality. There are two problems with this. First, we’ll need an account of how guilt works: if it’s culturally variable, then we’re tempted to think that it’s constituted, in part, by various concepts, and, given its close ties to moralizing, this provides some evidence that the concepts it relies on are moral concepts. Second, we’ll need a story explaining why it looks as though we’d use different sorts of evidence to assess the extension of guilt and of moralizing.
come to be a participant in moral discourse though she initially isn’t. There may be arguments that convince this person to adopt moral concepts, to go over to our way of evaluating things rather than her own way, or the change might happen as the result of a kind of conversion experience. Details aside, it’s plausible to imagine that this person would be able to pick up moral concepts and as a result take part in moral evaluation, discussion, and argument. Call this point the *acquirability of moral concepts*.

The question is how norm-expressivism can make sense of this point. It’s clear how Gibbard has to reply: if guilt-norms are *not* acquirable in the way moral concepts are, then his analysis is faulty. A hypothetical evaluator might take on moral concepts, but couldn’t take on guilt-norms, and so could have one without the other. So Gibbard must say that guilt-norms are acquirable in the way that moral concepts are (or, in his terms: that acquiring moral concepts just *is* acquiring norms for guilt). This, however, leads him into a familiar difficulty, for this answer suggests that guilt depends on fairly sophisticated conceptual resources, and this, in turn, brings up the threat of circularity.

As an illustration of this problem, consider the relationship between an emotion like anger and something like ‘tenure rage,’ that is, anger felt at the denial of tenure. Anger is usually thought to be universal because it doesn’t seem to depend on any parochial, non-universal concepts, if it depends on any concepts at all. Hence it’s hard to see how someone could acquire anger. Tenure rage, on the other hand, requires a number of concepts, because, to experience tenure rage, you have to think that you’ve been

---

89 A person from some distant cultures might not have the concept of carburetor, either, though she could pick it up easily enough. My point is just that moral concepts are are similarly adoptable.

90 This example comes from D’Arms and Jacobson (2003).
denied tenure. This suggests that tenure rage could be acquired, since all that’s required is a fairly simple concept-acquisition.

The idea here is that it’s hard to see how someone without guilt, say, a member of a shame-culture, could come to have that emotion unless guilt is, like tenure rage, a ‘cognitive sharpening’ of existing emotions, or a new way of classifying more basic emotions, as Sabini and Silver suggest it is. These possibilities, in turn, make it look as though guilt depends on the concepts Gibbard uses it to explicate, and so his account falls prey to a more sophisticated version of the circularity objection. This argument is simply an extension of our earlier point that guilt and moral concepts have to be parochial in the same way. There, the motivating thought is that Gibbard has to explain how guilt is culturally limited without relying on the sorts of conceptual content that form the basis of the most plausible explanation. Here, that thought is extended to include not only guilt’s actual extension, but its possible extension as well. If guilt is to serve as the basis for an account of moral concepts, it must be acquirable in the same way. But the best explanation for this modal feature is one that Gibbard cannot invoke.

In passing, it’s interesting to note another guilt-based argument we might use against Gibbard. According to many psychologists, an emotion’s universality is evidence that it’s innate or hardwired (see, for example, Ortony and Turner 1990). If guilt is universal, then, Gibbard will have even more trouble accounting for the modal status of moral concepts. Our guiltless interlocutor—given the assumption of guilt’s status as a cultural universal, we’ll have to make her the hokey old intelligent alien—can acquire

---

91 I take this term from D’Arms and Jacobson (2003)
moral concepts, but not guilt. Hence guilt’s universality is a point in favor of our general modal distinctness strategy, and against Gibbard.

The underlying idea is that Gibbard’s account of moral concepts sits uncomfortably beside certain ways of thinking about guilt. In order to increase the pressure on his analysis, we’ll have to extend our line of argument by showing that one cannot have *norms* for guilt (in the sense Gibbard requires) without having guilt itself, that is, without having guilt in one’s store of accessible emotions. One might think that our alien interlocutor might very well be, as a biological matter, incapable of acquiring guilt, while also thinking that he could still come to understand guilt well enough to make judgments about when it’s warranted or when it makes sense. If so, then our alleged modal asymmetry between guilt and moral concepts comes to nothing, because what really matters here is an asymmetry between norms for guilt and moral concepts. In order to make this a problem for Gibbard, we need to establish that moral concepts have a feature that norms for guilt and anger don’t, namely, moral concepts can be acquired by agents who can’t acquire the capacity for those sentiments. If Gibbard can make sense of the acquisition of guilt-norms even when guilt itself is not similarly accessible, then he can think that guilt is universal (and innate, immutable, and so on) while still maintaining that guilt-norms are acquirable in just the way that moral concepts are.\(^9^3\)

---

\(^9^2\) This assumes, questionably, that different physical realizations of a functionally defined mental state must share a common status as innate or acquired. That is, we’d have to assume that since guilt is innate in humans, it must be innate in any agent.

\(^9^3\) Cheap shot: Gibbard’s view relies on the psychological state of norm-acceptance, which is identified through its characteristic coordinating function, its evolutionary history, and so on. Is this too particular to our species? We might think that the issue of univocity arises at the level of rationality as well, if the state of norm-acceptance is unique to humans—though the imperative function of the norm itself is shared between distinct mechanisms for enabling certain attitudes toward it.
We can begin examining this second point by noting just what the acquisition of guilt-norms would be like. When we think that this is possible, we’re imagining a scenario in which our alien visitor hears us talk about guilt, finds out what situations we think warrant guilt and which don’t, and so on. On the basis of this instruction, she comes to have views of her own about when guilt makes sense—views that aren’t necessarily identical to those in place in our community. (That is, she must be able to think that, in some hard cases, we’ve got it wrong when we endorse guilt, because she’s not simply learning sociological truths about what we think is guiltworthy.) Forming these views allows her to make prescriptions of the relevant sort concerning guilt (and, by extension, anger): she can now endorse the moral sentiments that, ex hypothesi, she cannot feel.94

It’s hard to know just what to say about this possibility. One worry has to do with the familiar problem of circularity. Gibbard, after all, is seeking to explain how this agent without the concept of guilt can nonetheless have norms for guilt and thus can have moral concepts. Thus, he can’t allow that our agent is relying on some independent moral concept, say, [wrong], in order to make these judgments of warrant. Yet, at the same time, our agent can’t be using, as grounds for these norms, some feature of the sentimental response available only to those who experience it, because that’s just what she’s not doing. That is, she can’t appeal to something about the sentiment itself in order to justify or ground her guilt-norms.95 Where else would we turn?

94 Bear in mind, too, that this endorsement must be of a sort relevant to the associated property-ascription, and not based on, say, strategic considerations about the pragmatic benefits of feeling guilt—see D’Arms and Jacobson 2000a.

95 What to say here? To what extent is endorsement of a sentiment going to be free from the troubling concepts even if an experience of the sentiment is? For example, consider this suggestion:
If we could show, in a clear and convincing way, that it’s not possible for an agent to acquire norms for guilt and anger if she does not have guilt in her emotional palette—or at least that this is impossible on the norm-expressivist story—we’d then increase the pressure on Gibbard. He could no longer think of guilt as (biologically) basic, because this would force him to deny our earlier point about the acquisition of moral concepts. I think it’s plausible that moral concepts can be acquired in just the ways that other, unproblematically descriptive concepts are: learning to think moral thoughts shouldn’t be any stranger than learning to think carburetor thoughts. But Gibbard needs a story about how this is so. The problem is that, once we assume that guilt-norms require guilt, in some interesting sense, we need to say that guilt is acquirable in just the same way moral (or other) concepts are. And this adds to the temptation to distinguish guilt by appeal to some propositional or conceptual content of the sort that Gibbard must avoid. So the mystery is how to make guilt sufficiently malleable (in the way that various

The fear system, for instance, can plausibly be described as monitoring the environment for threats to the organism, even if (as neuroscientist Joseph LeDoux claims) there are distinct pathways into the syndrome known as fear: a syndrome of directed attention, physiological changes, affect, and motivation that can be functionally understood as constituting a kind of appraisal of the circumstances. There may be no better way of articulating that appraisal than by saying that it involves construing oneself to be in imminent danger; but it does not follow that, in order to feel fear, one must deploy this or any other concept. (D’Arms and Jacobson 2003: 139)

This is exciting because it promises to save the legitimacy of the intuitive connection between, say, a bout of fear and the judgment that something is dangerous while nonetheless stripping fear of its potentially problematic (for the sentimentalist, anyway) conceptual content. Yet it raises the question of endorsement: what is it to endorse fear but to throw one’s weight behind the motivation, or to grant it reason-giving force? Is this the same as judging to be dangerous? The endorsement, one might think, involves higher-level cognitive faculties than does the system responsible for fear-responses.

Additionally, it’s not clear how this applies in the moral case. If we say, in the case of fear, that fear has a motivational role much like that of the judgment that one is in danger, though it does not itself contain or rely on that bit of content, we’re saying that there are two things with roughly the same motivational force, generally and for the most part: the judgment (or construal, or some other attitude) that one is in danger and the experience of fear. In the moral case, the analogous move would commit us to thinking that there are two mental items with the same motivational import, namely the experience of guilt and the judgment of wrongdoing. This won’t do if we’re trying to reduce one to the other, or explain the judgment in terms of the sentiment. Then we can’t rely on the appeal to judgment to identify the distinct appraisal-like syndrome.
concepts are) without making it depend or include various concepts it’s being used to explain. But, again, this is mysterious only if we think that guilt-norms require guilt, which, thus far, we haven’t demonstrated convincingly.

Thus far our examination of guilt has played a role in several arguments against Gibbard. If moral concepts and guilt are found in just the same cultures, Gibbard will need to convince us, more effectively than he has so far, that guilt doesn’t depend on moral judgment or concepts. Similarly, the acquirability of moral concepts suggests that, if guilt is also acquirable in the same way, it’s so because of a problematic relationship with moral concepts. Finally, the alleged universality of guilt raises questions about how Gibbard’s account might preserve the modal status of moral concepts. All of this suggests that the sentiments are a troubling resource for expressivism, because they threaten the account’s reductive ambitions.

2.10 Blackburn

So far our attention has been on the specification challenge, the demand for an identification, and an account, of the mental states at the core of the noncognitivist’s theory of moral discourse. One response to this challenge is Gibbard’s: we could take up the demand and argue that some particular set of mental states (e.g., norms for guilt and anger) is up to the task. Another response is less direct, and, at first glance, less satisfying: we could reject the challenge, and claim that the expressivist doesn’t need to provide a specification of the mental states at work. This reply must be justified, of course, because on the surface it looks to be a shirking of philosophical responsibility. But if the reply can be vindicated, the expressivist has neatly sidestepped many of the objections we’ve been considering.
Here, I’ll examine the plausibility of this second approach by way of a related puzzle that I’ll call the problem of conative variation. Suppose moral judgments and claims are understood in terms of some noncognitive state or other, be this sentiment, attitude, preference, or whatever. For whatever state is chosen, it seems we can imagine a competent speaker making genuine and sincere moral judgments without that state. Perhaps there are cultures without guilt but with a recognizable moral discourse. Here at home, committed evildoers have constances toward moral demands. Yet both seem, at least initially, to be making real judgments about what’s right. To put it another way, the objection suggests that moral judgment cannot be explicated by appeal to conative states because these conations vary while moral judgment remains constant. The difficulty for the noncognitivist, then, is to explain how it is that nonstandard speakers can (at least appear to) make moral judgments that are of a piece with our own without having the appropriate relationship with the conative states at the theory’s heart.

Notice that these are two distinct issues. Particular answers to the demand for specification might leave the expressivist vulnerable to the problem of conative variation, if, for example, the states in question look to be neither necessary nor sufficient for moral judgment. But the demand brings with it independent dangers as well—e.g., the threat of circularity. If the best account of the expressivist’s mental states of choice sees them as being partly constituted by the moral judgments they’re being used to explain, the specification has failed, and in a way that’s not related to conative variation.

There are several different strategies available to the expressivist. The most obvious is to accept both puzzles and respond with a proposal detailing how certain mental states can account for all instances of moral judgment. One way of making this
response denies that there can be cases of genuine moral judgment apart from the conations on which the theory relies; apparent counterexamples must be reinterpreted or rejected. But such a commitment may not be necessary. Alternatively, we might accept that the demand for specification is legitimate, while thinking that attitudinally deviant speakers are kept within moral conversation—and within the scope of real (dis)agreement—by some sort of appropriate link to more conventional speakers. That is, deviant moralists mean what we mean via some sort of ‘parasitic’ connection to the conversational mainstream. They can engage in moral discourse without the characteristic noncognitive states that are nonetheless fundamental in accounting for that discourse. According to a view like this, we do need to identify characteristically moral attitudes, but we don’t need to argue that every participant in moral discourse is tied to them in just the same way. What’s needed instead is an explanation of the linguistic mechanisms by which deviant speakers are kept in the game. A third option is to reject both challenges. Simon Blackburn, for example, suggests that the demand for specification rests on a confused view of meaning; once we understand that, he thinks we’ll see that conative variation isn’t a problem, because it’s a mistake to require a conceptual connection between moral judgments and some particular noncognitive states. Because there are a number of ways to respond to the problems, we’ll have to be careful to keep track of just what work is being attempted by the replies we’ll consider here.

Let’s put off examining these strategies, though, until we’ve taken a closer look at the problem of conative variation. The problem comes from Nicholas Sturgeon (1986), and it’s a novel twist on the old noncognitivist argument that’s served as the background for so much of our discussion. In its classic form, the argument goes something like this.
Interlocutors have different standards for applying moral terms, since they take different facts to be decisive in moral judgment, and yet they have real (that is, univocal) disagreement. If we understood moral claims to be referring to real properties or entities, then we couldn’t make sense of this disagreement, because the differences in moral standard are great enough to ensure that speakers would be referring to different properties in their dispute about, say, rightness. And so we’d have only verbal disagreement. Thus, to preserve real conflict, we should understand these terms as having something like prescriptive or emotive meaning, because this provides a way of making sense of our intuition that there’s real incompatibility between these speakers’ views. 96

The strategy is simple and powerful. Reference to common entities is supposed to preserve meaning across disputes. The noncognitivist replies with a case in which it appears as though there’s no shared reference in the descriptivist’s sense (because, in these examples, the normative standards in play are so different), and yet we’re intuitively drawn to say that meaning is preserved anyway. This paves the way for the noncognitivist’s claim that something else is going on in moral discourse—because only this additional element could preserve the univocity of these conversations.

The problem is that the same strategy can be used against the noncognitivist. Sturgeon’s objection is simply a reconfiguration of this classic argument. The starting point is the apparent fact that we have to account for constant meaning across variation in attitude as well as belief: some speakers go in for the right and the good, while others are

96 This argument appears in Hare (1952) and at several points in his later work. In Chapter 3 I’ll suggest that it’s also the inspiration for Horgan and Timmons’ Moral Twin Earth argument.
against them because they favor wickedness and evil. Cynics suggest that virtue is something to be scorned, not prized; morality is to be distrusted, they say, because it’s nothing more than a coercive tool. Though they condemn what we value, we still want to say that we share terms with these skeptics and engage in real argument with them. Intuitively, we share meanings and disagree in our normative commitments. If so, then it cannot be that we share meanings because we share noncognitive attitudes (say, about virtue) with our interlocutors, because that’s just what isn’t happening here. If I can manage to have a boo-stance, rather than a conventional hurray-stance, toward doing the right thing, then my judgments of rightness can’t be simply a manifestation of a hurray-stance. Thus, the noncognitivist proposal is faced with just the same objection raised against descriptivism. We can’t save real disagreement by appeals to attitudes if those attitudes can vary between speakers who are nonetheless engaged in moral dispute.

The difficulty is neatly illustrated by Sturgeon’s example, the conversation between Socrates and Thrasymachus in the first book of the Republic. The two interlocutors have different ideas about what justice is, and, of course, they have different attitudes toward it: Socrates favors it, while Thrasymachus does not. The task for the expressivist is to explain how they can be talking about the same thing. This, clearly, is a close relative of the problem we’ve been discussing in relation to Gibbard. There, and here, the issue is whether or not there’s a single attitude of the right sort that accompanies all moral judgments. I’ve set the problem in terms of the alleged cultural parochiality of guilt, and Sturgeon’s objection hinges on the possibility of making moral judgments despite having attitudes that are out of step with those shared by one’s community, but
the central idea is the same.\textsuperscript{97} In both cases, we have ostensibly moral judgments without the mental state that the expressivist claims is constitutive of those judgments. If there’s no one attitude associated with all moral judgments, it looks as though expressivism’s classic argument has turned on its creators.

Let’s be clear about just what’s at issue here. Expressivism tries to account for moral judgment and discourse by appealing to noncognitive states: attitudes, sentiments, desires, norm-acceptances, and so on. The problem is simply that these mental states seem to come apart from the judgments they’re being used to explain. Thrasymachus and other amoralists (the Foole, the Sensible Knave, and so on) are graphic ways of illustrating the challenge, because the stances they take toward rightness, virtue, justice...are diametrically opposed to our own. But it’s important to see that these are just the most extreme examples of a common phenomenon. The underlying problem is simply that there seems to be an impressive degree of conative variation in speakers who are making \textit{bona fide} moral judgments, and this fact requires explanation.\textsuperscript{98}

This qualification is meant to highlight two noteworthy lessons. First, what we say about Thrasymachus in particular (or about other radical amoralists) is less important

\textsuperscript{97} In order to make this objection work, Sturgeon will have to argue that amoralists (of the sort who endorse wickedness or at least fail to say ‘hurrah!’ for virtue) are making genuine moral judgments in spite of their deviant attitudes. On the other hand, I’ll have to argue that guiltless peoples are engaging in moral discourse, in the appropriate sense, if I want to make troubles for Gibbard.

I find the idea that amoralists are making genuine moral judgments quite plausible, though, of course, there are many people who reject it. One thing that makes Sturgeon’s task a bit easier is that we do have a strong intuition that the problem with people like Thrasymachus isn’t a \textit{linguistic} problem. We think there’s serious normative disagreement here, and people who think that his judgments aren’t really about justice owe us a story as to what’s going on. To further complicate matters, it’s hard to say that he’s just making an ‘inverted commas’ judgment, because he also has nonstandard ideas about what justice is, and so he’s not just using the standards commonly accepted in the community. For more on this, see Sturgeon (1986), Smith (1994), Brink (1989).

\textsuperscript{98} The literature is a bit misleading on this, since critics have tended to focus on dramatic cases of conative variation, e.g., speakers who condemn the right and the good. I suspect this is so partly because of the noncognitivist’s failure to provide an answer to the specification challenge—without a firm grasp on which state is being put forward, it’s hard to give a subtle counterexample along these lines.
than our account of the general phenomenon. We shouldn’t think that the problem is with *this speaker* in particular, or with speakers equally deviant, because the issue is broader than this. The same issue is raised by more moderate cases of conative variation.

Second, we should be aware that there are a range of cases here that might be best treated in different ways. Speakers like Thrasymachus might be ruled out of the conversation, but other, less deviant speakers might be more plausibly interpreted as making actual moral judgments despite having (somewhat) nonstandard conative responses (for example, some moralists might have unusual sorts of pro-stances toward rightness). Of course, we’re then owed a story about how *these* speakers stay in the conversation. As we’ll see, the variations on this basic theme require us to think carefully about just what’s demanded of the expressivist. Just how much variation, and within what parameters, is consistent with a univocal moral discourse? Which deviant speakers are still within the conversation, and how are they connected to the conversational mainstream? We’ll have to keep in mind that there are a range of responses available here and that the most plausible expressivist position may be a compromise between more extreme options.

In order to account for conative variation, the expressivist must either convince us that, despite the initial appearances, attitudes (or sentiments, stances...) *don’t* vary across speakers who share moral terms in the way that Sturgeon’s argument suggests, or show how a recognizably expressivist view can account for the variations in attitude that are ostensibly out of reach. There are a few different possibilities. First, we might deny that Thrasymachus and similar amoralists are making real moral judgments. Thus, we’d discard or reject alleged counterexamples of conative variation by claiming that these cases *don’t* present us with speakers competent with moral discourse. Second, we could
reject traditional demands for a theory of meaning. We might say instead that there’s no
question about real agreement or disagreement apart from our norms for interpreting
other speakers. As we’ll see, this appears to be Blackburn’s response to the problem,
though it lands him with some additional difficulties. Third, we might look for ways to
accept the specification challenge, admit that individual speakers can make moral
judgments without a particular conative state, and show how those speakers could remain
in the conversation, perhaps through their ties to other speakers. I’ll examine each of
these strategies in turn and say something about their costs and benefits.

Let’s start with what might be the simplest, but also the least intuitively attractive,
line of response: rejecting the amoralist’s challenge to expressivism on the grounds that
such a character isn’t plausibly thought to be entering into real disagreement with more
traditionally-minded interlocutors. According to this line of argument, we don’t have a
case of genuine moral assessment in the absence of the characteristic pro-stances that are
said to explain these judgments, so the expressivist’s initial response is the correct one,
and the objection fails. Instead of explaining how the expressivist can accommodate
certain deviant speakers, the reply insists that they oughtn’t be accommodated.

This strategy might be particularly appealing when it comes to Thrasymachus.
Here we have a character who disagrees radically with conventional use both in what he
takes justice to be (the advantage of the stronger, the interests of the ruler, and so on) and
in his responses to it (contempt, derision). Granted, we don’t know a great deal about
how he would defend his view in extended conversation, what reasons he gives for
thinking that justice is simply the interest of the stronger, which acts he thinks are just,
and so on. But as it stands it’s certainly fair to wonder if and exactly why we interpret
Thrasy machus as sharing our moral terms and concepts rather than as someone engaged in linguistic perversity. After all, there are some types of eccentricity that encourage us to label a speaker as incompetent: if Thrasy machus had condemned justice and then pointed toward acts that we consider treacherous or cruel in support of his condemnation, we’d certainly be tempted to think that he’s just confused about how we use the term. He isn’t really making any claims about justice, we’d think; he’s making claims about something else and mislabelling it. All of this is just to say that we need some reason to credit him with our moral concepts and terms.

For now, I won’t say anything about the complex and interesting issue of just how we should decide whether Thrasy machus is a speaker within our conversation or outside of it. While this is an important question, it’s not crucial to our task here, because there are more intuitively compelling examples available. Thrasy machus may (or may not) strain our interpretive intuitions, because of the degree of his eccentricity, but there are cases closer to home that do the same work. For our purposes in examining the basic problem, we need only to think that some sort of variation along these lines is possible among competent speakers. Consider a character with an unusually exigent take on the demands of morality: by his lights, we’re required to give much of what we have to the poor, and while we’re at it we should be nicer to animals. Faltering under his extraordinary burden, he becomes disenchanted with the practice of moralizing. This person—call him Cold Peter—scorns morality as oppressive, because incompatible with a good life, and comes to frown on doing the right thing. Cold Peter has nonstandard beliefs about the extension of moral predicates, an unusual view about the foundations of morality, and atypical conative attitudes toward its commands. Yet it’s plausible to
maintain that he shares our terms. The general point is that we shouldn’t let qualms about Thrasyrmachus’s deviance undermine the force of the anti-expressivist argument, which has intuitive impact even if particular would-be amoralists are ruled out of our conversation.\textsuperscript{99} Even supposing that ruling out Thrasyrmachus in particular is a plausible reply, this general strategy has a mark against it, insofar as there are relevantly similar cases that do the same work.\textsuperscript{100}

What’s more, this reply doesn’t do very much work in making sense of conative variation generally. Thrasyrmachus and Cold Peter are at one end of the spectrum of possible affective responses to rightness judgments, but, of course, there are others. Speakers might respond to moral judgments with different sorts of pro-attitudes, or with complete indifference. Simply ruling out a certain level of deviance doesn’t provide a general mechanism that accounts for this sort of variation. By suggesting that a

\textsuperscript{99} Note that variation in the amoralist’s standards (his account of what acts are right and why, say) makes a familiar noncognitivist strategy difficult to invoke. It won’t be possible to say that the amoralist and the more traditional speaker share descriptive meaning if there are different standards at work. If the noncognitivist allows for shared descriptive meaning across different standards, then the standard noncognitivist argument, that descriptive accounts cannot preserve univocity in cases of disagreement, is undermined. See Sturgeon (1991: 24–25).

\textsuperscript{100} Suppose we did want to say that Thrasyrmachus fails to make genuine moral judgments in virtue of his conative deviance. What positive story could we offer about what he is doing? One traditional answer is the ‘inverted commas’ reading, which, as far as I know, first surfaces in the work of Hare (1963). According to this interpretation, the amoralist’s claim that an action is “wrong” should be glossed as meaning something like “this action is of a type that will be judged wrong [or condemned, met with con-attitudes...] by conventional members of my speech community.” What’s interesting about this interpretation, for our purposes, is that it relies on a tie between the deviant speaker and more conventional community members to provide an account of the meaning of the amoralist’s utterance. As we’ll see, this basic idea will be invoked by other responses to the conative variation problem.

The inverted commas interpretation is generally thought to be unsatisfying for two reasons. First, it changes the meaning of the amoralist’s use of “wrong,” because it understands the amoralist to be committing himself to a kind of sociological point about the reactions of various community members. So this reading doesn’t preserve univocity. Of course, if we think that Thrasyrmachus isn’t making real moral judgments then this is the right outcome. Second, this reading doesn’t make sense of the interesting amoralist, for example, Thrasyrmachus himself, who scorns morality while knowing that he has nonstandard views about the right- and wrong-making features of actions. It’s perfectly coherent for Thrasyrmachus or Cold Peter to acknowledge that most speakers, confused as they are, disagree with him about the extension of “right,” and thus it’s unsatisfying to say that their uses of “right” are merely predicting the responses of others.
difference in attitude is sufficient for a difference in content, this line of reply implicitly buys into the idea that moral judgment is what it is in virtue of its link to some conative state. The challenge to account for (a more modest degree of) conative variation, and to specify just what states are involved, is still on the table. This isn’t to say that the first reply goes wrong because it fails to address important problems with expressivism, but it is to say that the response leaves many important questions unanswered.

The first reply, then, has a few drawbacks. It leaves the projectivist committed to an implausible interpretation of deviant moralists such as Thrasymachus and Cold Peter, and it gives no help with the specification challenge. As we’ll see, the other options are more ambitious, because they attempt to counter the objection from conative variation by showing how no particular noncognitive attitude is required for participation in moral discourse. If these replies work, then the expressivist has defused Sturgeon’s objection, and may have also changed the terms of the debate over the specification problem.

Unlike the reply we’ve just discussed, the next responses reject the idea that, in order to make moral claims, speakers must have the appropriate relationship with the relevant mental states. Instead, these responses claim, individuals needn’t stand in these relations in order to moralize. There’s a sense in which this strategy is a plausible and natural one for the expressivist to take up. Developments in post-positivistic philosophy of language have urged us to think that meanings ‘ain’t in the head’ and that the content of our thoughts and utterances is determined partly by facts external to individual

---

101 To make the first response is to address a particular instance of a general problem, but it leaves the problem in place and it doesn’t so much as hint at the resources we’d use in solving it. My suspicion is that responding to the general point will require some of the moves discussed later in this section, which render the first reply more or less irrelevant.
An upshot of this thought is that individual variations are less important in determining univocality than we might have previously thought. There are ways to keep nonstandard speakers in the conversation; you can talk about something without having any particular belief about it. This thought, of course, has been important to recent moral realists, because it helps explain how widespread disagreement about moral questions is compatible with shared meaning. It provides room for some independence between what’s in the head (in terms of your beliefs or theories) and what you’re talking about. The problem for expressivism, on the other hand, is that it looks as though there’s no single attitude (or sentiment, feeling, norm-acceptance...) common to all moral judgment and conversation. If we want to maintain that moral discourse is based on attitude in some interesting way, we’ll need to show that this can be so despite the fluctuations in conative states that Sturgeon’s argument brings to mind. Perhaps some of the same moves in the philosophy of language will be useful here as well, because they might enable the expressivist to make room for the required independence between moral judgment and occurrent attitudes. There would be poetic justice if this strategy can be made to work: Sturgeon’s objection takes up an old noncognitivist argument and turns it against that position, only to be met by the same resources in the philosophy of language that gave new life to naturalistic realism.


Tyler Burge’s work has dealt with this issue extensively. See (1979) and (1986) in particular. We don’t need to agree with all of Burge’s views to agree with him in thinking that an adequate account of concepts and meaning will allow for speakers with nonstandard theories about things to nonetheless think and speak about them.

The similarity between the expressivist’s needs and those of the realist is roughly this. The realist is often tempted to rely on various externalist suggestions in order to argue that diverse cognizers can have radically different beliefs about the same subject matter, whether the object of belief is arthritis, gold, or rightness. This requires a view about how individual deviant speakers can have beliefs about those things, perhaps in virtue of their ties to the stuff in question, to more competent members of the speech
With this thought in mind, let’s look at the next possible reply, which is a bit more radical. It involves rejecting the idea that there are semantic facts, prior to our interpretations of speakers, that determine whether or not we share meanings or talk past one another. This reply is suggested by Simon Blackburn in some remarks on Sturgeon’s argument, and it’s worth considering here because, if it’s successful, it answers that argument while also giving us a way of rejecting the specification demand that we’ve been discussing. Blackburn’s proposal seems to be this, more or less: there are no independent facts about meaning; our interpretive norms aren’t tracking some independent semantic features of the world. In interpreting others in our speech community, we start with a default position that reflects certain paradigmatic uses. We presume that others use words in this way, and mean what we mean, just so long as they remain within a range of acceptable standards. Wild idiosyncracy overturns the presumption of univocal conversation. Nowhere in any of this, though, is there a fact about whether or not interlocutors mean the same thing. Instead, we have only interpretations. This, Blackburn thinks, is a way of avoiding the Thrasymachus community, or to the relevant experts. That is, what’s needed is a story about how individual speakers relate to various parts of the community in a way that allows them to have beliefs about things they misunderstand to a significant degree. Likewise, the expressivist needs a similar account of how a deviant speaker might share terms with a larger community while differing with them in important respects, in this case, in his conative states. What’s required is a way of divorcing moral claims from particular attitudes in the individual case, though not necessarily the communal case.

There’s also an important difference between the two challenges. The realist wants to account for moral judgment in terms of belief, and so what’s needed is an account of how cognizers can have radically different beliefs that are nonetheless about the same thing (say, a particular moral property). The puzzle is to show that there’s a shared object of belief and reference: how can these speakers share a subject matter if their beliefs are so different? The expressivist, on the other hand, doesn’t need to show that Thrasymachus really has a pro-stance toward justice despite evidence to the contrary. What’s needed is a way of showing how particular speakers can have moral concepts (think thoughts about what’s right, make moral judgments, and so on, with all of these potentially problematic phrases translated into some appropriate noncognitivist idiom) without having characteristically moral attitudes.
objection, and the broader problem of conative variation. It’s also a way of getting out of
the specification challenge: the reply

diminish[es] the force of one attack often made on projectivism. This is the
difficulty over defining the state of mind projected in other than ethical terms
(approval, guilt, etc.). I reply that...we would not want a definition, serving only
to rule out cases that we need to rule in. (Blackburn 1991: 9)

If Blackburn’s view is right, many of the expressivist’s problems disappear.

Before evaluating this approach, we should look at how Blackburn supports his
view, because doing so will give us a better idea of the challenges facing the projectivist.
He begins by noting that the problem here is a familiar one from other parts of language:
in ethics, as elsewhere, we have to preserve “uniformity in sense, even in the face of
massive differences of cognitive shape” (Blackburn 1991: 5). People have different
standards for applying a term such as “elm,” and they have various responses to elm-
judgments, yet we remain confident that our talk of elms is not mere linguistic confusion.
So the challenge is not one peculiar to ethical discourse.

He then suggests, plausibly enough, that the same devices used to preserve non-
evaluative content can be taken on board in ethics as well. In both cases, individual
speakers aren’t the final authority on how to use terms. The mechanisms are more
complex:

[A] group speaking a language holds itself together with what Putnam called
‘division of linguistic labor’ and what in Spreading the Word I called deferential
conventions. In the clearest cases, we institutionalize an authority, and anyone
using a term is liable to be regarded as having used it with the meaning vested in
it by the authority...Anyone using a term is therefore in principle up for being told
what it means, and his own peculiar configuration of knowledge and ignorance is
not the arbiter of it. We might say that in this picture sameness of meaning or
content is something that is imposed, although only on someone who himself
abides by, or recognizes the authority of, the convention. (Blackburn 1991: 5)
We can apply this in the ethical realm as well; there, as in other parts of language, we need to preserve univocity despite differences in individual speakers. In our moralizing, as elsewhere, we do this by

presuming a common, communicable topic. The ‘default’ position in any conversation is that we are discussing the same issue, even if we have different opinions about it. Only wild idiosyncracy overturns the default. (Blackburn 1991: 6)

The projectivist, Blackburn claims, has no problems adopting all of this. He can do so by “imposing as much identity of content as the need for communication and debate requires, hence, respecting the possibility of a common topic through different standards and responses” (p. 6). Thus, the response involves two claims: that ethical discourse is no different from others in its approach to the problem of preserving content, and that in ethics, as elsewhere, what keeps conversation going is a presumption of common subject matter.\(^\text{105}\)

There are a number of ways in which this strategy abandons traditional expressivist arguments and themes. Let’s begin evaluating this by looking at an odd feature of Blackburn’s discussion: his use of deferential conventions or the idea of a division of linguistic labor. Originally, of course, these ideas were part of a story about how a speaker might manage to refer to elms (for example) despite her inability to give a

---

\(^{105}\) Further support for this reading of Blackburn comes from an objection he makes against realism. The problem he poses concerns the possibility of misinterpretation: if the realist is right, there might be ‘hidden facts’ about the reference of a speaker’s moral term. We might take two speakers with similar (or identical) standards and reactions to be talking about the same thing, though in fact they’re “locked on” to different properties. This is thought to be implausible. “For the anti-realist,” he writes, “there is no such possibility [of interpretive error]: the interpretation is, as it were, downwind of our interpretative needs” (Blackburn 1991: 11). I take this to mean that, on Blackburn’s view, the interpretation is driven by our “interpretative needs” (such as the need to preserve genuine argument, and so on) rather than by some antecedent semantic fact.
sufficiently exacting account of what makes something an elm. It’s partly in virtue of this speaker’s willingness to defer to the elm-judgments of the community’s experts that she’s making claims and judgments about elms. All of this is to say that Putnam (along with Burge, Rey, and others) emphasizes deference in order to explain how incomplete understanding doesn’t (always) prevent speakers from referring to, and thinking about, things they misunderstand. Thus talk of deference originally surfaced as a way of saying how there is a fact of the matter about meaning and reference, a fact that relies on, among other things, various causal relations, and not just what’s ‘in the head’ or internal to the speaker. So it’s at least a little bit odd to adopt this talk as part of a view that denies the existence of meaning-facts.

There’s a problem here that’s deeper than simple incongruity. One of the most persistent, and most powerful, motivations for expressivist readings of moral discourse is the idea that our best solutions to various problems about the workings of empirical

---

Note, too, that both the experts and the neophyte can be wrong about elms, and that the neophyte’s claims are about elms, not about what some privileged group of speakers would say. That is, her utterances shouldn’t be understood to mean something like “this tree will be called ‘elm’ by the appropriate authorities.”

There’s an additional, and interesting, oddity here concerning authority. The idea of scientific authority is relatively unproblematic, but in the normative case it’s at least controversial. Moral discourse doesn’t have experts in quite the same way that, say, physics does. Consider the vaguely existentialist thought that one must choose to submit to some moral authority (the Pope, Peter Singer, or whoever), and so even those who obey authority do so by some kind of volition—the were free to do otherwise. This thought doesn’t have the same resonance in the case of empirical inquiry: it doesn’t make a great deal of sense to say that one could believe what Richard Dawkins tells you about Darwin, but then again one could always choose not to. This is right, in some sense, but it seems clear that there’s a difference between the two cases. Or, for another way of making the same point, contrast these two claims: (a) I haven’t seen the proof myself, but it’s valid; (b) I haven’t seen the film, but it’s brilliant. The first is kosher, the second isn’t. (This example is due to Pamela Hieronymi).

Examining the philosophical significance of these points must wait for another day. I raise them here to suggest that there’s something puzzling about Blackburn’s invocation of the idea of deference, because the kinds of deference often discussed in the literature about concepts and meaning is just the sort of thing that’s unproblematic in the case of elms or arthritis but quite tricky in the case of value quite generally. And, to make things worse, the features of value discourse that suggest authority relations are more dubious are just those that often motivate expressivist views of moral discourse. So Blackburn’s view here is unconventional in the resources on which it draws.
language *fail* in the case of normative talk. That is, we might be tempted by expressivism about morality precisely because it’s so different from everyday empirical discourse *and* because those differences (apparently or allegedly) make it impossible to apply the same strategies in the normative case. One classic instance of this argumentative strategy involves the issue of shared content and real disagreement: expressivists (and other opponents of a certain kind of realism) claim that what works for “arthritis” and “gold” doesn’t work for, say, “justice” or “right.” Blackburn hints at a newfound respect for the seamlessness of language; he seems to think that the same mechanisms can explain shared content in normative and non-normative cases. But this undermines one of expressivism’s traditional motivations. To say that univocity can be preserved by just the same means in both sorts of cases is to give up on the argument that motivates expressivism by appeal to the *failure* of various mechanisms when it comes to moral discourse.

Blackburn’s remarks about conative variation suggest another retreat from an expressivist stronghold. If he is serious about allowing speakers such as Thrasymachus to remain within moral conversation, then he has to give up some popular *internalist* arguments for projectivism, arguments which he had endorsed in earlier writings. It follows from his response to Sturgeon that one can make moral judgments without any particular conative state, and, more specifically, without any pull toward action. Thrasymachus, after all, scorns justice, so his judgments about just acts won’t have the

---

108 Horgan and Timmons’ “Moral Twin Earth” argument is a good example of this kind of strategy, though they aren’t expressivists. Their argument is meant to establish something roughly like this: the best semantic strategies for doing certain kinds of work with empirical discourse (e.g., a causal view of reference) fail to apply successfully in the moral case. Hence, the naturalistic realist can’t rely on these semantic strategies as a way of showing that moral language works the same way as empirical discourse.

same motivational pull as Socrates’ judgments. If we take on the challenge of Thrasymachus by attempting to show how he too is using our moral terms, we can’t claim that his judgments fail to qualify as moral in virtue of his negative conation, and thus we must admit that moral judgments can be motivationally inert. Granted, the question about the motivating power of moral judgment is a controversial one, but it’s a bit odd that Blackburn would come to endorse the traditional (naturalistic) realist’s view of the phenomenon, if not his explanation for it.

Next, and along the same lines, consider that Blackburn’s new position prevents him from offering the traditional expressivist account of what is going on in moral judgment. If no single response is required, then it can’t be that when we engage in moral discourse we’re gilding and staining a normatively drab reality with the colors of sentiment or projected response. No conative element (not just no particular element, but none at all, presumably) need be present for moral judgment, and so it looks as though making the judgment is something distinct from expressing or evincing that response. There may be other ways of accounting for what is going on that give pride of place to these sorts of mental states, and which are compatible with Blackburn’s response, but one traditional way of marshalling these resources is now off limits. We can no longer say that individual instances of moral judgment are best understood as traditional noncognitivists would have us believe.\footnote{See also Sturgeon (1991: 26).}

Another problem comes from an issue we’ve been considering in connection with the early noncognitivists, the problem of distinguishing moral discourse from other forms of evaluation. Obviously, Blackburn can’t do this by appealing to particular sentiments,
as Gibbard does, because that’s just what’s missing from Thrasymachus, who, for the
time being, we’re trying to accommodate as a (somewhat nonstandard) user of moral
language. Nor can we appeal to some kind of distinct role in practical reasoning,
independent of sentiment, for the same reason. So, at the very least, this view leaves us
with a puzzle about how we might distinguish moral commitments from those of
aesthetics, etiquette, prudence, and so on. Resolution of the puzzle seems difficult, given
that some of the most plausible contenders (common reference and response) have been
eliminated from the running.

Finally, there’s a methodological worry here that suggests we should see
Blackburn’s proposal as, at best, a last resort. His remarks suggest that this is a view
about all ostensibly assertoric discourse; it’s not limited to moral talk in particular. (If
it’s simply a view about moral language, Blackburn can no longer claim that he’s simply
adopting familiar work from other debates, though, as I’ve suggested, this claim is
problematic; what’s more, he’s committed to an apparently ad hoc distinction). Adopting
this view, then, involves all sorts of controversial theses about the workings of language
in general, and the expressivist takes on the burden of showing how more conventional
views of meaning are untenable even when it comes to nonmoral matters. Now,
meaning-irrealism may turn out to be right, at the end of the day; that it’s a controversial
thesis doesn’t mean that it’s indefensible. But its motivations should come from general
views about language rather than from ethics. Blackburn’s remarks put the expressivist
in an awkward dialectical position: he’s committed to defending a controversial view
about language based not on considerations having to do with language generally but
with the workings of one of its parts.
All of these problems suggest that the position we’ve been considering is a problematic one.¹¹¹ Let’s turn to a more moderate view that might do the same work, one that’s hinted at in some of Blackburn’s remarks on Sturgeon’s objections. The starting point is an intuitively plausible, if somewhat unclear, remark about the purpose of moral discourse. The point of having a moral practice is a practical one: we moralize in order to influence action and character, among other things. As Blackburn puts it, though he doesn’t do much to develop the thought, moralizing and valuing are different from descriptive talk in that “[t]heir point is conative...[t]hat is what they are for” (Blackburn 1991b: 39). The role of the discourse is not to describe but to direct.¹¹²

This is a natural thought, foreshadowed in themes from Stevenson and Hare, and suggested again in some of Blackburn’s later work. What we need now is a way of developing this exciting idea into an account of the meaning of moral terms, or, more loosely, an account of “what it is we’re doing” in engaging in moral talk. Perhaps the thought is something like this: just as deviant speakers rely on the conversation mainstream in their use of, say, “arthritis,” the attitudinally nonstandard moralist is kept within the conversation by some sort of tie to other, more conventional members of the speech community. The rough idea, presumably, is to institutionalize a particular conative function as the specific bit of work performed by a moral term, and then allow for speakers on the fringe to use this term, and mean by it what we mean, despite their nonstandard attitudes.

¹¹¹ I should note some unease about my interpretation of Blackburn, though I think it’s natural to read him this way given what he says.
¹¹² See Blackburn (1999: 213- 214) for another articulation of this idea.
(Blackburn suggests something along these lines when he says that amoralists, or those who condemn moral commitment, are “parasitic” on those with better attitudes [Blackburn 1998: 61]. In a brief discussion of the internalism/externalism dispute, he grants that “externalists can win individual battles” by pointing to cases like Thrasymachus, but “internalists win the war for all that, in the sense that these cases are necessarily parasitic, and what they are parasitic upon is a background connection between ethics and motivation” [Blackburn 1998: 61]. This is promising, but he never says just what this parasitic relationship is. To make matters worse, he later hints that he has in mind nothing more than the old inverted commas reading, as when he claims that “[t]he [deviant] agent is using evaluative vocabulary in a parasitic way, as mere labelling for what other people regard as good” [Blackburn 1998: 65]. This, as I’ve suggested, is not sufficient.)

Suppose we start by identifying, at least in a rough-and-ready way, functions for particular moral terms, for example, our old friend “right.” One thing to notice is that the functions of the individual terms embedded in moral discourse will be narrower than the function of that discourse generally. The point of having a predicate such as “right,” for example, is to do some of the work of moral discourse: “right,” we might think, is used to commend actions and put them forward as choiceworthy, while other terms will allow us to condemn and forbid. The idea here is that if moral discourse as a whole has the function of shaping conation, then individual parts of that discourse will have more

---

113 Blackburn’s own example in this discussion is Satan in *Paradise Lost*, but the point is the same.
finely-specified conative roles. The term can fulfill this function only in virtue of linguistic conventions governing its use, just as a term like “table” can perform certain descriptive functions only if conventions tie it to particular objects. In the moral case, this tie between term and conative function might work in one of several ways: for example, Stevenson’s view of meaning accounted for a term’s impact on conation in terms of its causal dispositions, while more straightforwardly expressivistic accounts saw a term such as “right” as being the linguistic vehicle for the expression of certain moral sentiments. But there’s some linguistic convention in virtue of which “right” plays this role and not some other. Without this sort of convention, it’s hard to see how moralizing would serve its distinct function, because its constitutive terms couldn’t be linked to particular parts of that function.

Notice, as a qualification of this point, that we’ve so far said nothing about just how fine-grained these functional specifications must be. The point, for the time being, is simply that “right” must serve some function that’s narrower than that of moral discourse generally: if the discourse is to do its work, we need to specify a more detailed role for, e.g., “right” than for moralizing as a whole. Just how narrow that role must be is, so far, an open question. This basic strategy is compatible with various ways of limiting the conventionally accepted uses of “right.” Inspired by Gibbard, we might think that it’s tied to (norms for) particular sentiments, or we might think that the conventions governing its use tie it only to some endorsement or other, regardless of just what mental states are involved. And, of course, there are various points between these

---

114 If not, then it’s hard to see how moral discourse as a whole will perform its function. It won’t do to say that the point of “right” is sometimes to commend, other times to condone, while still saying that it should be understood in terms of conative function.
two options. Because of the flexibility of this suggested third response, we should keep in mind that different developments of it will have different things to say about the scope of moral judgment, disagreement, and so on, and will have different responses to the particular challenges presented by Thrasymachus, Cold Peter, and their cohorts.

With this qualification in mind, let’s look at how these points bear on our puzzle. The mystery is how a deviant speaker (Thrasymachus, for example, though the same points apply to others) can do three things simultaneously. First, he must make sincere moral judgments, rather than engaging in some kind of linguistic subterfuge. Second (possibly required by the first), he must be in the right sort of relationship with various linguistic conventions governing (or constituting?) the meanings of the terms he uses to express his judgment. Finally, he must have attitudes that go against the grain: he must have, and urge on others, attitudes that go against the conventionally enshrined functional role of a term such as “right.” It’s hard to see how these three are compatible, which is why attributions of insincerity or linguistic incompetence are so tempting if our amoralist is to hold on to his deviant views.¹¹⁵

What makes the amoralist so puzzling is that he refuses to accept the sorts of linguistic conventions that set aside the term “right” for hortatory purposes. Contrast the amoralist with Burge’s arthritic patient. His willingness to defer to his doctor is one reason we’re willing to understand him as meaning what we mean by “arthritis.” Accepting the right authorities is important because it signifies a commitment to use the term as others do, that is, a willingness to respect the conventions governing its use. Without this, we’d think, the speaker would not be making claims about arthritis at all.

¹¹⁵ Compare using assertoric claims in a way “contrary to the point of the practice.”
His deference, though, allows him to be hooked up in the right sort of way with the ailment, if indirectly, despite his strange views about it. The amoralist, of course, isn’t like this. His recalcitrance would be unproblematic, for this sort of proposal, if it were restricted to the descriptive features of disputed acts—but it also extends to the attitude he takes toward the objects of his moral judgment. He doesn’t take the conventionally-mandated stance toward them, and he refuses to concede that the term he uses is for commending actions (at least according to the view we’re considering). We might say that he isn’t playing by the rules. Nonstandard speakers can stay in the conversation in spite of their unusual beliefs, but this isn’t to say that there are no requirements on their participation. Attempts to keep the amoralist in our discussions will be frustrated by his refusal to grant any concession on either front. Since he doesn’t stand in the right sorts of relationships with more conventional speakers, it’s hard to see he’s semantically on a par, at least according to the expressivist story.

The general problem is that the deviant speaker seems to fail to respect the linguistic conventions relating a term to its conative role. Since some sort of convention is required to connect a term and a particular function, it’s not clear how a speaker who doesn’t stand in the right sort of relationship with it should count as using the same terms as mainstream speakers. The amoralist who uses “right” as a device for condemning rather than commending is one example, but, depending on the specificity of the conventions involved, there may be others.\textsuperscript{116} If the requisite mental states are identified

\textsuperscript{116} There’s another way of approaching the problem. The basic strategy here is to offer an account of our moralizing that’s based on conative states without committing to the claim that every individual speaker must be tied to those states in the same way. One way of doing this is to shift attention from the speakers to the language they share. If we could give an account of the meanings of terms that’s not so closely tied
in a fine-grained or quite specific way, then more speakers will count as deviant. On the other hand, if the conventions lay down only a broad range of appropriate response (for example, by linking a term like “right” just to some form of endorsement) then only the seriously marginal cases, such as Thrasymachus or Cold Peter, will qualify as conversationally suspect.

This suggests another expressivistic alternative. Perhaps we cast the net too wide when we tried to capture shared meaning in these extreme cases. The obvious remedy is to give up on Thrasymachus (or to reinterpret him, as I suggest below) and take on more modest ambitions. If we grant that only speakers who endorse, say, justice and rightness count as competent speakers, we take on an easier task of showing how attitudinal deviance within the family of pro-stances is compatible with univocity.

This would provide some important benefits, though it leaves an obvious explanatory gap. The benefits are these. First, it would restore the internalist argument to the mental states of particular speakers, we might preserve semantic continuity in the face of (individual) conative variation.

Consider the development of Stevenson’s first pattern of analysis. In its first appearance, it viewed moral claims through the lens of individual attitudes and commands; to say that something is good is to report one’s own attitudes and command other to feel similarly. Later, though, as we’ve seen, Stevenson switches to a linguistic vehicle that doesn’t wear its coercive force on its sleeve—emotive meaning, understood as the causal dispositions of a word to affect the mental states of its hearers.

This is an exciting proposal for our purposes here because it allows for a close but not overly intimate tie between individual speakers’ conations and the meaning of moral language. Because the emotive meaning of a term involves the effects it typically has on its audience, meaning remains constant through occasional variation. The meaning of the term is understood by appeal to attitudes, but that meaning remains constant despite deviant speakers’ use of the term in conjunction with atypical conative responses.

Unfortunately the Stevensonian view of meaning is not a very plausible one, and we might suspect it errs too far in the other direction: rather than ruling out radical amoralists, it might rule in too many speakers who have no grasp of a term’s meaning. But it’s worth pausing to see what it gets right. It sees the conventions of meaning as partly independent of individual speakers while still depending on the community as a whole. Because of this, it can allow for variation around the fringes of moral talk—that is, it can preserve moral meaning in the case of the deviant speaker—while still holding on to the claim that noncognitive attitudes are the foundation of our moral practice. What’s needed now is a way of preserving these insights in a more appealing theory of meaning.
for noncognitivism, a perennial favorite among supporters of that position. If correct use of, say, “right” demands some sort of favoring of actions that are judged to be right, then there is a necessary connection between rightness-judgments and pro tanto motivation. Second, it would allow for at least some conative variation between speakers, which seems plausible—different moralists might respond to wrongdoing or virtue with various sentiments and remain in the conversation.

There are, however, corresponding problems. First, this sort of proposal doesn’t provide us with a way to distinguish moral discourse from other forms of evaluation. All we know so far, on this proposal, is that moral judgment consists in some form of pro-attitude, but there’s no particular attitude that need be involved. This rules out Gibbard’s strategy of discriminating between judgments of, say, aesthetics, propriety, and morals by appealing to the differences between guilt and anger, aesthetic admiration, and shock. This is not to say that these individuations are impossible on this scheme, but they can’t be made via one plausible avenue, and it’s at least initially unclear how they will be secured.

Second, there’s a related problem here concerning the preservation of disagreement. As we noted in thinking about the beautiful but shocking artwork, different sorts of evaluations pull in different directions without thereby being genuinely incompatible. This compatibility is preserved, on Gibbard’s picture, by appealing to the relationships between mental states underlying the judgments. Competing moral evaluations presumably will conflict, but, according to this latest suggestion, there’s no

---

117 Perhaps someone could violate this sort of rule out of ignorance, in the way that Burge’s patient violates rules about the application of “arthritis,” while still counting as a user of the term.
single attitude undergirding them. Hence their incompatibility isn’t ensured by saying that interlocutors have ordered incompatible sentiments on each other. The obvious alternative is to rely on something like a command to action—different pro-attitudes toward ing all urge us toward ing—but this, as we’ve seen, is unsatisfying, because of the complicated relationship between moral decisions and endorsements of action.

Finally, there’s a dialectical cost to limiting the extent of conative variation. According to this view, competent speakers can’t condemn morality with the kind of obdurate commitment displayed by Thrasymachus and other committed amoralists. I’ve suggested that with Thrasymachus himself this may not be so counterintuitive, but there’s a serious problem here. In saying that a competent speaker can’t condemn morality on pain of conceptual confusion, we’re ruling out an important question. Thrasymachus (or whoever) poses a serious challenge to morality when he asks if virtue, or justice, or rightness, or anything else we traditionally value should be valued. A speaker who rejects the idea that virtue is to be prized is raising a serious question; he’s not just making a linguistic error. The challenge has an impressive philosophical pedigree, yet the view we’re considering answers it with an appeal to the conventions of language. Or, more accurately, it writes off the challenge as incoherent.

To strengthen this objection, we’d need to argue for a particular reading of just what challenge is brought by, say, Thrasymachus, Nietzsche, or other radical critics of morality. Our objection hinges on their challenge targeting, say, rightness or morality itself, instead of a particular conception of what morality demands. If their critique is best construed this way, as questioning why we should endorse right acts or virtuous

---

118 This criticism, made against expressivism generally, is discussed in Sturgeon (1986).
people, then the expressivist view we’ve been examining gives an unsatisfying response. But we might read the challenge of Thrasy...
must, that the critic endorses morality at some level, or that the target of his criticisms is a perceived misunderstanding of moral discourse rather than the institution itself. Granted, making this criticism requires identifying moral judgment not by appeal to distinct endorsing attitudes but in some other way, because we need a critic who makes moral judgments without endorsement. The range of cases (Thrasymachus, Nietzsche, Williams...), and the plausibility of interpreting some of these critics as attacking morality itself, should suggest that we’re not simply begging the question against the expressivist. If so, then the expressivist has left us unable to ask, or answer, an important question. This is a cost of limiting the range of conative variation in moral judgment.\footnote{Another way of putting this point is via the machinery of the open question argument. For any proposed noncognitive attitude or range of attitudes, a speaker might conceivably judge that something is morally right without having (or being the correct relationship with) that attitude.}

Let’s return to the larger question that began this section. We started by wondering about the specification challenge—we’ve seen, in our discussions of Gibbard and elsewhere, that there are costs associated with meeting the challenge head-on. Doing so means taking on the challenge of showing how one set of mental states can account for the range and nature of moral disagreements, the acquisition of moral concepts, and so on. The attending headaches suggest that the aspiring noncognitivist might want to avoid the challenge by showing that he needn’t specify particular mental states. In order to see if this move could be legitimated, we examined the related problem of conative variation. How can moral judgment be explained in terms of noncognitive states if those judgments can be associated with all sorts of different conations? Some answers to the conative variation problem simply add force to the specification challenge. For example, if the expressivist responds by simply denying that, say, Thrasymachus is making a moral
judgment, in virtue of his nonstandard attitudes, we’re in a position to press our demand for a specification of just what attitude is the appropriate one.

On the other hand, the expressivist might attempt to respond to the conative variation problem by showing how such attitudinal diversity is consistent with a univocal set of judgments in virtue of nonstandard speakers’ ties to the conversational mainstream. If moral judgments are all related in some, possibly indirect, way to a canonical sentiment or attitude, then we’re still justified in asking just what this state is. Now, the expressivist’s move to a more sophisticated account of the relationship between moral judgment and conation does help to explain some of the phenomena we’ve been discussing, but it comes with two costs. First, the specification challenge is still a legitimate demand, though the requirements on meeting it are changed by the more permissive view of how the specified attitude must connect with moral judgment. Second, that permissive relationship must be given an adequate explanation; that is, the specified connection must hold in cases of deviant moralists and the like.

Finally, notice that, of the possible expressivist rejoinders we’ve discussed, only Blackburn’s initial proposal allows him to dismiss the specification challenge. Yet this response turns out to have so many drawbacks that it hardly represents a move forward from Gibbard’s straightforward approach to the specification issue.

2.11 Some conclusions

The expressivist tradition has been motivated, in part, by the need to make sense of the univocity of moral conversation, as we’ve seen from our discussion of Stevenson and Hare in particular. Yet expressivism has not yet provided a satisfying way of preserving and explaining that disagreement, as it must if the semantic argument is to be
successful. If, as I’ve suggested, the expressivist must meet the specification challenge, it’s hard to see what mental states are up to the task. Gibbard’s proposal, which is informed both by reasonable psychology and by knowledge of his predecessors’ failures, comes closest; but it too fails, for reasons that should make us suspicious of the expressivist project generally. It’s simply not plausible to maintain, in the wake of these failures, that noncognitive mental states can provide the basis for an account of moral discourse and disagreement.
3.1 Introduction

As we’ve seen, the expressivist’s attempts to make sense of disagreement are less than completely satisfying. In this chapter, I’ll turn to the negative part of the semantic argument, which claims that other views of moral language are unable to preserve shared meaning across moral disputes. Traditionally, this argument was offered against ‘descriptivist’ views that attempted to account for the meaning of moral language in non-normative terms. Recently, the argument has been revived against more sophisticated versions of descriptivism. Here, I’ll examine the new, improved argument, compare it to its ancestors, and argue that it fails.

This new argument is best understood as an elaboration of R.M. Hare’s old line against descriptivism: if descriptivism were right, then speakers with different moral views would be talking past one another. But their disagreement is real, not merely verbal. The shared meaning this requires can be preserved, Hare thought, only by a view
that takes the evaluative meaning of “good” that’s primary.\textsuperscript{123} Recall Hare’s parable of the Christian missionary and his conversations with his cannibal hosts (discussed in 2.7). Despite their differences of opinion about what makes for a good man, they can argue, sensibly and coherently, about judgments of goodness, and this allows the missionary “to teach the cannibals Christian morals” (Hare 1952: 149). Regardless of our views of the missionary’s pedagogical efforts, we should recognize that this is a serious problem for any meta-ethical view that urges us to understand moral language as ascribing unproblematically naturalistic properties to acts, persons, and other entities. A view of this sort seems to threaten the univocity, and hence the legitimacy, of apparently genuine dispute.

Let’s call such a view naturalistic moral realism, or NMR. According to this view, moral terms ascribe and refer to moral properties, which are either identical with or supervene on metaphysically unproblematic lower-level properties. Hare’s challenge, simply put, is that naturalistic realism must show that its account of reference and meaning is able to preserve and legitimize intuitively compelling cases of evaluative dispute. The danger is that NMR will be forced to say that certain disputes that seem to be genuine, such as the one between the cannibal and the Christian, are really based on linguistic confusion. This is a counterintuitive result, and it leaves moral language unable to fulfill one of its chief purposes. If we can’t enter into discussion with interlocutors who disagree with us, then we can’t use our moral discourse as a tool for reaching consensus or for discussing our opposing answers to crucial questions about how to live.

\textsuperscript{123} This kind of argument comes up in several places in Hare’s work. See (1952) and (1989a). Similar considerations seem to be behind Allan Gibbard’s noncognitivism, especially (1990) ch. 1.
At first glance, Hare’s argument looks as though it’s easily countered by some familiar moves on the realist’s part. Haven’t we learned from Kripke, Putnam, Burge, and others that meaning isn’t a matter of ‘what’s in the head’ and that we might share terms while having significantly different beliefs about how to use them? What’s more, we’ve learned from reflecting on the failure of G.E. Moore’s Open Question Argument that an account of moral properties is not faulty simply because it’s susceptible to doubt by competent speakers. In light of these developments, we might be tempted to view Hare’s challenge as nothing more than a quaint relic rendered obsolete by the current state of the art.

Such a quick dismissal would be a mistake. In a recent series of articles, Terence Horgan and Mark Timmons have clothed Hare’s argument in contemporary garb. By way of an example they call ‘Moral Twin Earth,’ they argue that the realist’s use of developments in the philosophy of language is really of no help whatsoever in solving the problem. Even with the lessons of Kripke et al. in hand, NMR still faces the same basic difficulty, according to Horgan and Timmons: it cannot preserve shared meanings in cases where it appears, intuitively, that we have real moral dispute. This, in turn, shows that NMR relies on a faulty account of the reference of moral terms. For any realistic specification of the properties to which moral terms refer, we can imagine a community referring to different properties while still using our!moral terms. While the argument is often developed in response to a particular version of NMR, an amalgam of the views of Richard Boyd and David Brink, it purports to have force against any view of this sort.

124 Of course, their conclusions aren’t uncontroversial. The ethical naturalist, though, tends to like externalist views of content and reference, for fairly obvious reasons. See Boyd for discussion.
Fortunately, the realist’s prospects aren’t as bleak as an initial visit to Moral Twin Earth would have us believe. Though I’m not satisfied with other attempts to deal with the argument, I think that NMR can be defended. Here, I’ll offer three arguments against the Horgan-Timmons challenge. In addition, I’ll offer a diagnosis of the underlying problem that makes the challenge so powerful and then make some suggestions concerning what the realist must do in order to resolve the underlying problem.

3.2 The Moral Twin Earth Argument

Let’s begin with a close examination of the challenge, beginning with its relationship to Hare’s earlier argument. We can think of Hare’s objection this way:

1. Suppose that moral terms could be defined by appeal to their descriptive meaning.
2. Descriptive meaning is determined by a community’s use of a term, and so a community’s practices determine descriptive meaning (see Hare 1952, 1989a).
3. So radical differences in practice and belief concerning a moral term (e.g., between Christian and cannibal) would yield different meanings.
4. But people with radically different views about how to use “good” nonetheless share a single term (that is, their disputes about good are univocal).
5. Thus, moral terms cannot be defined descriptively.

The weak point here seems to be the second premise. As I noted in passing above, a defender of NMR can appeal to, say, a causal theory of reference in order to show that people with very different beliefs about something can nevertheless be talking about the same thing. To use an old example, perhaps it’s because of our causal ties to electrons that we manage to preserve meaning despite changes in theory. Just as there’s been disagreement between old and new science about electrons, there might be two

---

125 For other responses, see Sayre-McCord (1997) and Copp (2000).
communities with the relevant kind of tie to a single property, such as rightness, such that
their term “right” referred to that property, in spite of their different beliefs about what
something had to be like in order to be right. What’s more, this ethical extension of the
so-called New Theory of Reference shouldn’t be surprising, since one of the motivations
for that view is the need to explain how shared meanings are compatible with different
beliefs, associated descriptions, and so on. While the view might have taken its
inspiration from cases of theory-change in the sciences, it’s a fairly natural move to
extend it to the analogous problem in moral language, a move that seems at least initially
plausible. So Hare’s argument seems to be hobbled by its reliance on a deservedly
unpopular view of reference.

Now for Moral Twin Earth. Horgan and Timmons’ insight is that Hare’s crude
account of the reference relation (that is, the relation between the term and its referent in
virtue of which the term refers to that thing) can be replaced with any other, more
sophisticated account while leaving the basic argument intact. In Hare’s presentation, the
descriptivist’s problem looked as though it rested on the peculiar view of meaning
invoked in premise (2). But the problem can be generated without relying on this view,
and so the realist can’t escape simply by rejecting that premise of the argument. Horgan
and Timmons show that the problem is much deeper than Hare’s presentation would have
us believe.

Here’s why. The realist, as we’re imagining him, thinks that moral terms admit of
synthetic naturalistic definitions. That is, moral properties can be defined in terms of
other, more basic natural properties—perhaps, as Brink suggests, these are multiply-
realizable functional properties supervening on base-level physical properties (see Brink
Just as heat can be defined in terms of molecular kinetic energy, or water in terms of H2O, the moral properties to which moral terms refer can be given naturalistic definitions as well. If so, then we’ll need a story about how moral terms refer to the realist’s natural property of choice. That story will include a specification of some relation R such that moral term t refers to natural property (or property-cluster) N in virtue of standing in R to N. But, given this specification, Horgan and Timmons say, we can imagine a scenario in which we encounter a speaker or a speech community using our moral terms (that is, terms that mean the same thing as our “right”, “permissible,” and so on) even though, for this speaker or set of speakers, term t stands in R to some other natural property N’. That is, we can imagine a case where we have moral disagreements with someone whose use of “right” is, on the realist’s account of reference, referring to some property different from the property we’re referring to. Yet our semantic intuitions tell us that we’re talking about the same thing; we’re not involved in a dispute that’s ‘merely verbal.’ Since we can hold meaning constant while altering the natural properties standing in relation R to our moral terms, the meaning of our terms cannot be accounted for by appeal to those natural properties. This argumentative strategy, Horgan and Timmons say, can be invoked for any specification of the reference relation R. Thus it can be used against any specification of the details of NMR—its force doesn’t depend on any particular account of the natural properties involved or on a specific account of reference. So, they conclude, any specification of R will leave the realist vulnerable to a potent question: couldn’t we have meaningful evaluative disagreement with people whose terms were so related to some other properties? The answer, they claim, is yes. If so, then the realist has some serious difficulties.
All of this is too abstract to be intuitively compelling, so let’s consider how this strategy applies to a particular brand of NMR. Horgan and Timmons create a kind of Frankenrealist out of the views of David Brink and Richard Boyd and level their objection against this version of the position. Because the details of their example are important to my argument, I quote their description of Moral Twin Earth at length. (Note that, according to the ‘Brink-Boyd’ view, our moral terms are causally regulated by functional properties that are correctly characterizable by a single normative theory, and it’s this causal regulation that’s responsible for our terms referring to these properties. Also, we’re assuming for the sake of argument that the properties regulating our moral discourse are those described by the consequentialist theory Tc.)

Now for Moral Twin Earth. Its inhabitants have a vocabulary that works very much like human moral vocabulary: they use the terms ‘good’ and ‘bad,’ ‘right’ and ‘wrong,’ to evaluate actions, persons, institutions, and so forth...But on Moral Twin Earth, people’s uses of twin-moral terms are causally regulated by certain natural properties distinct from those that (as we are already supposing) regulate English moral discourse. The properties tracked by twin English moral terms are also functional properties, whose essence is characterizable by means of a normative moral theory. But these are non-consequentialist moral properties, whose functional essence is captured by some specific deontological theory; call this theory Td. These functional properties are similar enough to those characterizable via Tc to account for the fact that twin-moral discourse operates in Twin Earth society and culture in much the manner that moral discourse operates on Earth....In addition, suppose that if Twin Earthlings were to employ in a proper and thorough manner the same reliable method of moral inquiry which (as we are already supposing) would lead Earthlings to discover that Earthling uses of moral terms are causally regulated by functional properties whose essence is captured by the consequentialist normative theory Tc, then this method would lead the Twin Earthlings to discover that their own uses of twin-moral terms are causally regulated by functional properties whose essence is captured by the deontological theory Td. (Horgan and Timmons 1992b: 245-6)

I should note a bit of unease about the compatibility between the views of Brink and Boyd, but this doesn’t affect my argument, and I’m not convinced it affects Horgan and Timmons in a significant way.
Just as you might expect, Moral Twin Earth looks a lot like Earth to the naked eye. The crucial difference is that use of moral (or twin-moral) terms is causally regulated by the properties specified by Td, while our own moral terms are regulated by the properties described by Tc.

Suppose that Earth moralists encounter Twin Moralists, and engage in what looks to be an evaluative dispute—say, over a claim like “It’s right to execute murderers.” How should we interpret what’s going on? There are two options. First, we could take twin-moralists to share our terms, and thus understand our conversation as a genuine disagreement. In this case, we’d have a real dispute about what’s right. Second, we could interpret them as using different terms, and conclude that our conversation is based on linguistic confusion. In Putnam’s original Twin Earth case, of course, we tend to think that a debate over whether “water” really refers to H2O or XYZ is illegitimate, because “water” means different things on Earth and Twin Earth. So too with “right,” according to this second line of interpretation. According to Horgan and Timmons, the choice is clear: “[t]here is no hermeneutical pressure to revise the original interpretation of Moral Twin Earthlings as having a moral vocabulary that they use to express moral beliefs” (Horgan and Timmons 1992b: 247). We initially think that we have real disagreement with the twin-moralists, and learning that twin moral terms are causally regulated by different properties isn’t enough to shake our judgment that there’s one set of terms here, not two. If this is correct, then causal regulation cannot be the reference-
fixing relation that connects moral terms with particular natural properties. And similar MTE-style cases can be devised in response to other attempts at specifying the relation.\textsuperscript{127}

This argument is as elegant as Hare’s in its construction and more potent for being generalizable. It’s not limited by a reliance on one particular account of reference—for \textit{any} proposed account of how moral terms refer (or any account of which properties they refer to), we can construct a scenario that looks just like good old Earth, except for the fact that in our imagined scenario our terms are hooked up by the candidate relation to \textit{other} natural properties. We think that this difference doesn’t prevent us from sharing moral terms with our imagined interlocutors, since we think of our disagreements with them are about morally substantive matters. Hence, \textit{pace} the realist, it can’t be those proposed natural properties that are essential to our moral terms.

Or so it seems. Later in this chapter, I’ll argue that much of the intuitive force of this argument rests on the underdescription of Moral Twin Earth and its twin-moral practice. Once we understand the constraints on its workings, we’ll be less willing to grant the interpretive point that’s central to the argument. Before turning to these arguments, though, I’ll look at some replies that \textit{don’t} succeed.

\textsuperscript{127} I’m taking the central issue to be the preservation of disagreement, which may seem odd, given that the focus of Horgan and Timmons (1992b) concerns the explanation of moral-nonmoral supervenience relations. There’s no conflict, though, because the argument about supervenience goes roughly like this: NMR typically appeals to an ‘innocence by association’ strategy in order to show that moral supervenience is no more problematic than is the supervenience of the mental, economic, political, etc. on the physical. But the moral case is special, Horgan and Timmons continue, because, in all of these cases, the supervenience needs to be explained, and the explanation that works for the mental, the biological, and so on doesn’t work in the moral case. They conclude that the ‘innocence by association’ strategy is not enough. The Moral Twin Earth case is supposed to show that the semantic constraints that serve as part of the best explanation of other cases of supervenience won’t work to explain moral supervenience.
3.3 Copp’s first response

In the next two sections, I’ll consider David Copp’s responses to Horgan and Timmons. Though I think that Copp’s replies go wrong in important ways, I also think that they get a lot right—in particular, they highlight a variety of resources the realist might use to respond to the challenge posed by Moral Twin Earth. He offers two lines of reply; I’ll consider both in turn.

Copp’s first reply begins by noting that Horgan and Timmons assume that, in order for there to be genuine disagreement between moralists and twin-moralists, “the Earthling term ‘wrong’ must express the same property as is expressed by the Twin Earthling term ‘wrong’” (Copp 2000: 124). He then challenges the assumption by suggesting that moral disagreement is possible even if “right” on Earth and on Moral Twin Earth express different properties. That is, he argues that disagreement is possible, in some important sense, even if “right” on Earth ascribes *rightness* while “right” on Moral Twin Earth ascribes a different property, *twin-rightness* (Copp 2000: 123-126).

How could univocal conversation be possible if the terms express different properties? Copp suggests we think of how we might *translate* twin-moral terms into English. Because of the extensional near-equivalence of “right” in both communities, and because of the similar evaluative role played by twin-moral and moral judgments, an English translator might choose to render the twin-English term as the English word “right.” This seems plausible enough: think of cases where English has no exact match for a term in another language, and we’re forced to settle for an inexact translation. We just do the best we can, keeping in mind that there are nuances lost in translation. Thus,
even though corresponding moral and twin-moral terms express different properties and therefore have different “meanings” in the philosophically preferred sense of the term, there is also a sense in which they might have the same “meaning.” Given this, we can make sense of our intuition that Earthlings and Twin Earthlings disagree morally even if corresponding moral and twin-moral terms express distinct properties. (Copp 2000: 125)

So, Copp thinks, there is a sense in which we disagree with the twin-moralists—it just isn’t the narrow philosophical sense of meaning that remains constant in the dispute. The terms have the same meaning in the broader, more ordinary, sense of being intertranslatable (Copp 2000: 128). We have real disagreement in one sense of meaning and not the other. He concludes that

our intuitions in the Moral Twin Earth scenario do not undermine the apparent implication of Putnam’s theory that the English and Twin English term have different meanings in the preferred philosophical sense. In an ordinary sense of the term, terms have the same ‘meaning’ if they are intertranslateable. (Copp 2000: 128)

According to this line of argument, Horgan and Timmons’ mistake lies in thinking that our intuitions of shared meaning—intuitions that are central to their objection—involve the narrow ‘philosophical’ sense of meaning, while they actually involve a broader, rough-and-ready notion. Since terms can have the same meaning, in this second sense, while expressing different properties, the Moral Twin Earth argument fails, since it relies on the idea that shared meaning requires shared properties.

This reply is unsatisfying, because it comes too close to explaining why we think there’s disagreement without vindicating that disagreement. This is a serious concession, and the realist should turn to it only as a last resort. To see why this line of reply does not adequately preserve disagreement, think of cases where translations give us identity of meaning in the broad, loose sense, but not in the narrower, more technical philosophical
sense. Disputants arguing about the application of one of these terms, it seems, are a bit confused: if they were informed of the different properties being expressed by their ostensibly, but not actually, univocal term(s), they would stop arguing. The dispute is terminological, not substantive, because there is something lost in translation. The argument shouldn’t persist in light of better information about the workings of the terms at issue. In cases like this, the right thing to do is to accept that translations are not exact, and stop arguing. This is just to say that the argument between moralists and twin-moralists rests on linguistic confusion.

Of course, it is natural that, at first, we’d see their terms as meaning what ours do, because of their orthographic identity and functional similarity, but this presumption would disappear on learning that the two terms express different properties. We will find this satisfying only to the extent to which we are willing to see our initial intuitions about the case as being mistaken. If, as I suspect, our intuitions are firmer—that is, we think there really is disagreement here—Copp’s first line of reply is insufficient.

3.4 Copp’s second response

Copp realizes that his first reply will leave many philosophers unsatisfied, because, in effect, it grants Horgan and Timmons’ central point: in the narrow, philosophical sense of ‘meaning,’ “right” doesn’t mean different things on Earth and on Moral Twin Earth. His second reply is designed to show that Horgan and Timmons cannot secure even this point. The argument relies on a close examination of Putnam-style semantics, and, in particular, how those semantics leave room for error. When we take Putnamian approaches to error and false presupposition seriously, we’ll have a powerful response to the Moral Twin Earth argument, because we’ll be in a position to
say that, if “wrong” expresses any property, then it expresses the same property on both Earth and Moral Twin Earth. While this isn’t supposed to show that there is such a property, it does purport to show that, in any coherent example of the type Horgan and Timmons give, “right” will have one meaning across both contexts (Copp 2000: 130).

The second argument goes roughly like this. According to Putnamian semantics, the reference of our terms is determined by “the actual nature of the stuff in the samples we use in explaining the meaning of the term together with the actual content of our referential intentions” (Copp 2000: 133). Since Earth and Moral Twin Earth are (almost) identical, and the content of our intentions is the same, “right” must refer to the same acts on both planets. To see how this argument unfolds, though, we should look at the ways in which Copp thinks Horgan and Timmons go wrong.

Consider again the differences between Earth and Moral Twin Earth. On Earth, moral evaluators use “wrong” in a way that, for the most part, conforms to Tc, the consequentialist moral theory; on Moral Twin Earth, however, “wrong” is used in a way that follows the dictates of deontological theory Td. So, different actions are called “wrong” on Earth and on Moral Twin Earth. Speakers on both planets, though, presumably intend to refer to acts that are of the same moral kind as “the local samples of action that are called ‘wrong’ in [the speaker’s] language” (Copp 2000: 134). Since the local samples have a different nature, then “right” will express different properties in English and Twin English. This is just the result Horgan and Timmons need: if the terms express different properties, we have no genuine disagreement with the speakers of Twin English.
Copp identifies two problems with this line of reasoning. First, it does not take Putnamian theories of error into account: it doesn’t allow for widespread error in use of terms such as “wrong.” Putnam says, for example, that my ostensive definition of “water” presupposes that the sample I’m pointing to is in the same-L relation to the stuff that most speakers in my community call “water.” Ostensive definitions of moral terms, though, do not work this way. It’s perfectly sensible to think of me teaching the use of moral terms even though I’m a moral nonconformist, with views different from that of the community around me (see Copp 2000: 134). Second, the proposed Horgan-Timmons argument fails to take into account that not all similarities shared by members of a term’s extension are relevant to the term’s reference. For example, all water may just happen to be liquid at a given time, but being liquid isn’t essential to being water; the relevant similarity is being H2O. Or, to use Copp’s example, imagine that humans were the only kind of mammal ever to live on Earth. Then, all milk would have had one basic chemical structure, but it’s at least plausible to think that what’s essential to being milk is a functional or genetic property such as ‘liquid produced by a female mammal used to nourish its young’ (Copp 2000: 134). Likewise, all or most of the acts called “wrong” on Earth share the property of being prohibited by Tc (or they share the property of producing suboptimal consequences, which amounts to the same thing), and all or most of the acts called “wrong” on Moral Twin Earth share the property of being prohibited by Td (or the property of being in violation of certain conduct-governing rules). But this does not show, in itself, that these properties (being contrary to Tc and Td, respectively) are what’s essential to being “wrong” on Earth or Moral Twin Earth. Horgan and Timmons have failed, that is, to identify the relevant kind in question.
How, then, can a Putnam-style semantics preserve univocity between Earth and Twin Earth? Copp thinks that the presuppositions of moral discourse are crucial to this stage of the argument. I noted above that a speaker need not think that her applications of terms such as “wrong” have to be directed at acts that are of the same kind as those acts her fellow-speakers label “wrong.” She would not withdraw the term, that is, if she discovered that she applied the term to acts that are of a different kind from the acts that are called wrong by other speakers. She would, of course, if she’s being sincere, withdraw the term if she came to believe that the action in question is not!wrong, or lacks the property of wrongness.128

This, Copp thinks, shows, first, that the use of moral terms presupposes that there are moral properties or kinds and, second, that these presuppositions show us something

---

128 Copp says that the relevant kind is the kind “wrong action,” and the property we intend to ascribe is the property of “wrongness” or, as J.L. Mackie might say, “not-to-be-doneness” (Copp 2000: 22). Copp’s language here and elsewhere, as well as his approving citation of Mackie’s views about what speakers intend to do in calling actions wrong, led me, on first reading, to think that the property of “not-to-be-doneness” had some kind of Mackie-style ‘objective prescriptivity.’ That is, it gave reasons (or motivations) for acting in certain ways to all agents, regardless of their desires and ends (here I’ll avoid distinguishing between the host of positions bound up in this idea, because the distinctions don’t matter for present purposes).

The problem with using this feature of ‘objective prescriptivity’ or “to-be-doneness” as a way of ensuring univocity is that it saddles the naturalistic realist with problems he would otherwise like to avoid. Objective prescriptivity plays a central role in Mackie’s ‘argument from queerness’ precisely because it’s hard to see what in the world might have this feature. If we think that moral properties have to have this power to move us, and if we think, sensibly enough, that no natural property can have this power, we’re forced to say that moral properties cannot be (identical with or constituted by) natural properties. One popular solution for the naturalist has been to deny internalism: that is, many naturalistic realists simply reject the idea that moral properties need to have this feature. My objection to this reading of Copp’s reply is not that it’s incompatible with a line that many naturalistic realists have, in fact, taken. The problem, instead, is that externalism and naturalistic realism are closely-connected views, and much of realism’s plausibility depends on externalism.

128 Though certainly this is contentious, and begs the question against some of the realist’s most potent opponents. One can imagine a philosopher like Simon Blackburn, for example, using moral terminology without these presuppositions. Nonetheless, Copp claims that “[i]f we came to believe that these suppositions [about there being moral properties] to be false, we would not engage in moral discourse and moral teaching” (Copp 2000: 135).
about our referential intentions in using moral terms. He then goes on to make an illegitimate move in identifying English and Twin English referential intentions:

In Putnam’s view, this point about the presuppositions of moral discourse implies that we have corresponding referential intentions. For example, it implies that, in using “wrong” (in a non-deviant case), an English speaker intends to refer to actions that are of the kind, or that have the property, that he and most speakers in his linguistic community intend to refer to in using “wrong.” Or, more simply, in using “wrong,” an English speaker intends to refer to actions that are of the kind, wrong, or that have the property of being wrong or of being not-to-be-done. And speakers of Twin English would have to have the same intentions. If they did not, the relevant terms in their language would not be moral terms. (Copp 2000: 135)

We can see how this move, if it were right, would save the naturalistic realist. Recall that, according to Putnam-style semantics, reference is determined by the actual content of our referential intentions in conjunction with the actual nature of the world. Moral Twin Earth is, ex hypothesi, just like Earth in all morally relevant respects, and the speakers of Twin English have referential intentions with the same content as the relevant English speakers’ intentions. Thus, “right” on Earth and on Moral Twin Earth must have the same referent, because the factors that determine reference are exactly the same in both cases.129

There’s something suspicious about Copp’s argument. His second reply to Horgan and Timmons rests on two claims: (i) that terms such as “right” and “wrong” on Moral Twin Earth are moral terms, and (ii) that, for a term to be a moral term, as opposed to, say, a term of etiquette, it must be used with certain referential intentions—that is, it

---

129 This is essentially the argument Copp gives. The worlds are the same by stipulation, the referential intentions are the same because these are required for the terms in question to be moral terms at all (or so Copp thinks), and English and Twin English speakers have the same interests in using the terms (that is, both terms play the same practical role in our practice, and so on). I suspect that this last point of similarity is redundant, given the identical content of our referential intentions, but whether this is so or not does not affect my argument.
must be used to refer to properties such as “wrongness” or “not-to-be-doneness”. This conjunction of claims ensures that Horgan and Timmons’ example is clearly incoherent.

If being a moral term requires that the referential intentions behind its use are to refer to acts that have, for example, the property wrongness, then simply to be a moral term is to have the same meaning as the English term. As Copp understands their argument, Horgan and Timmons stipulate that “wrong” on Moral Twin Earth is a moral term, which, given the similarity between Moral Twin Earth and Earth, requires that it refer to the property wrongness in virtue of the referential intentions that make it a moral term, and this, in turn, guarantees that it means just what the English term “wrong” means, and hence that our disagreements with twin moralists are genuine, even though Horgan and Timmons intended to show just the opposite. This, it seems, is too close to simply insisting that moral and twin-moral terms mean the same thing. Now, it may be that Moral Twin Earth is incoherent. In fact, later in this discussion I’ll suggest that there’s a sense in which it is incoherent. But it certainly seems to make sense, and, what’s more, it seems to be a powerful challenge to synthetic realism. All of this should lead us to expect that Horgan and Timmons have not, in fact, been generous enough to give the realist all he needs in order to show that the argument self-destructs. Surely there is a more charitable way to understand the Moral Twin Earth argument.

The problem, I think, lies with Copp’s claims about what’s required of moral terms and his claim that “right” and “wrong” on Moral Twin Earth must be moral terms. There are two ways of formulating my objection: first, we might say either that the Twin English term “wrong” is a moral term, but that being a moral term does not require what Copp thinks it does. On the other hand, we might say that Copp is right about what’s
required of moral terms and then deny that the Twin English orthographic equivalents of our moral terms are *themselves* moral terms. Saying either of these things is compatible with the spirit of the Moral Twin Earth argument. I’ll present both of these options quickly.

First, take Copp’s claim that, in order for “wrong” to be a moral term, it must be used with the intention to refer to actions that have the property of wrongness. This seems to be a fairly, and perhaps needlessly, strict requirement on what counts as a moral term. Expressivists might continue to use the term despite having no intention to refer to any property whatsoever; surely it’s still a moral term in *their* use. Suppose we come across Moral Twin Earth and observe that the term “wrong” plays a certain functional role in their practice, attaches to acts meeting certain descriptive criteria, and so on. Not knowing anything about the referential intentions of Twin English speakers, we still know enough to say that this term is a moral one, at least in an ordinary sense: we think it means the same as ours, we think that it can serve as the focal point of real disagreement, and so on. The intuition of disagreement does not rest on any claims about referential intentions. Should we learn that twin moralists lack the right kinds of referential intentions, we would still think we have genuine disagreement when we apply “wrong” to different actions.

The second way of putting this reply comes to the same thing. Suppose that Copp is right about moral terms. Horgan and Timmons can build their case just as effectively without stipulating that Twin English terms “right,” “wrong,” and so on, are moral terms. In fact, in other statements of their case, they avoid—consciously, it seems to me—the
troubling stipulation that lies at the heart of Copp’s argument. Consider their specification of the differences between our two planets in another version of the example:

Now for Moral Twin Earth. Its inhabitants have a vocabulary that works very much like human moral vocabulary: they use the terms ‘good’ and ‘bad’, ‘right’ and ‘wrong.’ to evaluate actions, persons, institutions, and so forth...But on Moral Twin Earth, people’s uses of twin-moral terms are causally regulated by certain natural properties distinct from those that (as we are already supposing) regulate English moral discourse. (Hargan and Timmons 1992: 245)

The central point of this case, it seems to me, is that we think we have real disagreement once we look at the surface features of twin-moral practice: the formal role played by “right” and “wrong,” the rough agreement in extension between English and Twin English speakers as to what acts deserve what label, and so on. There is no need to stipulate anything about Twin English’s terms being moral terms or not, in Copp’s restricted sense, or, for that matter, in any other sense. The example will do its work without this added complication. Hence, Copp’s second argument fails.

3.5 Disagreement, here and elsewhere

Let’s begin evaluating the Moral Twin Earth argument by looking at what it must do if it’s to succeed. The argument is meant to show that NMR’s view of moral terms leaves it unable to account for real dispute, because there are cases in which we think we have univocal disagreement with another community in spite of the fact that their terms and ours are connected to different natural properties. Thus, the Horgan-Timmons

---

Grant, in the long description of Moral Twin Earth quoted earlier in this paper, Horgan and Timmons do call Twin English’s terms “moral” in passing. But this isn’t enough, it seems to me, to convict them of anything more that sloppiness, and, what’s more, it doesn’t matter much what the real Horgan and Timmons wrote—we can simply imagine a Twin Horgan collaborating with Twin Timmons to produce a closely related but not identical example...
objection rests on two conditions. First, we have to think that we have a real moral
dispute (that is, about what’s right) with the twin-moralists. We have to think that our
terms mean what theirs do, or the realist’s (purported) admission of equivocation seems
intuitively plausible. Second, it has to be true that twin-moral terms really are hooked up
in the right way (e.g., by causal regulation) to natural properties that don’t stand in this
relation to our moral terms. If not, then the realist can happily agree with our linguistic
intuitions, and there’s no objection. In this section, I’ll use these two conditions to pose a
dilemma for Horgan and Timmons. The basic idea is that their thought-experiment has
trouble satisfying both of these requirements simultaneously: if it meets one, it fails the
other.

Before examining this charge, let’s note an initial difficulty with the example that
drives the Moral Twin Earth argument. The problem is that we don’t know very much
about what twin-moral practice looks like. We know that twin-moral terms play roughly
the same formal role as our moral terms: they evaluate actions, their application is
connected to patterns of praise and blame, and so on. But we don’t know anything about
the content of twin-moral discourse. What kinds of acts are labelled “wrong” by our
alien evaluators, and why? What kinds of justifications are offered in defense of these
assessments? What broad theoretical commitments inform their reasoning? These
details matter because they affect both what we say about the reference of twin-moral
terms and our intuitions about twin-moral meaning.

With this worry about underspecification in place, let’s turn to the dilemma.
Suppose Horgan and Timmons attempt to meet the first requirement by making sure that
their example prompts strong intuitions about the univocity of moral and twin-moral
terms. One way to do this is to make Moral Twin Earth exactly like Earth in every respect, so that twin moralists make the same kinds of judgments with “right” that we do, appeal to the same kinds of justifications in defending these judgments, and so on. But if this were the case, then the twin-moral term “right” couldn’t refer to a different (functional) natural property given that the same relation fixes reference in both places. A term will refer to a property in virtue of some fact about the world, the term, and its use; if Moral Twin Earth is just the same as our own planet, then their terms cannot but refer to the same properties.\footnote{\label{fn:1}Here’s another way of supporting the claim that a difference in reference must be accompanied by some difference in practice. According to Boyd’s view of reference, “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 $k$ (excuse the blurring of the use-mention distinction).” Such mechanisms include investigative procedures, deference to experts, and so on. When these relations obtain, “we may think of what is said using $t$ as providing us with socially coordinated epistemic access to $k$” (Boyd 1988: 116). Those causal mechanisms will include features of our practice, for example, the existence of a body of fairly stable judgments about particular cases, a methodology for handling disputes, and so on. All of these features of moral (or twin-moral) practice have an impact on the reference of our moral terms on a view like Boyd’s. If twin-moral practice were identical to our own in every respect, we’d end up, at the end of the day, saying exactly the same things, given that other factors will be held constant. Hence there must be some difference in the initial positions that accounts for the difference in end-of-the-day theory.} The realist can argue, plausibly enough, that Horgan and Timmons’ stipulation about reference is at odds with the facts about the twin-moral practice that serve to fix reference.

On the other hand, suppose Horgan and Timmons want to force the realist to admit that, by his lights, twin-moral terms do refer to different properties, thus ensuring that their example meets the second requirement I’ve set out. The way to do this is by making twin-moral practice different from our own moralizing. It’s in virtue of these differences that “right” can refer to different properties in different places. Fleshing out the example in this way, though, weakens the univocity intuition. We think that we share the term “right” with Moral Twin Earth because we imagine that twin-moralists use the
term in roughly the same ways. Suppose we learn that this is false, because there are
significant differences between our uses of the term and theirs (differences, say, not just
in the extensions of the terms but the kinds of reasons they give and accept, and so on).
Twin moralists apply “right” to acts we’re sure are abhorrent, they offer completely
different kinds of justifications for their claims, they see our reasons as irrelevant, and so
on. It seems to me that then we’d either withdraw our initial judgment that we share
terms with the twin-moralists, or at least offer it with less conviction. The content of the
twin-moral practice will affect our judgments of shared meaning.

The basic strategy is simple: first, we press Horgan and Timmons for a more
detailed account of the workings of Moral Twin Earth, and then we argue that either (a)
the differences in content between moral and twin-moral practice are significant enough
to undermine our conviction in the synonymy of moral and twin-moral terms or (b) the
similarities are sufficient to ensure that, according to the realist’s view of reference, twin-
moral terms refer to the same properties as our own moral terms. It’s easy to miss this
worry about the details of twin-moral practice if we’re thinking of Putnam’s original
Twin Earth example, where the only difference lies in the chemical composition of the
watery stuff in lakes and rivers. There, the human (or twin-human) side of things can
remain the same. Not so on Moral Twin Earth, where perfect agreement with Earthian
moral practice will fail to secure the difference in reference that the Horgan-Timmons
argument requires.

As it stands, this is only the start of a compelling argument in defense of NMR.
The dilemma is a problem for Horgan and Timmons only if it’s impossible for them to
slip between its horns. It might seem that they can easily escape unscathed—all they
need is a story about twin-moral discourse that’s able to meet the two conditions specified above. So far we’ve seen nothing to make us doubt that there’s a specification of Moral Twin Earth that garners the right intuitions while forcing the realist to acknowledge the needed difference in reference. Indeed, we might think that Horgan and Timmons have reason for optimism when we notice that our intuitions about shared meaning are fairly permissive. Actual-world moralists with strange views don’t often tempt us to say that they’ve merely changed the subject. When Peter Singer argues with Kantians, we don’t think they’re talking past one another. Just as differences between Earth speakers don’t strain our semantic judgments, small but important differences between moral and twin-moral discourse might leave us with the required intuition that speakers on the two planets are talking about the same thing. Yet these differences could be enough to ensure that, according to the realist’s account, “right” refers to different things in different places. If so, then our first argument comes to nothing.

To add force to our complaint, then, we need reasons for thinking that there’s no way to specify the details of Moral Twin Earth in a satisfying way. In order to provide these reasons, I’ll pursue what may seem like an odd dialectical strategy. First, I’ll exploit a concession that Horgan and Timmons make regarding disagreement here on Earth. Once we take this concession seriously, we’ll see that it has interesting implications for the details of their case. Of course, this will be of little interest if the concession is one that needn’t be made, so I’ll give some reasons for thinking that the point they grant for the sake of argument is one that’s plausible on independent grounds. The end result, I’ll argue, is that the Moral Twin Earth problem is intimately tied to another, more familiar worry about the resolution of moral dispute. Seeing the
connections between these two puzzles requires a look at a cluster of interesting problems that the realist must address if she’s to answer the general problems suggested by the criticisms of Hare and of Horgan and Timmons. While treating these problems with the care they deserve is beyond the scope of this paper, I hope to take some initial steps toward that goal.

To begin, though, let’s take a look at what happens if we assume, with Horgan and Timmons, that the terms of our moral discourse here on Earth are causally regulated by, and so refer to, one set of natural properties, despite variations in individual speakers’ moral views.\(^{132}\) We’re supposing, then, that even Earth-speakers with significantly different normative commitments are hooked up in the right sort of way to the same cluster of properties, even though they make different moral judgments, offer different sorts of justifications, and so on. This is granting, for the sake of argument, that the conversations between Kant and Mill, between Christian and cannibal, and between the moral communities of various places or times are all within the realist’s semantic reach. NMR’s theory of reference and meaning, whatever its details turn out to be, can preserve univocity between speakers with diverse views on topics such as the divine right of kings, the permissibility of chattel slavery, the obligation to segregate people of different races,

\(^{132}\) For example, they suppose that the following claim is true: There is indeed a unique family of functional properties that causally regulates the moral judgments and moral statements of human beings in general, despite the fact that humans widely disagree among themselves about matters of morality. (Horgan and Timmons 1992b: 245) Later in the paper I’ll argue that the concession is something we have reason to believe anyway. But why do Horgan and Timmons concede it? They present the Moral Twin Earth argument as a ‘knockout punch’ to NMR, regardless what happens with actual-world disagreement. This suggests that they think their argument is an additional problem for the realist. If I’m right, this is false. This conclusion is interesting for another reason: it focusses attention to the issue of disagreement and univocity here on Earth, which is where the realist wants it.
and the proper social role of women—or so our assumption lets us say. This is an impressive claim precisely because of the range of actual-world moral disagreement.

At the beginning of this section, I pointed out that the twin-moral practice must differ in some way from our own moralizing, or else we couldn’t make sense of the idea that twin-moral terms refer to different properties. Taking Horgan and Timmons’ concession seriously lets us strengthen our objection because it allows us to say that these differences between moral and twin-moral discourse must be significant. The realist can preserve meaning across actual-world disputes, we’re assuming, though, if Horgan and Timmons are right, she can’t preserve meaning between Earth and Moral Twin Earth. Thus the twin-moral practice must be different from our moralizing in a way that puts it beyond the realist’s reach. So there’s pressure on Horgan and Timmons to make the details of twin-moralizing much different from those of our own moral discourse; in an important sense, it must be like nothing we’ve ever seen before. If it gave us only the views of a past time-slice of our moral discussions, or views that we see as fringe participants in our own conversations, then the realist could reply, plausibly, that the twin-moralists offer no more of a challenge to her view of meaning than do the Earthian interlocutors we’re assuming don’t cause problems.

Now let’s examine the effects of these differences on our intuitions about the univocity of moral and twin-moral terms. Earlier, I suggested that the content of twin-moral judgments is relevant to our assessment of whether or not the twin-moralists mean what we do in calling acts “right.” This seems, at least to me, to be an intuitively appealing view. After all, we need some warrant for interpreting these speakers as referring to our old friend rightness, and the mere orthographic identity of our terms is
insufficient grounds for the univocity judgment. One might appeal to similarities
between the evaluative role of “right” on Earth and Moral Twin Earth, but this formal
resemblance doesn’t seem sufficient. Imagine our response on encountering some
radically foreign culture whose evaluative practices are quite different from our own. We
could understand their terms as meaning “right” and “wrong,” or we could think that they
have different concepts that play roughly the same role that the concepts of
right and wrong play in our discussions. We might think that they evaluate action in
terms of honor, or fierceness, or some untranslatable notion that (at least generally and
for the most part) carries with it the kind of endorsement that rightness-judgments have in
our own case. We can still apply our moral notions to their actions—that is, we can
judge that they’re doing wrong—and we might think they have reason to take up our way
of doing things, but all of this is compatible with thinking that their evaluative terms
don’t mean what ours do.

I’ve argued that Moral Twin Earth must be a case like this, since it presents us
with agents whose evaluative practices differ from our own in remarkable ways. If so,
and if I’m right about the effects of these differences in content on our intuitive
assessments, then it’s not at all clear that we should interpret twin-moralists as talking
about right and wrong. The differences between moral and twin-moral practice
undermine our univocity intuitions, and, by extension, the force of the Moral Twin Earth
argument. The realist’s view of the case starts to sound like just the right one.133

---

133 As always, there are some additional nuances here. Why can’t some small, subtle difference in twin-
moral practice ensure the difference in reference that Horgan and Timmons need? (After all, they need
only to establish that twin-moral terms refer to different properties, not that they refer to drastically
different properties). If the small difference in question is between our current Earthen moral practice and
some variant, we’d need to hear why this small change is enough to force the difference in reference, when,
The argument of this section so far shows that if the realist can account for all, or at least enough, of moral disagreement here at home, then Moral Twin Earth poses no additional threat. At the very least, it adds to the philosophical debt that must be paid if the argument against NMR is to work. Horgan and Timmons must explain just how the following three claims are consistent: (a) the realist can account for shared meaning we’re supposing, the realist can account for the wide variety of different positions represented in past and present moralizing here on Earth. The realist’s (assumed) ability to cope with the wide scope of real variation should make us confident that a small change from our current theorizing is not enough to raise problems. (That is, as long as the view of reference in question is tracking roughly the same phenomena that determine meaning—or else we’ll have a counterintuitive view of meaning, and a failure to respect the moral inquiry problem, which is discussed below. In that case we have reason to suspect that the view of reference we’re working with is one that’s ill-suited to the realist’s needs.)

Surely, though, we can imagine a case where we have one scenario that prompts judgments about shared meaning that’s just subtly different from one that doesn’t. Think of a sorities series of possible worlds arranged so that there’s only a very small degree of difference (in the relevant respects) in the (purportedly) moral practices on those worlds. Somewhere in the series there would be a point at which “right” stopped referring to rightness and began referring to some other property. Yet, if we said “right” was univocal at one world, consistency pressure forces us to acknowledge that it’s univocal at a world almost exactly like it. True enough, but it’s plausible to think that our intuitions get weaker and less significant as the cases become more and more different from our home world. Here, it’s not clear that (a) we’d have any intuitive reaction at all or (b) what intuitions we did have would count for much.

A final wrinkle: the final part of this section of the paper will argue that the realist should use the notion of an idealized, end-of-the-day moral theory in addressing issues about univocity. If this is the right strategy, then there’s an additional problem, because practices that look the same (or very close) now might end up diverging by the end of the day (it’s more plausible to assume that they’re causally independent communities). Then we’d have a case where our intuitions clearly stood in opposition to the realist’s semantics. There are a few responses available to the realist. First, we’d have to hear more about how suitably idealized moralizers would start with the same raw material, use the right process of investigation, and end in different places, since we’d think, initially, that this wouldn’t happen. But it might turn out that way, perhaps because the seemingly trivial differences between our starting points ended up mattering a great deal. In this case, I think the considerations raised by my second argument against Horgan and Timmons (in section 3.4) would apply. Roughly, the response would be this. We should respect the views of suitably improved thinkers in epistemically superior conditions. Hence, we have reason to think that their moral views are better than ours. We also have reason to think that they (the idealized moralizers) would reject the idea that the idealized twin-moralists mean what they do by terms like “right.” Hence we discard the intuitions.

This reply works only if the two end-of-the-day theories are suitably distinct. If they’re quite similar, yet different enough to force the realist to admit there’s a difference in reference, this reply won’t cut any ice; we, along with our idealized, better-informed selves, would think that the other camp’s terms do mean the same thing. In order to head off this challenge, we’d need some arguments in normative ethics—we need to show that it’s implausible to think that two similar theories are both maximally coherent and stable end-of-the-day views. I think this is implausible, but giving convincing arguments for this will have to wait for another day. If, contrary to my suspicions, this is indefensible, the realist has another move available: in a case like this (with subtly different limit theories), we might think that a term like “right” is used in different senses, between which our current use remains ambiguous. (Compare the subjective and objective senses of “right”). Conceptual refinement is a plausible outcome of further inquiry. (Thanks to Nick Sturgeon for making this point and suggesting this example).
across diverse Earth contexts; (b) Moral Twin Earth looks just (or enough) like Earth; (c) the realist cannot account for shared meanings between Earth and Moral Twin Earth.

It’s hard to see how this can be done.

At this point, it’s natural to wonder about the realist’s prospects with disagreement here on Earth. After all, Horgan and Timmons’ generous concession is made for the sake of argument, and perhaps reflection on Moral Twin Earth is meant to show us that even Earthian dispute, as the realist construes it, is equivocal. What case can NMR make for preserving univocity in garden-variety cases of moral disagreement? Answering this question in detail must wait for another day. For now, I’ll offer some programmatic remarks intended to clarify the realist’s task.

As a preliminary step, we should distinguish three interrelated problems for NMR. The first is the one that’s occupied us so far, the univocity problem: how can a naturalistic account of our moral discourse earn the right to say that our terms mean the same thing in the mouths of speakers with such serious differences? Distinct from this, but connected to it in interesting ways, is the practical role problem: how can the answer to the univocity problem—in terms of a view of reference, for example—allow for the distinct practical and evaluative role of moral discourse? Third, there’s the moral inquiry problem, which is the problem of showing how an account of the univocity of moral terms can preserve and legitimate (at least parts of) the procedures we use in our moral inquiry. Even if NMR can give an answer to the univocity problem, it hasn’t really made much progress until it can show how its answer is compatible with preserving the practical role of moral discourse and giving us a satisfying picture of moral investigation. Conversely, the best way for the realist to respond to these worries, and to address some
of the underlying unease that drives them, is to show that the right kind of approach to reference can also preserve these other features of our practice.

Just for the sake of illustration, consider how a simple causal view of reference might go wrong in addressing these problems. If the reference of moral terms were fixed by some kind of initial baptism, then we wouldn’t be able to entertain certain seemingly important questions—for example, were the ancestors wrong to apply the term as they did initially? A view like this seems to undermine our means of ethical inquiry, since the answers to moral questions seem to be decided by causal-historical facts that appear to be irrelevant. There’s a related puzzle about the practical role problem. Suppose we learn that our terms stand in the right sort of relationship with a single cluster of natural properties. This discovery, by itself, doesn’t do anything to vindicate the evaluative or practical role of these terms. In fact, that role is undermined by the thought that our moral terms are held hostage to historical contingencies that seem not to matter in our moral deliberations.\footnote{If this objection were just the familiar line that any natural properties would fail to play the right sort of action-guiding role, simply in virtue of being natural properties, realism would be dead in the water, though the requirement itself would look suspiciously question-begging. What I have in mind, in talking about the practical role problem, is just the thought that the realist needs to do more to show how his account allows for moral terms to play their distinctive role—and questions about how reference is fixed are relevant to the practical or evaluative importance of these terms.}

I offer these observations only to focus our attention on what might make an alternative view more attractive. The lesson to be learned here is that whatever approach we take to the question of reference must not settle the wrong questions, or dictate the wrong means of investigating these questions. In addition, it must at least allow room for an account of why these properties (the ones picked out by moral terms, according to the
account of reference under scrutiny) are interesting and relevant to our practical decision-making. That is, it can’t preclude adequate answers to the practical role problem or the moral inquiry problem.

What would a solution to these three problems look like? One promising thought involves connecting our views about the reference of moral terms with our first-order normative inquiries. If it’s our end-of-the-day moral theory that supplies part of the story about the properties to which we’ve referred all along, then our normative inquiry and discussion become relevant to the question of the reference of moral terms, and so the problem of moral inquiry can be solved. It’s not just by accident that Boyd proposes a version of the causal theory of reference that invokes what we end up saying rather than some initial baptism. A view of this sort promises to preserve the normativity of moral questions by taking our deliberations about them to be part of the process that reveals (or determines?) the reference of moral terms. In addition, a connection between our moral inquiry and the reference of our terms helps to solve the practical role problem. Why should we care about these properties? In part, it’s because they’re revealed by our best prolonged attempts at reasoned inquiry into pressing questions. This doesn’t guarantee that we’ll continue to find these properties to be relevant to action, and it doesn’t ensure that all agents will be motivated to act in accord with their moral

---

135 The idea that our end-of-the-day moral theory is part of what determines the correct account of moral properties is a fairly popular one. See Brink (1984, 1989), Sayre-McCord (1997), and Sturgeon (1985). See also Rey (1983) for a related suggestion about concepts generally.

I should also note that Horgan and Timmons wouldn’t find this objectionable, since they assume that, at least on the Brink-Boyd view, the functional properties to which our moral terms refer are those described by the moral theory that’s arrived at via applications of a coherentist methodology. Though they don’t say much about the connection between (a) the properties that causally regulate our moral practice and (b) the properties described by an idealized moral theory, it seems that they’re happy to grant that the methodology we use in arriving at (b) will be a way of getting at (a).
judgments. Nonetheless, it makes it likely that these judgments will continue to play their distinct practical role in our deliberations.

Connecting issues of reference with questions about an idealized moral theory looks to be a promising direction for the realist, because the connection will help to explain how natural properties could play certain action-guiding roles, and because it helps to preserve the relevance of our procedures of moral inquiry. How would this strategy help with the univocity problem? If the correct account of our moral properties—the referents of our moral terms—were given by the end-of-the-day moral theory, then the issue of shared meaning comes down to, or at least essentially involves, the question of convergence. On such a view, we share meanings in virtue of the kinds of shared canons of evidence and argument, and deference to future theory, that would, under slightly idealized conditions, result in a convergence of our moral opinions.

This approach has several attractions. First, it fits neatly into common and appealing ways of thinking about shared meaning in cases of nonmoral terms. For example, we think that speakers share terms across changes in theories, or despite nonstandard beliefs involving the term. One part of the explanation for how this is possible involves the idea of deferring to the right authorities, or to the right kinds of reasons and evidence. And ethical convergence seems to involve similar phenomena—though in this case what we’d be deferring to is not some particular expert but to the deliverances of a refined future theory. So there’s a neat continuity between the moral and nonmoral cases, which should make this kind of view attractive to the naturalistic realist, who, after all, thinks that there’s far less of a difference between moral and nonmoral terms than is usually supposed. Second, this approach takes some of the
intuitive sting out of judgments of equivocality. On a view like this, we’d say that disputes that looked as though they were genuine turn out to be merely verbal only when it turns out that the parties involved would not, under suitably idealized circumstances, move toward agreement. And this is just the outcome that does make us suspect that disputes that we had thought were univocal are not, in fact, real disagreements.

So far, this is just an initial gesture on behalf of the realist. Though there’s not a fully developed view on the table, I think NMR’s most promising strategy is to invoke the limit of moral inquiry as a resource to use in answering the problems we’ve been discussing. This offers us a way of preserving the univocity of moral terms across disputes while retaining recognizable forms of moral inquiry and making room for the practical role of moral judgment. If this is the right approach to the univocity problem, then the realist’s hope for accounting for shared moral terms rests with the convergence of moral inquiry under suitably idealized conditions.

This, of course, raises larger issues that are beyond the scope of this paper. For now, I’ll make a few quick points in an attempt to bolster optimism about the prospects of such an end-of-the-day moral theory. There are a number of reasons why hope in convergence isn’t quixotic: the relative youth of secular moral inquiry, apparent examples of progress in spite of considerable distorting factors such as ideology and self-interest, and so on. We also need to bear in mind that what’s needed is not universal agreement among all people with (ostensibly) moral views. Just as the views of flat-earthers or creation scientists don’t count against the idea of scientific progress, the obdurate moral commitments of, say, white supremacists may not be evidence against the prospects of
moral progress. In addition, if we had independent reason to think that moral properties had causal powers, then we might think that convergence is made more likely by the influence these have on our inquiries.

On the other hand, it’s difficult to offer strong positive arguments for the prospects of convergence—as opposed to negative arguments showing that the usual considerations offered against this are inconclusive or irrelevant—without looking at particular normative theories. Here, we might appeal to work in normative ethics that attempts to show us that our ordinary practice is seriously confused, and that there are only a few appealing and consistent methods of reforming it. If these efforts are on the right track, we might hold out hope that serious critical scrutiny would produce convergence because of the appeal of some well-worked-out moral views.

A full examination of this issue must wait for another time. I gesture toward these arguments here only to point out that there are reasons for the realist to hold out hope for the resolution of moral dispute. Our broader argument against Horgan and Timmons must rest, for now, with the conclusion that if such resolution is plausible, then realists have nothing to fear from Moral Twin Earth. That argument began with a dilemma: if

---

136 This point is made in Gewirth (1960).
138 Think, for example, of Kagan’s (1989) arguments concerning the instability of ‘hybrid’ views that seem, initially, to be quite plausible. As another example, think of the efforts of Peter Singer, Peter Unger, and other radical moralists to convince us that we’re deeply confused about what morality requires.

Reflections on these considerations might tempt us to think that there just aren’t all that many coherent and stable end points when it comes to moral theorizing, because views that combine consequentialist and deontological elements will collapse under pressure. Furthermore, we might think that once we admit some consequentialist element into our moral theorizing—say, the idea that there’s a pro tanto reason to promote the good—we’ll have a hard time keeping ourselves from ending up with a straightforward consequentialism once we subject our views to critical scrutiny. If there are few stable endpoints, and if the consequentialist elements in our thought are deeply rooted and hard to contain, then all sorts of initial positions will end up producing Tc on reflection. If so, then surface differences between these starting points, which may look quite significant, are irrelevant to the question of shared reference that occupies us here.
twin-moralists are too much like us, NMR can guarantee the identity of moral and twin-moral terms; if they’re too different, then our linguistic intuitions are undermined, and the realist is out of danger. When we assume, with Horgan and Timmons, that the realist can account for the univocity of actual-world moral disagreement, we see that twin-moralizing is alien enough to bring our initial intuitions into doubt. (Notice that if this assumption can’t be made good, then the realist has real trouble here at home—there’s no additional problem posed by Moral Twin Earth). I suggested that the best way to establish the truth of that assumption, while preserving other advantages of realism, is to look to the end of inquiry. Hence, the realist’s full answer to the problem of Earthian univocity, and to the Horgan-Timmons challenge, depends on the issue of convergence. For the reasons I’ve mentioned, I think the realist’s prospects are fairly good.

3.6 The second argument: revising our intuitions

Unlike the argument we’ve just discussed, my second argument accepts the Horgan-Timmons example as it’s given. It then attempts to cast doubt on the significance of our intuitions concerning Moral Twin Earth. In this section, I’ll argue that when we take some important details of the case seriously—details that Horgan and Timmons must supply if their argument is to work—we’ll see that our initial judgments should be revised.

For the purposes of this argument, it’s important to keep in mind a few key features of the thought experiment that relate to our earlier discussion of the realist’s semantic commitments. When we suppose that the consequentialist theory Tc is the correct moral theory for us, we’re supposing that it “is discoverable through moral inquiry employing coheren list methodology” (Horgan and Timmons 1992b: 245). We try
to bring our views into ‘reflective equilibrium’ by weeding out overlooked contradictions and theoretical tensions; in doing so, we end up with a consequentialist theory, or so the example asks us to imagine. Meanwhile, twin-moralists are using this same process of inquiry, and it’s leading them to their deontological theory. The key point here is that we take this process seriously as a means of clarifying and improving our views. That is, we grant it normative authority; we think it makes our views better than they were before undergoing this kind of critical scrutiny.

With this in mind, we can proceed with the second objection, which has two steps. The first step is an argument (well, an appeal to intuition, at least) about what our better-informed selves would think about a better-informed set of twin moralists. The second step is an argument concerning the epistemic status of these judgments, and what effect they should have on our current intuitions. The conclusion is that our initial intuitive judgment concerning the univocity of moral and twin-moral terms should be viewed as spurious or misleading.

Let’s begin with another thought experiment. Imagine Earth and Moral Twin Earth not as they are now, but as they will be if we continue our moral inquiries in a serious and philosophically honest way. Imagine that we make efforts to engage in discussion with our peers, to discard moral commitments at odds with more fundamental and better-supported views, and so on, and that our counterparts do the same. In short, we engage in the kind of coherentist inquiry that Brink suggests. For the Moral Twin Earth argument to work, it must be true that twin moralists end up saying something like “well, it seems that deontological theory Td is true; this is the best account of morality” while moralists here on Earth say the same thing of their consequentialist theory. What’s
more, imagine that the results of these inquiries are widely accepted; these claims, and
some of the theoretical apparatus surrounding them, become part of our standard lore,
just as claims like “water is H2O” have. Earth and Moral Twin Earth are alike in having
a maximally stable, well-justified, and widely accepted moral (or twin-moral) theory.
This, of course, makes them different from the Earth of today, since, here and now, we’re
stuck with a discourse that seems torn between several incompatible theories with
roughly equal backing from firm intuitions and plausible justifications.139

Sometime after reaching this breakthrough in normative inquiry, we’ll imagine,
moralists and twin-moralists encounter each other for the first time. They discover that
they speak what appears to be the same language, and eventually talk turns to moral
issues. Then the conversation begins to go off the rails. Moralists are shocked to learn,
for example, that twin-moralists favor a policy of executing murderers, a practice that
they think should be condemned. They raise objections based on the minimal (or
nonexistent) deterrent value of executions, the economic cost of such a policy, its
contribution to skepticism about the justice system, and so on—that is, they appeal to
‘forward-looking’ reasons for being suspicious of the death penalty. Their efforts are met
with blank stares. “We know all that,” say the twin moralists, “but what has that got to
do with whether it’s right or not?” You get the idea. These two camps not only seem to
disagree about what’s right, but they offer completely different kinds of reasons for their
views.

139 Think again of the ease with which a radical moralist like Singer can use our seemingly innocent
intuitions and principles to drive us to conclusions we find utterly bizarre. Like Peter Unger, I think that
this is evidence for the claim that our ordinary thought is confused—though I don’t follow him past that
point. See Unger (1996) in particular.
What should we make of their conversation? It’s plausible to imagine that
moralists and twin-moralists would attempt to engage in serious inquiry to resolve their
differences; after all, each group is faced with an equally insightful, equally conscientious
body of thinkers who have reached different results.\footnote{Thanks to an anonymous reviewer for insisting on the importance of this point.} In the spirit of epistemic humility,
our interlocutors should be inspired to re-examine their old views and to consider the
positions and arguments offered by their new colleagues. Suppose that moralists and
twin-moralists decide to sit down together to hash out their (putative) disagreements.
Such an inquiry might result in changes in view on the part of one or both parties, or it
might leave them both completely unmoved.

First, imagine that the ensuing conversation alters the views of some of the
interlocutors. We might become deontologists, they might become consequentialists, or
we might reach some new synthesis that we both find more satisfying than our earlier
views. It might be that this change in theory represents a change in meaning: we’ve
suddenly gone over to a new way of talking as the result of some kind of conversion
experience. This is an unsatisfying response, since, after all, our idealized reasoners
changed their views in what looks to be a reasonable way, as the result of careful
discussion. It’s more plausible to say that the change in theory shows us that the speakers
involved (be they moralists, twin-moralists, or both) did not have a fully coherent and
stable end-of-the-day theory, precisely because they were susceptible to the justifications
for some alternative view. That these idealized reasoners found the rationale for a
different view compelling is some evidence that their pre-encounter position didn’t meet

the stipulations of the case, insofar as it wasn’t maximally coherent, stable, and so on.\textsuperscript{141} A change in view just shows that we weren’t \textit{really} at the end of the day after all.\textsuperscript{142}

If, on the other hand, continued conversation among our idealized inquirers doesn’t produce any change in their views, it seems increasingly reasonable to think that moralists and twin-moralists would be warranted in interpreting each other as using different terms. The conversation that at first appeared genuine might now be seen as equivocal. If their conversation remains at loggerheads, there’s a strong temptation to view the disagreement as spurious; it’s increasingly plausible for them to judge that their terms “right” and “wrong” simply mean different things. They have intractable disagreement despite knowledge of all the relevant facts of the case, they can’t really engage in meaningful argument with one another, and they each think the other has offered irrelevant reasons for its position. One interpretive option is mere dismissal: moralists (for example) might think that twin-moralists are deeply confused or simply incompetent with the relevant concepts. This isn’t attractive, because our idealized twin-moralists certainly \textit{appear} to be rational and coherent. It’s more sensible—and more

\textsuperscript{141} Keep in mind, too, that it’s essential to Horgan and Timmons’ argument that Moral Twin Earth and Earth have \textit{different} maximally stable and coherent end-of-the-day views. This is crucial because this difference is the means by which they force the realist to admit that “right” refers to different properties on different planets.

\textsuperscript{142} It may look as though I’m cheating by relying on an overly robust idealization. In this particular case, that doesn’t seem right, since, after all, our idealized moralists have already considered something like the twin-moralists’ theory Td, and rejected it. We’re assuming only that the idealized theorists have ironed out the wrinkles of their views and come to stable consensus. It seems odd that moralists might purge themselves of the deontological elements of their moral thought, commit themselves fully to Tc, and then turn around and find the justifications for Td compelling enough to change their minds.

Generally, we need make our inquirers only idealized enough to consider fully the reasons for competing moral views, actual and imaginable. That a view doesn’t have real defenders, or doesn’t have many, shouldn’t make it impossible for improved reasoners to think through its possible justifications. If so, then it’s hard to see what new material a distinct group (whether idealized or not) would bring to the table.

For now, I’m ignoring the significant problems involving the idea of a limit of inquiry or a theoretical end of the day. This notion faces serious challenges, yet I’m confident that any appropriate and defensible construal of it can do the work I need it to do here.
charitable—for each group to think that the other is simply talking about something else. Later in the paper I’ll argue that they do have a disagreement (of sorts) with each other, though it’s not a disagreement about what’s right. For now, I’ll claim only that the natural thing to say is that their disputes over “right” are equivocal. It’s hard to see why we should think otherwise, except for the orthographic identity of the words they use. At least, I think that the sensible thing to do, were one a member of the exploration party, is to dismiss the idea that their terms mean what one’s own do.\footnote{Again, the judgment of equivocation is made by a speaker who’s undergone this kind of idealization process. As I’ll argue below, part of the reason that we think the Earth-Moral Twin Earth disputes are univocal is because both Tc andTd get at something true, as far as we can tell. That is, both of these theories get a part of our (current, unrefined) thinking right. But Horgan and Timmons don’t account for the fact that our current moral theory is improved by the process that yields Tc; we get to consequentialism because we think, on reflection, this is the moral theory that fits best.}

If we take the stipulations of the case seriously, then, we see that we shouldn’t imagine fruitful discussion between moralists and twin-moralists. If their discussions led to changes in view, then they brought to the table something less than a maximally coherent and stable end-of-the-day theory. Furthermore, thinking about the ways in which a conversation between the speakers of Earth and Moral Twin Earth would break down gives us reason to think that our idealized selves would judge their ostensible dispute to be equivocal.

Now for the second stage of the argument. Here, the idea is that the judgments of the idealized moralists (concerning the equivocation involved in conversations with the Moral Twin Earthers) give us reason to reject the significance of our original intuitions about univocity. As a general principle, it seems sensible to give greater weight to one’s fully informed, carefully considered opinions than to one’s hasty, ill-informed judgment. Our judgments in idealized circumstances carry greater epistemic authority. Similarly,
it’s reasonable to grant more authority to one’s moral views after this process of inquiry than to one’s current views, because we take the process seriously from a normative point of view. We should take our idealized selves’ moral views seriously simply because such views result from a starting point we accept being modified by a process of sound inquiry. And our improved selves reject the idea that moral and twin-moral terms mean the same thing. Hence, we have testimony from an authoritative source against our initial judgment; since we think that the source is better positioned to make this judgment, we have reason to reject our own intuition.

It might seem as though I’m illegitimately extending the range of our idealized thinkers’ expertise. After all, we’re looking for semantic guidance, not moral guidance, and our enlightened future selves are experts only with respect to the latter topic. Are their semantic judgments only as reliable as our own? This objection misunderstands the nature of their authority. These future moralists are experts in the use of our moral concepts; they understand how to apply our moral terms better than we do. Just as experts with some concept in the natural sciences are better positioned to see which disputes about that concept are genuine, and which are spurious, our moral experts’ full understanding of our concepts—the ones we currently don’t fully understand—allows them to see more accurately which moral disputes are genuine.\textsuperscript{144}

\textsuperscript{144} I’m assuming that our concepts and the concepts of our improved selves are the same, or that our inquiry doesn’t change the meaning of the terms involved. This strikes me as the most plausible way to interpret the case, but, of course, other readings are possible. One might think that the results of our moral inquiry are so radical that we’ve changed the very meanings of the terms. Though I won’t say anything about this here, I’ll note that Horgan and Timmons cannot take this line, because doing so undermines their argument. Their objection requires that the meanings of our terms remain constant through theory-change, because they need for the meanings of our terms now to be fixed by the end results of this normative inquiry. If the end-of-the-day-theory doesn’t say anything about the current meanings of our moral terms, then a difference in end-of-the-day theory doesn’t entail a difference in meaning between moral terms and twin-moral terms now.
The argument has this form: idealized moralists and twin-moralists will see their putative disagreements as equivocal. But these idealized interlocutors have the same concepts (or mean the same things with their terms) as their present-day counterparts; they’re just better-informed about them. Because of this, we have reason to take seriously the judgments of the idealized future moralists. Hence we have reason to think that our initial judgment about the Moral Twin Earth case—that the disputes are univocal—is faulty. That is, we have reason to conclude that, contrary to initial appearances, we didn’t!mean what the twin-moralists meant with their terms.

It’s easy to see why we have this misleading intuition. After all, Earth and Moral Twin Earth are supposed to be very much alike (this ignores, of course, the considerations of the first argument). Because the practices look the same, we think that the terms involved have the same meanings. Keep in mind, though, that the similarities on which we base this judgment are just those parts of the practice that are rejected, on reflection, by thinkers engaged in moral (or twin-moral) thought. So, for example, we might imagine that current-day moralists and twin-moralists share a cluster of intuitions to the effect that the consequences of an act are somehow connected to its moral status. Moralists decide, on reflection, that this is right, but twin-moralists reject these thoughts as incompatible with “what they’re really getting at,” namely, the constraints of some deontological theory. So our first reaction to the Horgan-Timmons case is driven by details of the scenario that will, ex hypothesi, be rejected as superficial or misleading by careful inquiry. This, I think, explains why we have the intuitions that we do in a way that lets us conclude that such intuitions are less significant than they appear to be.
3.7 Preserving (nonmoral) disagreement: the appeal of Moral Twin Earth

The preceding sections have given independent objections to the challenge presented by twin moralists. Here, I’ll offer a broader diagnosis of the problem that lurks behind the unease of Hare, Horgan, and Timmons. While we should remain confident in our first two arguments against the particular difficulty presented by Moral Twin Earth, we should also seek to defuse the anti-realist’s general worry. In this section, I’ll take some initial steps toward that goal. This will allow us to offer a third defense against Horgan and Timmons. More importantly, we’ll be able to understand more clearly just what’s at issue between the naturalistic realist and her opponents.

Think again of the strategy that the Moral Twin Earth argument shares with Hare’s challenge to descriptivism. Both rest on the idea that we really do have something to talk about with the twin-moralists, the cannibals, or any other group with evaluative standards, no matter how different these are. Because we think that there’s something important at issue in these conversations, we’re unsatisfied with the idea of dismissing the dispute as mere linguistic confusion. The two challenges provide us with just the kinds of disagreements that normative language should help us resolve: we need to decide to take scalps or to turn the other cheek; we need to decide whether to act rightly or twin-rightly. An account of normative language is in serious trouble if it doesn’t allow for that language to play its distinct and important practical role.

It’s understandable that NMR appears vulnerable on this issue. The naturalistic realist identifies moral judgments by their content, and offers a picture of moral properties as metaphysically ordinary and motivationally inert. The practical role of these judgments—something that strikes noncognitivists as their core feature—is not, by
the realist’s lights, internally or conceptually connected to moral discourse. On this view, it’s possible to imagine some group invoking different properties in their deliberations about how to act; their judgments might have the same practical role as our moral judgments while having different content. The realist tells us that the terms used to express these judgments simply don’t mean what ours do. If the realist is right, then the foreign practices in which these terms are embedded aren’t properly called moral practices, in an important sense. Yet we’d think that there is something important at issue between us and the participants in this rival practice. Now the question is whether the realist can make sense of this quite general conviction once she’s conceded that we don’t have a case of moral disagreement.

What’s at issue seems to be this: we disagree about what to do.\(^{145}\) We don’t think of these disputes as spurious because the two parties offer competing endorsements or contradictory directives. They’re not talking past one another because they’re offering incompatible answers to the same question about action. Our hypothetical community might not talk about moral properties at all, but they do offer advice about what to do. Our disagreement, on the realist’s view, is not about what’s right, it’s about how to act—but it is, nonetheless, genuine, and not equivocal, dispute. If we can give a satisfying account of what’s going on when speakers disagree about what to do, we’ll earn the right to say that there is something at issue in cases where radically different evaluative standards come into play.

\(^{145}\) This locution is suggested by Gibbard’s attempts to describe a sense of “bland, flavorless” endorsement that captures “the last ought before action.”
This point also enables the realist to offer another reply to Horgan and Timmons. Suppose, just for the sake of argument, that Moral Twin Earth did present us with a case where moral and twin-moral terms referred to different properties. Imagine a conversation between moralists and twin-moralists about a topic on which Tc and Td seem to give contrary advice. For example, they might be wondering whether or not to assent to a claim such as “it’s right to lie in exceptional circumstances.” We’re granting that there’s no argument about what’s right, because the sentence means different things in the mouths of different speakers. But there’s still a real dispute here. Insofar as they’re committed to doing what’s twin-right, the twin-moralists urge us not to lie, and good moralists urge just the opposite. Speakers disagree without sharing the term “right.” What’s more, this is a familiar form of conflict here on Earth. People sometimes agree about what morality requires, and then go on to disagree about whether to follow morality’s commands—perhaps because one party thinks that considerations of prudence or etiquette carry the day. In these cases, the dispute isn’t about what’s right; it’s about whether the properties of morality, prudence, or etiquette are the ones that settle what to do.

The strategy of this third argument urges us to revise our initial views of the disagreement’s location. Is this adequate? A few points suggest that it is. First, some disagreement is better than none at all. This reply at least preserves an important part of the phenomena, namely our sense that there’s something to talk about when moralists and twin-moralists get together. Second, we might note that the question of whether our dispute is centered on “right” or somewhere else is a fairly sophisticated one, and it’s not...
clear that our ordinary linguistic intuitions have much to say about the matter. Third, we’re in a position to give an explanation of why ordinary speakers might have trouble distinguishing between dispute about what’s right and dispute about what to do. Given the association between judgments of rightness and motivations to act accordingly, it’s easy to see how speakers would link a moral term with its broader evaluative accretion.

What’s needed now is a more detailed discussion of just how to understand the disagreement on which this third argument rests. In what follows, I’ll canvas some different ways of understanding the disagreement that’s being preserved. As we’ll see, this question involves some very deep issues that can’t be settled here. My goal for now is more modest: I want only to show that there are a few options available to the realist, each of which is enough to provide a satisfying response to the Horgan-Timmons objection.

The first option is to rest on the idea of clashing attitudes or commitments to action. Moralists’ attitudes are in favor of doing what’s right; twin-moralists favor the twin-right act. Since our two normative theories (Tc and Td) come apart in their assessments, agents who are committed to doing what their theory labels “right” will be motivated to act in contradictory ways. So dedicated moralists and twin-moralists have a kind of disagreement in attitude. This is, after all, a form of disagreement, and so it might look as though we’ve saved some genuine incompatibility. The problem with this suggestion is that we’re not really any clearer about what kind of attitude this is. It can’t be simply that we differ in attitude, the way we might differ in blood pressure; this isn’t yet a robust enough notion to ensure disagreement, let alone discussion and deliberation about broader normative questions. In order to capture the kind of incompatibility we’re
after, we’d need some kind of attitude with some ‘outward-reaching’ force. My pro-
stance toward doing the right thing has to conflict with, not merely be different from,
your endorsement of the twin-right act. That is, I need to do more than commit myself to
morality’s demands; I need to think that you should do the right (instead of the twin-
right) thing, as well.

These kinds of considerations might lead us to think that there are different (and
incompatible) judgments here—judgments about ‘the last ought before action.’ Earlier, I
alluded to a kind of conflict that’s familiar here on Earth: conflict between different
evaluative perspectives or standpoints. Morality demands one action, while prudence,
etiquette, and so on demand others. In deciding how to act, we need to know not only
what’s right, or what’s prudent—we also need to know whether to heed the demands of
morality or prudence in those cases where the two come apart. Different people
sometimes come down on different sides of this issue, even though they agree about the
answers supplied by these competing species of evaluation. One person might be
committed to doing the right thing, while another is set on the profitable thing, or the
polite thing, even if this requires violating moral demands that they recognize. We seem
to engage in conversation and disagreement about what to do. These conflicts also take
place within an agent, as when we wonder if moral reasons are outweighed, in some
instances, by other kinds of considerations. So it does seem that a further decision is
required once we know what’s asked of us from these competing evaluative standpoints.
And this, we might think, is where we disagree with the twin-moralists: what’s at issue in
this conversation is whether rightness or twin-rightness is the correct guide to conduct.
Should we listen to morality, or to twin-morality? That, of course, is a question that can’t
be answered *within* one or the other, as both claim to give us the answer to the question of what to do. Then again, so does financial discourse.  

Reflecting on this phenomenon raises a challenge to our third argument. However it accounts for our disagreements with the twin-moralists, it must ensure that the same problem doesn’t merely repeat itself at the next level. If we want to preserve dispute about *what to do* with our twin-moralist colleagues, we can’t allow for these disputes to be equivocal between, say, ourselves and the residents of some Normative Twin Earth. I’ve suggested that the conflict between moral and twin-moral advice is like the conflict between prudence, morality, etiquette, and various other kinds of evaluations. If this is right, we have grounds for confidence that our judgments of what to do are not parochial in the way we’re conceding (for the sake of argument) moral terms to be. After all, our confidence in the catholic nature of our disagreements about the last ought before action is unshaken by the variety of considerations that enter into that discussion. Most of us wonder, at one time or another, about what kinds of reasons to follow (moral, prudential, aesthetic, and so on). All of these have significant pull on our decisions about action. Yet despite the diversity of considerations taken into account in discussing what to do, we aren’t really concerned about the univocity of these conversations. Why should another normative standpoint, that is, twin-moral discourse, pose any additional threat?

---

147 One might think that twin-moral discourse differs from other evaluative perspectives (those of etiquette, economics, prudence, and so on) in having some kind of overriding purport. That is, like moral discourse, and unlike any other, twin-moralizing demands absolute allegiance; its requirements—from its point of view—*cannot* be overridden. Surely it does, but, we might think, so does the financial point of view, the prudential point of view, and so on. One might claim that prudential reasoning (for example) never really conflicts with moral requirements, because a correct understanding of prudential requirements shows that they respect the dictates of morality. This just isn’t plausible. When someone follows self-interest instead of morality, we think they’re wicked, or weak-willed, or overwhelmed by temptation, not that they’ve failed to understand the demands of prudence. On the contrary, it seems that these people understand prudential demands all too well.
This, of course, is just some initial reason for skepticism about this challenge facing our third argument. In order to say more about why that challenge won’t succeed, and in order to flesh out the nature of the disagreement between moralists and twin-moralists, we’ll need to say more about what’s involved in making a judgment of what to do, all-in. There are at least two options. A naturalistic realist about morality could offer an analogous account of the last ought before action. She could give an account of all-in endorsement that construes these judgments as being about motivationally inert natural properties. Or she could combine her view of moral discourse with a different treatment of our talk of what to do. The view I’ll consider here combines moral realism of the sort we’ve been discussing with expressivism about all-in endorsements. I don’t mean to suggest that these are the only options, and I survey them only to give a sense of how different approaches would handle the constraint on the strategy used by our third argument. Nonetheless, these issues are worth considering here not only because of their relation to the Moral Twin Earth argument, but also because of their more general importance in thinking about the connections between moral assessment, endorsement, and motivation.

Suppose we’re inspired to offer a Brink-Boyd style treatment of our discussions of what to do. What we’d be doing, then, is giving an interpretation of some notion of evaluation that’s broader than moral assessment, and we’d say that, in making these evaluations, we’re referring to natural properties and making (straightforwardly) truth-apt judgments about them. Now, in order to make this plausible, it seems that the realist should pick some substantively constrained ‘ought’ and identify that as the last ought before action. The most obvious candidate, I think, is the rational ought. (We don’t
think the moral ought is a good candidate here to the extent to which we think the moral answer doesn’t settle what to do, all-in. But we might think there’s less of a gap between what’s rational and what to do. If someone doesn’t think he should do what he has most reason to do, he’s just being irrational, and that’s the end of the story.) NMR defines moral properties in functional terms by reference to the best end-of-the-day moral theories—perhaps this related account of rationality judgments would follow the same strategy. Of course, as in the moral case, it’s crucial to specify the functional core of those properties at the correct level of generality. If the specification is too fine-grained, then we’ll mistakenly think that we have different properties; if it’s too coarse-grained, we’ll miss distinctions between properties. It seems plausible to think that our best end-of-the-day view about what we have (all-in) reason to do would be broader and more general than the functional role of narrowly moral properties. After all, our discussions about what to do admit of a broader range of considerations than does moral discussion. Because the task of our practical reasoning is partly to adjudicate between the competing claims of various normative standpoints (e.g., moral, prudential, aesthetic, etc.), we have reason to think that the relevant functional properties will be specified in relatively coarse-grained terms. If so, then it becomes increasingly difficult to imagine what the relevant Twin Earth-style counterexample would be like.

There is, I think, another way of thinking about the last ought before action. This combines realism about moral discourse with expressivism about all-in endorsement.

---

148 It may be that particular reasoners think that what’s morally right settles what to do, and that non-moral reasons are no reasons at all, at least when they compete with moral considerations. And they might be right. But surely we need to come to this further conclusion about what to do by a slightly different means from the way we answered the question of what’s right, since we have at least one additional question to settle, namely, “does it always make sense to do what’s right?”
According to this view, moral rightness is a matter of natural fact, but an answer to the question of what to do—not from a moral or a prudential or perhaps even a rational standpoint, but what to do, all-in—is not a factual judgment but an endorsement of one course of action or one set of reasons for action. When I get behind doing the right thing, I’m expressing my acceptance of certain norms, or urging others to act accordingly, or something along these lines.

Why would anyone be attracted to this combination of views? Admittedly, it may seem odd to defend naturalistic realism and accept expressivism at the same time. It’s worth keeping in mind, though, that every expressivist is either an expressivist about everything—not the most plausible view, on its face—or a realist about some discourse or other, whether logical, prudential, financial, or whatever. What’s being proposed here is realism about moral discourse and expressivism about something else, namely, judgments about what to do, all-in. There are a few reasons for finding this kind of picture appealing. The first has to do with skepticism about the ability of any substantively constrained ‘ought’ to play the role of the last ought before action. The second, related consideration has to do with the connection between judgments of what to do and motivation. The third reason concerns the differences in descriptive content between moral discourse and talk of what to do, all things considered.

First, we might be a bit uneasy about the identification of some substantive evaluative standpoint (morality, prudence, rationality...) with the last ought before action. The reasoning here is familiar: there’s a kind of open question that can be raised about whether or not the deliverances of any of those perspectives settle what to do. We’ve discussed this in conjunction with the moral ought, but we might have similar worries
about the rational ought as well. Suppose we think that what an agent has reason to do depends on his (intrinsic) desires, broadly construed. We might admit that a person with wicked desires has *reason* to do horrid things, and that he wouldn’t be irrational for doing them, but we might still say, sensibly, that he ought not to. What we’re doing is more than merely pointing out that from the moral point of view, which he cares nothing about, he’s doing something wrong. This is something to which he can assent. We’re also saying that this point of view settles what to do, that he should (all-in) do what’s right instead of acting as he desires. So it seems coherent to think that, when we evaluate others’ actions, we *don’t* take facts about rational requirement to settle the question of what to do, and this indicates that we think the two questions are distinct. This argument can be repeated against any substantively defined evaluative perspective, and this, in turn, might lead us to think that the function of the last ought before action is solely to commend. It plays a prescriptive, not descriptive, role, as Hare might put it.

The second consideration in favor of this combination of views has to do with motivation and action. Naturalistic realists have made a great deal of the fact that sometimes moral judgments *don’t* motivate agents to act accordingly; as they like to point out, the tie between judgment and motive mustn’t be too tight to account for the everyday phenomena of our moral experience. At the same time, we might think there’s a tighter connection between judgments of what to do and motivation. After all, we can more easily make sense of someone unmotivated to follow through on his moral judgments than we can of someone who invokes the last ought before action and then

---

149 It’s interesting to note that this kind of evaluation makes sense only from the third person; first-personally, the question of what to do is settled by one’s commitments. I owe this point and the example to Don Hubin.
fails to act. Certainly there will be cases of weakness of will and the like, but it’s plausible to think that there’s a very close connection between this ‘bland, flavorless ought’ and action. The larger gap between moral assessment and action makes it more plausible to think that moral properties are natural, while the tighter connection between all-in judgments of what to and action makes it plausible to think that these judgments aren’t about natural properties.

Third, moral realists have sometimes been impressed by the fact that our moralizing is fairly “thick” in its descriptive content: moral judgments can’t be about just anything; moral standards can’t be eccentric in certain ways without failing to be moral standards at all. When we think about morality, we have some rough idea of the kinds of considerations that are relevant (facts about benefits and harms to creatures like us, for example) and facts that don’t matter (the height of an agent, perhaps). This descriptive richness is important, since it suggests that the way to capture what’s distinct about morality is by appeal to a certain subject matter or content, not a particular set of formal features. At the same time, we might think that talk of what to do, all-in, is descriptively thin: people can take all sorts of considerations to settle how to act, and, as long as they satisfy certain formal constraints, we can think of them as coherent.\textsuperscript{150} And, of course, they’re still making a judgment of what to do, unlike the moral eccentric, whose deviance might prevent him from taking part in moral discourse at all.\textsuperscript{151}

\textsuperscript{150} So, for example, we can imagine people who make their decisions by thinking about beauty, or their own interest, or some other kind of consideration. These people have moral failings, we might think, but it’s not clear that they’re incoherent.

\textsuperscript{151} Hare tried to identify moral claims by their formal features alone, and this led him to some implausible views. We might think that Hare was on to something but erred in thinking that moral discourse could be identified without appeal to its content. The combination of naturalistic realism about morality and
We might think, with Hare and Gibbard, that some of the uses of language—even some that appear, on the surface, to be assertoric—have commendation and condemnation as their principal functions. If so, it may be that any descriptive account of these uses will fail to capture what’s crucial to them. These theorists err, though, in identifying the particular piece of language that plays this role consistently. Hare’s view is problematic because it fails to allow any room between moral judgment and all-in endorsement. Gibbard’s view strains natural language by stipulating a sense of “rational” that equates it with bland, flavorless endorsement. It’s better, I think, to identify this idiom as simply talk of what to do. Or, on the other hand, it might be more plausible to say that there’s no one idiom that consistently expresses a view about the last ought before action. Instead, different sorts of normative judgment can be expressed in conjunction with all-in endorsement. At times we think that what’s right settles the question of what to do, while at other times it’s what’s prudentially best that carries the day, and so on. On this model, the twin-moralists have simply thrown their weight behind a different evaluative standpoint. Even conceding that we’re not at odds about rightness or twin-rightness, we still have room for some form of genuine dispute.

All of this is quick, of course. Developing this suggestion more fully requires specifying just what noncognitive states are involved in these endorsements, and how they allow us to preserve the contours of practical discourse. We’d have to make sense of the canons of reasoning and consistency pressures that are at work in our talk of what expressivism about the last ought before action is a way of preserving the insight in Hare’s work while still holding on to a thicker, more contentful picture of our moralizing.

152 Note, too, that the same sentence is often used to express contradictory recommendations in different contexts. I might say “it would be financially crazy to take this trip!” to urge us to stay home, or I might say it as a way of encouraging us to throw caution to the wind.
to do. There’s a danger that we’d meet this challenge too well, and in doing so provide the makings for a satisfying expressivistic account of moral discourse, thus undermining our proposed combination of views. This is unlikely: recall that some of the most serious problems facing expressivism resulted from the need to constrain moral judgment so that it remains distinct from all-in endorsement. This, as we’ve seen, is a difficult task. Developing this argument in detail must wait for another day. All that matters, for present purposes, is that we have some way of preserving real (that is, non-equivocal) disagreement about questions of what to do.

3.8 Concluding remarks.

I’ve given three arguments against the Moral Twin Earth challenge. The first begins with the observation that we don’t know much about the workings of twin-moral discourse, and offers Horgan and Timmons a dilemma. If twin-moralizing is too similar to our own moral practice, then it’s not clear that the argument delivers the semantic result they need; if it’s too different, the result is no longer counterintuitive. The second argument questions the significance of our initial interpretations of our twin-moralist counterparts. It’s natural to think that suitably improved moral evaluators would interpret them as using different terms, and it’s natural to give more weight to these intuitions than to our own. Hence we have reason to think that our intuitions are spurious. The third argument takes a different strategy by abandoning moral disagreement and focussing on disagreement about what to do. If the realist can preserve this form of (nonmoral) conflict, then the dialectical force of Moral Twin Earth is undermined.

Together, I think that these arguments are sufficient to allay realists’ fears about this forbidden planet. More importantly, these responses highlight broader concerns
about realism, in particular, the ways in which such a view might accommodate the role of our moral inquiry within its account of reference, and—a related point—how it might make sense of the normative function of moral judgment and discourse.
4.1 Introduction

Our discussion of Moral Twin Earth establishes that the Horgan-Timmons argument is not a threat to NMR. The broader issue of univocity, though, is still a threat, albeit one that Horgan and Timmons misconstrue. In this chapter, I’ll argue that moral realism does face a serious problem in accounting for the univocity of moral conversation.

As we’ve noted, the realist is faced with two puzzles about disagreement. The first has to do with the apparent intractability of moral disputes, since this seems to sit uneasily with the realist’s claims about robust moral properties. The second puzzle is semantic: the realist must show how these sorts of disagreements are compatible with univocal conversation. We might be tempted to think, in response to this diversity of opinion, that speakers aren’t referring to properties at all, because it’s hard to see how to square their divergent theories with a shared subject matter. Perhaps they’re evincing competing sentiments or issuing incompatible prescriptions instead, as a familiar line of thought would have us believe. The realist must, in response, reconcile her moral semantics with the apparent univocity of our moral conversations.
My concerns here are primarily with the second problem and its relation to the first—that is, with how the realist’s answer to the semantic difficulty requires thinking about the nature and alleged intractability of actual moral dispute. I suspect that a satisfying answer to the semantic worry will require the realist to say that moral disputes are resolvable, and, what’s more, that they are resolvable in a particular way. Some realists might be happy to make this commitment, but, as we’ll see, there are grounds for skepticism. The final word on the issue of convergence or the resolution of dispute can’t be given here, but it will be illuminating to examine the connection between the two problems. My aim is to show that though it’s faulty in its most famous versions, the semantic worry involves a serious and interesting challenge to moral realism.

First, I’ll describe some important features of moral discourse—features that, I think, do not fit easily into the realist’s view. Then I’ll use these features, in conjunction with some worries about the philosophy of language, to suggest that a popular and initially plausible line of response to the semantic worry is insufficient. This leads us to issues concerning the resolution of moral disagreement and how some claims about moral convergence would be of help to the realist. Though I won’t take on the convergence question here, I’ll argue that only a particularly strong version of the convergence claim will help the realist solve the semantic worry.

4.2 How moral discourse is distinct: three disanalogies

Naturalistic realism views moral discourse as being roughly on a par with scientific inquiry: the view holds that both make truth-apt claims and that both use terms that refer to natural, motivationally inert properties. Given this claim about the continuity of language across the descriptive/normative divide, we’d expect that realism would face
some pressure accounting for what look to be distinctive features of normative discourse generally and moral discourse in particular. In other words, if we’re looking for problems for the view, the obvious place to start is with potential disanalogies between moral and empirical inquiry. In this section, I’ll canvass some of moralizing’s distinctive features and suggest why realism might have problems accommodating them. Keeping these phenomena in mind will help us to see that the realist’s traditional answers to the semantic puzzle are not entirely satisfying, because they fail to take these aspects of our moralizing seriously.

First, our moral concepts seem to be essentially contestable. They allow for a wide degree of latitude in their application, and competent speakers can disagree not only about borderline cases but about cases that one party views as paradigmatic. Moral discourse is permissive—speakers have great freedom in the content of their moral views, and can exploit that freedom while remaining participants in moral discourse. Without drawing their conceptual competence into question, speakers can disagree both about the extension of moral predicates and about how they are best characterized. Yet speakers with idiosyncratic standards for using terms like “right” are nonetheless interpreted as having moral standards; we’re not often tempted to think that they’re engaged in a different sort of evaluation.

The contestability of moral concepts seems to be connected, in interesting ways, with a deeply-rooted pessimism about the possibility of their successful analysis or explication. G.E. Moore’s Open Question Argument has fallen on hard times as a bit of

---

153 This is part of what makes moral disagreement appear especially problematic—we have radical disagreement about, say, right-making features as well as about the extension of the predicate.
philosophical argumentation, but many commentators think that it nevertheless tells us something interesting about our moralizing. In its simplest form, the argument points out that, for any proposed analysis of normative concepts, competent speakers can ask, “I see that [it has the analysis’ non-normative property of choice], but is it really [good/right/art…]?” What’s interesting about the questions generated by Moore’s schema is that they aren’t open only in the sense that they might be asked by a competent speaker. “I see that it’s water, but is it H2O?” is, or was, open in this sense, but that doesn’t seem to tell us anything particularly interesting about the concepts involved. Its normative counterparts, on the other hand, seem, at least in many cases, to be live epistemic options in ways that questions about the composition of water do not. Of course, the explanation for this phenomenon might be simply that we’re in a relatively impoverished position vis-à-vis the correct moral theory. On the other hand, the intuitive force of these questions, and the suspicion that their openness isn’t simply a product of our feeble efforts at moral science, might well make us uncomfortable with this optimistic line. Here, the salient point is that this open texture (and its accompanying analytical pessimism) suggest some potentially problematic differences between normative and empirical inquiry.

---

154 See for example Darwall, Gibbard, and Railton: 1992. Others, of course, think that the argument is mistaken through and through. Thanks to Don Hubin and Nick Sturgeon for insisting on the significance of this point.

155 Note, too, that the essential contestability considerations are important to our worries about the apparent intractability of moral disagreement. If moral concepts are distinct in this way, then we’d expect that moral dispute would look different from, say, disputes within the natural or social sciences.

It’s also worth noting that this contestability is closely related to the next putative disanalogy, the troubling nature of authority and deference in the moral realm. Both speak to a sort of independence granted to moral speakers that isn’t given to users of, say, the language of physics.
Second, our moralizing seems to lack the relationships of authority and deference that characterize non-normative disputes. We draw our competence into question if we neophytes don’t defer to the established experts when we’re arguing about, say, questions of chemistry, or medicine, or economics. But the notion of moral expertise is considerably more problematic: I might be making a moral mistake if I reject the advice of your favored moral compass, but I’m not incoherent or incompetent, as I might be if I rejected the views of, say, Richard Dawkins on questions of evolutionary theory. That’s not to say that we never recognize moral authorities or experts. Clearly, we often rely on the advice of others, and we sometimes see the input of external authorities as more than mere advice, as when we think, for example, that a religious leader’s pronouncements settle a contentious issue. But this seems different, in an interesting way, from authority in the case of the natural sciences. Obviously, there’s no body of agreed-upon experts in the case of moral dispute, as we’d expect given the apparent intractability of those disagreements. Furthermore, we might think that individual speakers grant authority in the normative case, and retain responsibility for their higher-order commitment to follow a particular moral guide, whereas acceptance of the authority of, say, scientific experts seems to be a requirement for competent participation in that sort of talk. Thus, in the moral domain, authority has a kind of dependence on individual speakers that isn’t present in empirical inquiry.

156 Just as an illustration of this point, consider that deference plays a large role in thought-experiments that hinge on speakers with confused ideas about some phenomenon.

157 It seems to me that at least some ostensible moral experts aren’t people with greater propositional knowledge about moral theory; instead, they’re people with more subtle emotional and ‘perceptual’ capacities—along the lines of a ‘knowing how’ rather than a ‘knowing that.’
This issue of moral authority will be important in what follows, because, we might think, the presence of fairly clear-cut deference relations in empirical inquiry is a useful resource in keeping speakers on the same semantic page when it comes to theoretical disputes. Moral talk, on the other hand, seems to grant greater autonomy to individual subjects. Because these deference relationships are more problematic, the naturalistic moral realist is faced with the task of picking up the slack. The semantic puzzle will be harder to solve because we can’t appeal—at least not in just the same way—to considerations of authority when it comes to making sense of univocity and normative dispute.

The third, related feature of moral discourse that puts some stress on the naturalist realist’s picture has to do with the relationship between moral judgment, motivation, and practical reasoning. Just what connection exists between these is controversial, of course. It’s probably too strong to say that moral judgment entails or involves some pro tanto motivation to act accordingly—but we still need to preserve and explain the tight (if contingent) connection between judgment and motivation. An account of our moral discourse that leaves us completely indifferent to moral judgment is, for that reason, deeply flawed.

The link between this feature and the semantic puzzle is a bit harder to see. Just for the sake of illustration, suppose that judgment-internalism is true, i.e., that making a moral judgment includes, as a conceptual matter, a defeasible motivation to act in accord with that judgment. This, in conjunction with familiar Humean psychology, would lead us to expect that moral judgments would be unconstrained by restrictions on content, either from substantive ‘ground rules’ or from deference relations. Were they so
constrained, we would be bound to have particular motivations, and this is antithetical to the Humean ur-thought that conation is independent of belief. In general, and for the same reasons, the tighter the connection between moral judgment and action, the looser the content restrictions on moral judgments. This suggests, though it does not show, that the realist will have a difficult time ensuring an appropriate action-guiding role for moral properties, as she understands them—a problem made more pressing by the diversity of moral opinion. How can speakers share reference to these properties—properties that are picked out by a moral theory some of them reject—if their conations pull them in such different directions, and if the right account of moral properties must allow them to play a suitably robust action-guiding role? We might worry that the realist can preserve univocity only at the cost of giving short shrift to the ties between moral judgment and motivation.

(It’s worth noting that the so-called action-guidingness of moral judgment helps to explain the other features we’ve canvassed here. Certainly, all three are closely related. This leads to a fairly obvious objection: hasn’t our discussion of these phenomena implicitly assumed either judgment internalism or some other strong thesis about the motivational force of moral judgments? The naturalistic realist, who typically rejects internalism of this sort, can make this point as a way of denying either the existence or the seriousness of the disanalogies we’ve been considering. Our attempt to put pressure on realism, its defender might say, has simply relied on confused ideas about the ties between judgment and action. The best response to make to this objection—which, I think, is both fair and serious—is to note that these phenomena are supported by the actual workings of moral conversation, not simply by an abstruse point
about moral psychology. While we need to be careful not to overstate the strength of these disanalogies, we can nonetheless insist that they are differences that must be accounted for and explained.)

All three of these phenomena—contestability, authority, and motivational force—are closely tied to what we might think of as a platitude about the point of moral discourse: its role is to guide conduct and to allow us to reason together about how to act and live. If moral discourse is too descriptively rich—that is, if there’s too much substance or content built into its very concepts—then it’s less well-suited for its distinctive task, because it seems to cut off certain kinds of serious challenges. The radical critic, for example, doesn’t get a fair hearing within morality if it’s simply a constitutive rule of moralizing that her view is faulty. We might read her as offering a critique of morality from the outside, rather than urging us to reconsider our moral views, but this response just underscores the way in which a contentful view of moral discourse might fail to live up to its job description.¹⁵⁸ The realist’s view of moral properties seems to be in tension not only with the phenomena we’ve been discussing, but with the broader point about the purpose of moral discourse. The naturalistic realist’s suggestion that we’ve been hooked up to the same set of robust moral properties all along seems to deprive us of some degree of autonomy. After all, the answers to our moral questions are fixed ab initio, rather than determined by our deliberations, so our inquiry reveals prior moral facts. Thus it undermines the freewheeling dispute that’s characteristic of moral discourse. As this Kantian allusion suggests, we also face a worry about the authority of moral facts, should they be as the naturalistic realist claims. A satisfying treatment of

¹⁵⁸ See for example Brink (1986b) for a discussion of this point.
these concerns must show that the realist’s view can accommodate or explain the appearance of serious differences between moral and empirical inquiry. Specifically, we’ll need to examine how these allegedly distinctive features of moral discourse stand in the way of a quick and familiar solution to the semantic puzzle.

4.3 The traditional response

The realist’s first line of response to the semantic puzzle is, by now, familiar. Moore’s notorious Open Question Argument relies on the idea that claims of property-identity are indubitable to competent speakers, but subsequent reflections—and attention to the results of various empirical inquiries—have shown that property-identity and synonymy are two different things. It’s a substantive discovery, not a bit of linguistic analysis, that tells us about the microstructure of water or gold. Saul Kripke’s arguments in Naming and Necessity suggest that it might be in virtue of causal ties, rather than associated descriptions, that proper names refer to individuals. This is motivated, in large part, by the thought that we can refer to a person despite variations and errors in our beliefs about her. Hilary Putnam extends this view to include natural-kind terms and argues that relationships beyond a speaker’s cognitive ken help to fix the content of his thoughts and the meanings of his terms. In making the case for semantic externalism, Tyler Burge has impressed on us the importance of the distinction between talking about something (having propositional attitudes about it, and so on) and completely understanding it. Buying into incomplete or erroneous theories is no block to concept possession: as Burge’s famous example reminds us, a medical neophyte might be
thinking of arthritis despite failing to understand that it’s a disease of the joints.\textsuperscript{159} We can leave aside the more controversial implications of this work in the philosophy of language; what’s crucial here, for the realist, is the underlying theme that variation in belief is compatible with a shared subject matter.

The problem we’ve been considering involves making sense of shared meaning given the considerable interpersonal variation in what we might call cognitive shape: the collection of beliefs, standards, epistemic norms, and so on, that make up a person’s (perhaps inchoate) moral view.\textsuperscript{160} Given these different commitments, how do we manage to refer to the same properties? Armed with the famous dictum that ‘meaning ain’t in the head,’ and some of its semantic-externalist corollaries, the realist can argue that univocity is compatible with this sort of variation between speakers.\textsuperscript{161} As Kripke, Putnam, Burge, et al. have helped to suggest, the realist need not be embarrassed by the diversity of moral opinion, because this is no threat to agreement in meaning. We can hold on to quite different theories and still have genuine, rather than merely linguistic, disagreement.

Note, as well, that the moral realist has benefited from a re-examination of the workings of empirical inquiry. Reflecting on last century’s physics reveals that even

\textsuperscript{159} Note that Burge’s point is not, primarily, that the patient and his doctor share the concept arthritis, but that it’s factors outside the head that fix content. We’re supposed to think that the patient has [arthritis] while his individualistically identical twin, in a world where “arthritis” picks out a broader range of ailments of joints and other body parts, does \textit{not} have the same concept. Thus we’re meant to see that individualistic description underdetermines content.

\textsuperscript{160} Putnam’s original point in “The Meaning of ‘Meaning’” is that agreement in “what’s in the head” is not sufficient for agreement in meaning (since, after all, Oscar and Twin-Oscar have the same thoughts, individualistically described, but possess different concepts). And my point here is that this work helped us to see that in-the-head agreement is not \textit{necessary} for shared meaning.

\textsuperscript{161} I take the term “cognitive shape” from Simon Blackburn.

One interesting issue here is whether the realist needs the ‘new theory of reference’ itself, or if his needs are as well-served by any view that makes sense of the phenomena that \textit{motivate} the new theory of reference, e.g., partial mastery, non-standard theory, and so on.
seemingly indubitable claims are open not only to doubt, but to rejection: “every particle has a mass and a velocity” has the ring of conceptual truth to it if anything does, but theory-change is often no respecter of our intuitions. These thoughts underscore the analogy between moral and scientific inquiry—if conceptual bedrock is missing from physics, then its absence from our moralizing is no special encouragement for the moral skeptic. This, in turn, bolsters realists’ efforts to show that their picture of moral language is compatible with the phenomena mentioned above.

Or so it seems. These familiar thoughts are right as far as they go, but I don’t think they go far enough, because they don’t solve the real problem. They show that some of the classic arguments against realism are insufficient, as stated, to drive home the worries about semantic continuity. At best, they buy the realist some time; they aren’t enough, by themselves, to show how a realistic construal of moral language is compatible with the phenomena that motivate the concern. And extant versions of moral realism have not done enough to develop the resources needed to deal with this problem. If we want a more satisfying answer, we’ll have to do more than rely on these useful but incomplete moves in the philosophy of language.

4.4 The traditional response is not adequate

In order to see why these familiar thoughts are not enough, let’s take a closer look at two details of the realist’s proposal. Once these are on the table, we can turn our attention to a broad worry for semantic externalism that will help to illustrate just why the approach outlined above is unsatisfying.

Realism’s response to the semantic worry must involve two commitments concerning the nature of moral properties. First, the realist must be committed to what
we might call the ‘non-neutrality’ of moral properties. That is, realism holds that the
correct account of rightness will not leave open the question of whether, say,
consequentialism or deontology is closer to the truth.\textsuperscript{162} We can see why this
commitment is required if we keep in mind that moral properties are to serve as the basis
for an explanation of what’s going on in cases of moral dispute. If the non-neutrality
claim were false—for example, if moral properties were suitably coarse-grained, or
disjunctive, or in some other way failed to decide between substantially different
normative theories—the realist would face a troubling puzzle. What’s at stake between
our old friends, the Kantians and the utilitarians?

When they disagree about what’s right, the initial realistic thought is that they’re
making competing, incompatible claims about the (moral) facts of the matter. But, if
moral properties \textit{were} normatively neutral, the realist would be forced to say something
unpalatable. The first option is to say that the dispute is confused, since, on this view, the
properties we seem to be disagreeing about simply underdetermine (or leave open) the
answer; so continued argument is senseless. The second is to ground the dispute on
something other than competing property ascriptions. Either route is disastrous. The first
does so much violence to moral discourse as to render it unrecognizable: our central cases
of moral conflict turn out to be covert instances of agreement, or, at least, confusion. The
second option leaves the realist with the obligation to say just how our interlocutors are at
odds with one another. This augmentation, in turn, threatens to render the original
position otiose, since the addition might well supplant the theory it was intended to save.

\textsuperscript{162} The realist is, of course, free to say that there are vague penumbra and the like, but this move can be
used only so often before it undermines the interest of the realist’s view.
Suppose we invoke some kind of clash in attitude or sentiment to preserve the disagreement: if successful, we might be tempted to try to account for all of moral discourse in these terms, leaving us without the need for our original view. For these reasons, it seems that at least the clear and paradigmatic cases of moral disagreement should be preserved by appeal to realism’s core doctrines. This, in turn, requires an account of moral properties that does not leave them neutral between at least most of our substantive positions. Were they otherwise, they would not provide the basis for an account of disagreement at all.

This brings us to the second relevant feature of natural moral properties: their ubiquity. If these properties are to be the basis for univocal conversation, they must be widely shared across times and across speakers. Moral terms must pick out the same properties in your mouth, mine, and those of our ancestors. To the extent that the realist wants to preserve what appears to be moral disagreement, she’ll have to say that we can have shared reference even between interlocutors (and communities) with wildly different views. We shouldn’t think that all disputes about evaluation are moral in flavor; we might think, instead, that some of these disputes involve predicates that do roughly the same work as our moral terminology but that really do mean different things.¹⁶³ But, insofar as the realist thinks that we have sharp disagreements within moral discourse, either between individuals or between communities, she must think that this variation is compatible with shared reference.

With these two features, non-neutrality and ubiquity, in mind, let’s look quickly at a worry for semantic externalism and then examine how this problem raises special

---

¹⁶³ See Chapter 3, esp. 3.5, and Chapter 2, esp. 2.7, for a more complete development of this point.
problems about the moral realm. According to an externalist view of content, we have the propositional attitudes we do not simply in virtue of ‘what’s in the head’ but, in part, because we stand in the right sorts of relationships with various things in the world. Content is determined partly by states outside of an agent’s body. So, for example, we have water-thoughts partly because of our connections (causal, covariational, or whatever) with actual H2O. Everything would look just the same, from the first-person point of view, if we were hooked up to XYZ, say, but then we’d be thinking about twin-water.\footnote{See Putnam 1975.} The nature of the relevant relationship is controversial, of course, but the general idea is that we have the thoughts we do partly because of relationships beyond our ken.\footnote{According to Georges Rey, Donellan, Kripke, Putnam and Devitt think that an agent has the concept [water] iff she has states that stand in the appropriate historical relation to water; Dretske, Millikan, and Papineau think that an agent has the concept iff water played a functional explanatory role in the presence of certain states or behaviors in the agent; an earlier Dretske, Stampe, and Fodor think that an agent has [water] iff certain states in her covary with water in the right circumstances. See Rey 1992: 316-7.}

It’s this central point that raises a concern, one that’s inspired by the thought that externalism has perhaps paid too much attention to what isn’t in the head and, in doing so, has shortchanged the importance of what is—namely, our conceptions, beliefs, attitudes, and so on. Consider what Georges Rey calls the fortuitous locking problem: if the relationships in virtue of which we have the propositional attitudes we do aren’t accessible to us, we might end up standing in those relations with, or being ‘locked on’ to, the intuitively wrong stuff. That is, we might stand in, say, a covariational relationship with something that doesn’t seem to be the subject matter of our thoughts and conversations. “Insofar as one is moved by standard externalist claims about the reference of [our] terms,” Rey says, “one supposes that they refer to whatever real
phenomen[a] the people in one’s community are ‘getting at’ in their uses of the terms.”

But people don’t seem to be getting at just any phenomenon that happens to be linked up in these ways with their brain states. These fortuitous lockings, Rey complains, “do not give rise to any meaning intuitions, they would not enter into cognitive deliberations, and, most importantly, they would not seem to figure in any cognitive psychological laws” (Rey 1992: 317-8). The worry is that any purely nomological, non-intentional account of reference and content might leave us ‘accidentally’ hooked up to things that aren’t really what we’re referring to and thinking about.

We must seek a safe passage between the dangers on two sides. If we say that reference is fixed by individual speakers’ beliefs, associated descriptions, and so on, then we have the familiar problem of accounting for shared reference in cases of variation in cognitive shape. If, on the other hand, we say that reference is fixed by facts that are completely independent of cognitive shape, then we’re stuck with an implausible and mysterious account of reference and its link to our thoughts and inquiries. The lesson of this second option is simply that what’s in the head must be an important part of the story about content and reference. It’s all well and good to say that variation in belief and theory is compatible with shared reference and meaning; surely there are speakers with incomplete understanding of, say, arthritis, who nonetheless succeed in talking and thinking about it. But we need to say how this is so while avoiding the counterintuitive results that arise out of fortuitous lockings.

The crux of Rey’s worry is the possibility of a troubling disconnect between our thought and talk, as viewed from the inside, and its content or reference, fixed by some nomological relationship beyond our ken. What’s needed, in response, is a way of
leaving the issues of content and reference firmly tied to the workings of our inquiry: our sometimes inchoate theories, our views about how best to revise those initial commitments, our standards of evidence, our norms for deciding whether some new claim elucidates or changes the subject. We need to avoid a theory that leaves us tied to things we don’t see ourselves as getting at. That is, the facts about reference and content can’t be completely divorced from what we do and think: they must manifest themselves somewhere on the surface of the practice, rather than remaining below, hidden and inscrutable. \(^{166}\) While we can acknowledge that interpersonal differences in standards and theories do not prevent shared reference to a common natural property, we also need to say how we can preserve a shared subject matter without relying, in a problematic way, on mysterious extra-cognitive relationships. This isn’t the place to take on the general problem, but we should keep it in mind when we turn our attention back to moral realism. This will help to demonstrate why the traditional realist responses to worries about semantic continuity aren’t sufficient.

Let’s look now at why the moral case might pose some special difficulties. If moral properties are both non-neutral and ubiquitous, some speakers are referring to properties best characterized by theories they find unappealing, even repulsive. This, in conjunction with the permissiveness or open texture of moral discourse, makes it more difficult for the realist to preserve univocity while respecting the force of worries like Rey’s locking problem. Given that moralists are allowed such wide latitude in their

\(^{166}\) Note, too, that Rey (and I) see this problem as essentially compatible with the spirit of externalism. As Rey puts it, externalism “is not really committed to saying that the external relations are sufficient for a state having a certain meaning, only that they are necessary; and so it is free to claim that the right kind of causal route involves not only a tie to the right stuff outside, but by the right route inside as well.” (Rey 1992: 319).
views, and that ties of deference and authority are more problematic in moral discourse than elsewhere, it’s puzzling how clashing interlocutors can be linked to robust moral properties by the sorts of bonds that matter. In Rey’s discussion, remember, the problem stemmed from the fact that various kinds of locking relationships (covariational, causal, and so on) might link speakers to things that, intuitively, they don’t recognize as ‘what they’re getting at.’ The permissiveness of moral discourse will make this problem even more pressing for the naturalistic realist, as we’ll see.

The issue, of course, is whether and how the differences between disputants in moral conversation are compatible with shared reference. Consider Simon Blackburn’s voicing of what I take to be essentially the same complaint:

[W]hy does the existence of a ‘real’ moral property help us to restore sameness of sense to differently functioning participants in a debate? [Disputants] have to be brought into some kind of relationship with any such property. But their different standards and views of the importance of the issues they are discussing makes that just as hard for the realist as it is for anyone else…(Blackburn 1991: 9)

Now, a difference in standards and views doesn’t necessarily prevent the right sort of connections to a shared property, as the realist is quick to point out. But keeping Rey’s worry in mind adds force to the problem, because it shifts our attention to what’s in the head, as well as what isn’t. The relevant connections between speaker and property seem to be ones that the speaker must acknowledge as relevant to ‘what he’s really getting at.’ But this makes the intractability of dispute even more troubling. When it comes to preserving shared meaning, and, by extension, real (dis)agreement, Blackburn thinks,

…the reality of the property is quite idle. In fact, if anything it makes matters worse, since for the realist there may be a hidden right or wrong, depending on the explanation of the subject’s thought…For all I can see, on the realist picture, one person may be misguided in standards and motivations, but ‘locked onto’ the right property; another with the same functioning may be ‘locked onto’ the wrong
one, so though we engage them equally and accuse them both of mistakes about justice, we might be wrong to do so. For the anti-realist there is no such possibility: the interpretation is, as it were, downwind of our interpretive needs. (Blackburn 1991: 11)

In order to share reference, participants in the conversation have to be connected in some way with the same property, and the realist’s opponent is right to demand warrant for attributing reference to that property, rather than some other, to both parties. The speakers’ intractable disagreement makes this challenge harder to meet. As we’ve seen, popular resources can take us only so far—if causal ties drag Kant, kicking and screaming, into the consequentialist fold, this looks like a point against a view of reference that relies on them.167 The deviant moralist finds our views anathema, and so we must say just how he’s tied to our moral properties, appearances to the contrary.

This might look like just another application of the fortuitous locking problem: we’ve got a troubling notion of content if our eccentric moralist is connected to our moral properties by bonds to which he’s indifferent. But there’s a problem here that’s peculiar to normative discourse: tying the deviant moralizer to a (substantive, robust) property against his will seems to be just the sort of linguistic or conceptual fiat that renders moral talk less interesting, and less suitable for performing its distinctive task.

This is where worries about contestability and open texture really hit home, because we have to make sure that the realist’s attempts to preserve univocity respect the special nature of moral discourse. Reflecting on Kripke, Putnam and Burge reveals that we can share a subject matter despite differences in our theories and beliefs. Clearly this

---

167 These concerns are similar to those raised in Chapter 3. The concern is with how the naturalist’s picture of reference will allow for our moral practice to go on as before, in particular with regard to the ways we investigate moral questions. This ‘moral inquiry problem’ requires, more or less, that questions about reference and shared meaning are settled by appeal to features on the surface of the practice, rather than by hidden facts that don’t manifest themselves in the space of reasons.
is right, in morals as elsewhere. But Rey’s fortuitous locking problem underscores that a purely nomological, pre-intentional relationship is not sufficient to connect speakers with ‘what they’re really getting at;’ we need some sort of contact with what’s in the head. This is harder to provide in this instance because of the permissiveness of our moralizing, which ensures that there are fewer points of agreement here than elsewhere. Without these, the realist’s claims about reference look disturbingly coercive: they bind speakers to substantive moral properties by causal ties alone. Extending the metaphor, we might say that, in order to connect speakers with these properties, we must seek the consent (or, better, the license) given by speakers’ conceptions, beliefs, norms, and attitudes—that is, things within their cognitive ken. In other words, we must make a case for saying that these are the properties they’re getting at. Rey’s demand concerns meaning generally, but the special character of moralizing makes this instance of the demand particularly hard to meet.

In light of our earlier discussion, we can see that part of the solution has to be in terms of what we might call ‘the surface of the practice:’ that is, the answer must be couched in terms of moralists’ conceptions, commitments, norms of reasoning, and so on. We must rely on what’s in the head, rather than appeal to some accidental causal contact with the property of rightness. This is the only way to make good sense of the phenomena of moral discourse in a manner that’s compatible with the realist’s view.

\footnote{Note that we don’t need to say that all apparently genuine moral disputes are univocal; it might be that some of our intuitions should be revised because the parties either are talking past one another, or because they’re arguing not about rightness but about something else, namely, whether rightness or some other property should play a particularly important role in our practical reasoning. As I’ve suggested elsewhere, the expressivist has loaded the dice here by assuming a broad and permissive construal of the boundaries of moral discourse and disagreement.}
4.5 The convergence claim; and how it helps

The realist needs to connect diverse speakers to moral properties that some of them find appalling, despite their protestations that this isn’t what they’re getting at. What’s more, she must do this in a way that respects the constraints we’ve just discussed. The realist’s best response to this demand, I think, involves a notion that he’s apt to like for other reasons: the idea that moral inquiry will converge under suitably idealized conditions. The claim is that some interesting and important range of moral disagreement will be resolved under conditions of corrected reasoning (and perhaps other capacities involved in moral investigation, such as emotional sensitivity, sympathy, imagination, and so on) and full information about the facts. Call this the convergence claim, and suppose for a moment that it’s plausible.

The convergence claim is one that fits naturally into the realist’s view of moral discourse. Realists, after all, tend to reject the idea that participants in our moral discourse are radically estranged from one another, sharing at best a meager store of platitudes that are insufficient to lead them to robust and interesting agreement on contentious questions. They tend to think, instead, that the appearance of intractability comes not from some failure of truth-aptness or some strange open texture, but from distorting factors such as ideology, self-interest, the shadows of misguided theistic metaphysics, and so on. Remove these, they often suppose, and many contentious issues would be resolved. As Nicholas Sturgeon has remarked, realism is an optimist’s position.

\footnote{Full awareness of which facts? Well, to say the moral facts would be to make the claim trivial. But the naturalistic realist usually says that the moral facts are either identical with uncontroversially non-moral facts or are constituted by them. So the best thing to say is that full information involves all of the facts, described in non-moral terminology, or something like this.}
The convergence claim might also be a part of the solution to the semantic worries under consideration here. Suppose we thought that the convergence claim is true in virtue of shared commitments and norms of reasoning on the part of interlocutors who otherwise appear to be at loggerheads. That is, the theoretical agreement that results at the end of the day, we’re supposing, is produced through revisions in both parties’ views that are reasonable by the disputants’ own lights. Despite their clashing commitments, speakers would come to agreement if they deliberated correctly, because the only things standing in the way of this resolution are impediments that can be removed by exposure to better information and (what can be fairly established as) better reasoning. If so, it becomes much more plausible to say that these speakers share terms and are hooked up, in a way that matters, to the same set of moral properties. We no longer need to worry that univocity is preserved by appeal to some feature that’s invisible from the surface of moral discourse: our explanation of shared meaning lives in the space of reasons, not causes. Or, to put things more clearly, invoking the convergence claim allows us to connect the give-and-take of our actual moral discussion with worries about agreement in meaning, which is as it should be. If idealized, but still recognizable, investigation would lead two speakers to theoretical agreement, we have grounds for saying that they really were getting at the same moral properties all along.

The realist’s appeal to the convergence claim is part of a sensible response to a serious problem. The puzzling disconnect between the real facts of reference and the workings of our moral conversations, the problem which rightly disturbs Blackburn, can be eliminated if these facts are tied to the procedures and outcomes of our investigations. There’s no gap, because we’ve forged an intimate connection between the semantic
question and what goes on within the discourse. At the same time, the realist can make room for the needed variations in belief and theory between speakers, because what’s required is not uniformity in moral commitment, narrowly construed, but agreement in those commitments, whatever they are, that are sufficient to produce agreement in the right sort of idealized circumstances.\textsuperscript{170} This uniformity is consistent with a wide degree of substantive variation.\textsuperscript{171} The dilemma we considered earlier was more or less this: how can we account for the obvious differences between speakers while avoiding appeal to the extra-cognitive relations that fuel the fortuitous locking problem? The answer suggested by the convergence claim is that we need to consider broader notions of what’s in the head, so that we include not only first-order moral commitment but, for example, higher-order epistemic norms as well. We can then argue that these broad points of agreement between speakers committed to different normative theories are enough to justify the claim that they share meanings. This gives us an opportunity to preserve univocity while respecting the constraints revealed by our earlier discussion.

The appeal to convergence, and to the features of interlocutors responsible for its occurrence, is, I think, the move that must be made in response to the worries we’ve been considering. Consider the dialectical situation if moral conversation \textit{didn’t} converge under idealized conditions. Then the realist is faced with the task of explaining how speakers are referring to properties that they \textit{don’t} see as ‘what they’re getting at’ even at

\textsuperscript{170} The class of commitments that are relevant here might be quite expansive. I have in mind things like higher-order attitudes concerning epistemic revision, commitments to interpret fellow speakers as sharing moral terminology up to the point of radical eccentricity, and so on.

\textsuperscript{171} Compare this with a case of scientific investigation, such as widespread disagreement about, say, atoms, which included purportedly definitional claims that would later be rejected. Why say that earlier speakers shared the concept? In part, the justification for thinking this has to do with the idea that their conceptions are revisable, step by step, in a way that relies on facts they admit to be relevant, until the conceptions are more or less correct.
the end of the investigative day. If the speakers in our optimal moral conversation don’t recognize these properties as their own, and we continue to insist that the properties are nonetheless the referents of shared moral terms, we’ve left ourselves open to the fortuitous locking problem and related considerations.\textsuperscript{172} The realist’s insistence on shared reference looks groundless, and he’s left without a compelling response to the univocity problem. Hence, the convergence claim is one the realist must make if he’s to meet the challenges we’ve been considering.

Another way of putting this is to note how the convergence claim allows the realist to respond to our earlier worries about coercion and consent. A speaker who’s tied to rightness by, say, causal connections, might reject the relevance of those ties to her deliberations about what to value and how to live. If the convergence claim is true, then the recalcitrant speaker is, at some level, implicitly committed to the (ex hypothesi) correct end-of-the-day theory. The only authority to which the realist need appeal, then, is the authority conferred by the speaker herself. This is important insofar as we think that moral discourse ought to remain free of the sorts of authority and deference relations involved in straightforward empirical inquiry.

(Though this is a bit removed from our main worries, it’s interesting to note that the convergence claim also provides at least part of a response to our concerns about the realist’s ability to provide an appropriately robust connection between moral judgment and action. The root of the problem, recall, is that it’s by no means certain that participants in moral discourse will be suitably committed to giving the realist’s natural

\textsuperscript{172} This assumes that the realist wants to maintain that speakers who disagree in the limit of inquiry are still sharing terms. Another option is to deny this and account for their dispute in some other way, as explored in Chapter 3.
moral properties—whatever these turn out to be—the right sort of place in their practical reasoning. “Why guide our lives by these properties, rather than some others?” they might ask, and understandably so. Once we have the convergence claim at our disposal, though, we can reply to this by pointing out just where these properties came from: they arose out of our careful and sustained reasoning into pressing practical questions. It’s important to keep in mind that the convergence claim concerns what would happen under idealized circumstances that we take seriously; that is, its idealizations are ones that we think have some degree of normative authority. And part of what guides that hypothetical inquiry is the notion that the properties we’re after have to play a certain role in practical reasoning.\footnote{Compare this to the moral functionalist strategy—i.e., build in ties to action as part of the role description that identifies the (descriptive) moral properties. See Jackson and Pettit (1995).}

4.6 Agreement on cases and on theories

I’ve argued that realism faces a worry about the univocity of moral conversation that cannot be defused by easy appeals to, say, causal views of reference and externalist theories of content. If our discussion is on the right track, the realist must make some strong claims about the resolution of moral dispute in order to preserve genuine moral disagreement. Putting aside, for the time being, worries about different sorts of convergence claims, we should look into a broader question: how much confidence should we have in the idea of end-of-the-day agreement? This question must be answered before we turn to the more subtle issues concerning the various ways in which convergence might occur. Realist discussions of this general issue about convergence (independent of the semantic worry) have often faltered because they have paid
insufficient attention to the distinction between disagreements about hard (or at least contentious) cases and more abstruse theoretical disagreements. If we pay too little attention to this distinction, we’re likely to focus our attention away from the problems that are both more serious and more important. In effect, we’re looking where the light is good.

To see the problem, consider some of the moral issues that are frequently invoked in discussions of convergence: slavery, abortion, the political role of women, perhaps gay and lesbian rights. These are often cited as problems that have served as the focal point for intense, seemingly intractable dispute despite the plausibility of saying that they admit of reasonable resolution. In discussing them, the realist attempts to raise our hopes about the prospects for convergence in moral opinion by pointing out, in effect, that past or present controversy is no guide to resolvability.

This is true, as far as it goes, but I think the realist is taking an unfair advantage by invoking disputes that are located primarily in our popular discussions. Many of these conversations—now, in particular, over the abortion debate—are hobbled by bad arguments, ignorance of relevant empirical facts, and the sorts of ideological and metaphysical baggage that pervade (and undermine) our everyday moralizing. These disputes look (or have looked) intractable in an important sense: there’s just not much hope for near-universal agreement on the question of abortion’s moral status. But they also appear to be eminently resolvable under idealized conditions and full information. If opposition to abortion is grounded, in practice, on religious conviction, and if questions about the existence, nature, and will of a divine being are themselves subject to reasonable scrutiny, then abortion begins to look like a topic that’s actually quite friendly
to the realist. These issues, after all, may be difficult, but they aren’t questions that tempt
us to say that theological discourse is not truth-apt or that its concepts are contestable in
some special way.\textsuperscript{174} All the sound and fury that suggests intractability (at least to
undergraduates) vanishes on closer inspection, and we see that convergence is quite
possible under the circumstances that we tend to think are relevant.\textsuperscript{175} Similar points can
be made for many issues that remain seriously controversial in our everyday moralizing.

But this, by itself, is not of much help to the realist. The problem is that the
wrongness of slavery (or the political status of women, or, perhaps, even the
permissibility of abortion) is overdetermined: chattel slavery is an abomination whether
you think that moral rightness has everything to do with the commands of pure reason,
utility functions, divine prescription, or the well-tuned soul. This explains how we might
agree on particular judgments—“slavery is wrong!”—without having any robust
agreement at the theoretical level about just what makes those judgments true.\textsuperscript{176} This is
both a part of the explanation of why the issue is resolved (since, after all, we need not
come to terms with differences in theory to see that the institution is indefensible) and an

\textsuperscript{174} We might take a sort of noncognitivist view about theological commitments and disputes but this isn’t the most plausible option, and, in fact, it seems a good deal less appealing than noncognitivism about ethics.

\textsuperscript{175} Nick Sturgeon has argued that common responses to moral disagreement (‘it’s not worth talking about this issue because…’) actually favor the realist, because speakers often explain these disagreements by appeal to features (e.g., emotional commitments) that, we agree, skew inquiry.

\textsuperscript{176} For example: the pivotal moments in the history of American abolitionism were provided not by theoretical discussions but by simple presentations of the facts of the case, one fictional (\textit{Uncle Tom’s Cabin}), the other consisting of carefully documented testimonials about various respectable people’s experiences of the institution (\textit{American Slavery as it Is}). The tremendous influence of these books suggests that the general public need not have any particular moral view—or anything more than the inchoate commitments of most everyday moralists—in order to see the moral indefensibility of slavery. And actual slaveowners had powerful economic incentives to preserve the institution, to the extent that we might plausibly construe them as infected by considerations of their own interests rather than as good-faith adherents to a reasonable moral view.
illustration of why it’s misleading to focus our attention on the resolution of disputes like these when thinking about the realist’s semantic concerns.

Recall that our concern here is with the threat of equivocation and that to counter it we need to show, in a way that satisfies the worries we’ve canvassed, that our moral terms are connected to the same properties despite our differences in theory. *We haven’t* done that if we’ve made a case for convergence on particular issues without arguing that this agreement is related to the subtler and more important issue of theoretical dispute between different moralists. If we can (ostensibly) agree on slavery while being committed to quite different moral theories (theories that, of course, give different and incompatible accounts of the properties to which we refer), then our agreement, by itself, isn’t a point in favor of shared reference. (Note again the importance of the point that the realist’s view of moral properties has it that the properties aren’t neutral between competing normative theories.)177 Our conversation might involve reference to properties that are distinct though roughly coextensional over some interesting range of cases in the actual world. If so, then the realist’s theoretical apparatus doesn’t really preserve univocal conversation after all.

The problem, then, is that convergence in controversial cases of the past and present seems to rest on simply clearing up ignorance and confusion about the nonmoral facts and removing irrationality, blind self-interest, and other distortions from the equation. This, of course, is extraordinarily difficult as a practical matter, but it’s not very complicated in theory, simply because the disagreements, in many cases, clearly

---

177 At least, the properties are not neutral between significantly different theories such as consequentialism and deontology. Maybe the realist will end up saying that the correct account of rightness fails to distinguish between, say, variants of consequentialism, but this is a much less serious weakening of the non-neutrality claim.
depend on non-moral mistakes of knowledge or reasoning. On the other hand, agreement about, say, slavery and abortion is not immediately and obviously connected to the central semantic question, namely, how disparate moralists can be hooked up with the same set of properties despite the variations in their moral views. Because agreement on particular cases is perfectly compatible with (intractable) disagreement in theory, convergence of opinion about these cases is not much evidence for the sort of agreement that the realist needs, namely, theoretical agreement. That, or something like it, is needed because the realist has to connect individual moralists to (shared) moral properties in some way that respects the practice of moral discussion and inquiry.

4.7 Responses to recalcitrant theorists

The realist needs theoretical convergence, not just agreement on problem cases, because only the former will help to resolve the semantic worry we’ve been considering. But theoretical convergence is a hard sell given the conjunction of familiar worries about intractable disagreement and our concerns about the fortuitous locking problem. We might admit that a variety of controversial issues are resolvable and still deny that such resolution is possible in the case of competing theories. For one thing, the arguments unfold differently here: participants in these conversations are better informed and more dialectically sophisticated, they don’t seem to be making simple mistakes, and they understand the ramifications of their positions. What’s more, much of the debate concerns hypothetical cases in which the relevant non-moral facts are fixed, so the

\[178\] It’s natural that opponents of realism should focus on the popular disputes about issues, because the disagreement is so obvious and unavoidable. And realists, as I’ve suggested, may find, in these cases, easy fodder for optimism about idealized inquiry along with grounds for pessimism about the actual world.
correction of factual errors won’t move us much closer to agreement.\textsuperscript{179} As I noted above, it’s easier to be optimistic about the resolution of (parts of) popular discussion \textit{because} of its flaws, which are responsible for some of the appearance of intractability. By contrast, the sophistication of the current theoretical debate makes the convergence case harder to make, because there’s less room for clearly mandated revision.

Let’s suppose, just for the sake of argument, that some variety of consequentialism is the correct moral theory.\textsuperscript{180} This makes the problem more vivid, because we can now imagine real philosophers who think that this view of rightness is deeply confused. Yet it nonetheless seems clear that they’re engaged in debates \textit{about rightness}. Just how are John Rawls and Barbara Herman connected to these moral properties, as we’re assuming they’re to be understood? The realist, if she is to preserve the appearance of substantive disagreement, must find a satisfying way of linking moralists like these idealized Kantians to consequentialist moral properties. And, of course, the connection must be one that respects the constraints we’ve been discussing. This is simply a special case of the semantic puzzle, of course, but it’s a problematic one, because it puts pressure on a useful and familiar realist thought, namely, the idea that appropriate revisions to our error-riddled moral thinking will produce agreement. Given the dialectical virtuosity, and the intransigence, of our interlocutors, it’s just less plausible to say that we’re implicitly committed to the same theory.

In this section, I’ll survey three different responses to the general problem of recalcitrant theorists, which I’ll call the ambitious, concessive, and reforming strategies.

\textsuperscript{179} I owe this point to Justin D’Arms.
\textsuperscript{180} Though, as I discuss below, this isn’t just an assumption for many naturalistic realists.
These three give different answers to the question of whether and how the realist can keep our recalcitrant theorists within the semantic boundaries of moral conversation. Seeing how this discussion unfolds will, in turn, help us to see the importance of convergence, to which we’ll return in the next few sections. The ambitious strategy gives the most obvious answer: it argues that these thinkers—or any idealized reasoners with a strong commitment to a view diametrically opposed to the one that, we’re assuming, is the correct theory—are implicitly committed to the right view, despite appearances. If the realist takes this position, she must argue that a failure to come around to our way of seeing things represents some kind of failure of inquiry, such as irrationality, insensitivity, or whatever. This may seem hopelessly naïve, or untenably arrogant, because the strategy seems to pay inadequate respect to the moral disagreement we’re trying to understand. After all, these are idealized Kantians, and they seem to have about as good a case for their view as the consequentialists have for theirs; so suggesting this thought about implicit commitment may look overly dismissive. Later on I’ll suggest why the ambitious strategy isn’t as unpalatable as it appears, but before turning to this and to the other strategies, we should look at why something like the ambitious response is necessary.

We might suspect that the optimism of this view is simply unsustainable (because, say, we have a hard time imagining our idealized Kantians switching sides in the normative debate) while thinking that the problematic speakers are still connected to our moral properties by some familiar mechanisms. Perhaps these speakers share reference

---

181 In other words, the realist is committed to saying that moral discourse exhibits what Crispin Wright calls cognitive command. See (1992) ch. 3.
with the rest of us because of their ties to others, as when medical neophytes manage to refer to arthritis because of their location in a larger community. Their own theories fall short, but, luckily, they’re positioned so that others can pick up the slack. The so-called ‘division of linguistic labor’ might prove useful here as well.\footnote{This phrase comes from Putnam (1975).}

On reflection, though, it’s hard to see how this initially appealing thought can be sustained in light of our earlier discussions. The appeal to the idealized Kantians’ relationships with other speakers (whether these are deference relations or some other sort of connection) is a move that, at best, buys a little time. If those ties elicit the endorsement or approval of our recalcitrant speakers, then it’s hard to see how we’d have a case of intractable disagreement. (For example, if the speakers were committed to deferring, then their deference would produce agreement). If the ties were viewed as irrelevant to what the speakers meant to be getting at, then we’d be back in the puzzle that rightly disturbs Blackburn and Rey: how can a reference relation coerce speakers into substantive views against their will? The general point here is that appeals to referential intentions, causal chains, and the like are stop-gap measures that must be replaced with something more substantial at the end of the theoretical day. So it looks as though the obvious alternative to the realist’s most ambitious option won’t pan out.

Before moving on to the other strategies, there’s one more point we should consider. This is a distinction concerning just what our recalcitrant theorists might agree on at the end of the day. It might be, of course, that they’d find our arguments so compelling that they’d be committed consequentialists, through and through; they would then think that this is the correct story about rightness \textit{and} that rightness is the property
that ought to play an important role in their practical deliberations. On the other hand, they might give us the semantic claim about “right”—it really does refer to utility maximization, they’d say—and then go on to question the claim that it’s rightness that should hold pride of place in our thinking about what to do. That is, they might give up on the term and then insist that the property they’ve had in mind all along (k-rightness?) is the one we should be interested in. These two options are importantly distinct: the second, but not the first, is an external criticism of morality as a whole. In addition, keeping this distinction in mind will help us to evaluate the prospects for the realist’s claims about agreement.

Let’s turn now to the concessive strategy, which grants that our blamelessly recalcitrant speakers aren’t referring to the moral properties we refer to when we’re all engaged in what appears to be moral disagreement. We might say, instead, that some of our conversations revolving around a term like “right” are equivocal. By taking this view, the realist is, in effect, giving up on some lofty ambitions, because she’s no longer committed to keeping just about everyone within moral discourse. But this isn’t the end of the line for realism. There are two important points to make in favor of this general approach.

The first is that we have a plausible story to tell about the appearance of real disagreement in cases like this: namely, one invoking the old expressivist idea of clashing attitudes or conative states rather than beliefs. We have conflict about what to do, about whether to put consequentialist or deontological properties at center stage in our practical

---

183 I owe this point to Justin D’Arms.
reasoning, but not conflict about what’s right.\textsuperscript{184} The second is that the admission of equivocation is reserved for just those occasions when our actual moral conversations make it most appealing. We’re assuming, when we make this response, that interpreting our conversations as univocal leads us to intractable disagreement without any obvious failure of rationality or moral sensitivity in either party.\textsuperscript{185} It’s just this outcome that makes the charge of equivocation plausible, since one explanation of how such disagreement is possible is that the parties involved are talking about different things. After all, they make opposing judgments by invoking different standards; they see the others’ reasoning as missing the point, and so on. Given this impasse, the interpretive move I’ve suggested is not at all implausible.

It’s important to see that the concessive strategy can be developed in a number of different ways, depending on just who we want to leave in or out of the conversation. Given that we’re faced with intractable disagreement between theorists, perhaps it’s most appealing to say there’s something suspicious about our idealized Kantians in \textit{virtue} of their status as fairly refined thinkers. Less sophisticated, less refined speakers, on the other hand, might be in good semantic standing because of the nature of their commitments, intuitions, and so on.

For example, we might think that the appearance of intractable disagreement at the theoretical level is explained by noticing that, say, both consequentialists and deontologists seize on different features of our everyday moral practice, and use these features as the basis for a unified, coherent theory. The problem is that our moralizing

\textsuperscript{184} This is discussed in chapter 3.
\textsuperscript{185} Just what isn’t breaking down is a little bit tricky; we might think the ‘do no injustice though the heavens fall’ Kantian is guilty of a sort of insensitivity to the demands of the situation—to the position of those who could be saved by a single rights violation, for example.
appears to be confused: it’s a mix of all sorts of intuitions that don’t seem to cohere well with one another. Thus, it’s possible to seize on one or other set of intuitions, develop these into a more coherent alternative, and claim that you’ve got it right about what we’ve been doing all along. Unfortunately, others can end up with strikingly different theories by prioritizing different sets of intuitions. Our problem arises from the observation that both camps seem to have more or less equally strong claims to capturing the core thoughts of our commonsense morality.

One way of developing the concessive strategy is to argue that this last appearance is misleading: in fact, one party has focused on a marginal cluster of phenomena that don’t get to the core of what we’re doing. In refining these intuitions, they’ve simply drifted so far from the mainstream of moral thought as to have ended up doing something else altogether. In other words, the problem with our idealized Kantians is that they’re working with the wrong raw materials; they see as essential what’s best explained or revised away (we add, snidely). Now, most everyday moralists haven’t refined their views in this way; they still have intuitions that strain toward a variety of incompatible theories. But our Kantians may have taken themselves out of moral discourse by centering their views on what are really marginal or ancillary data-points. Perhaps it’s best to see them as urging us to adopt their properties of choice to serve an important practical role, rather than as saying that they capture what we’ve been talking about all along.

---

186 See for example Shelly Kagan (1989) for arguments to the effect that our ordinary moral thought is confused and needs serious revision. See also Peter Unger (1996).
187 And, presumably, their commitments are such as to connect them in the right way with the right theory and hence the right moral properties.
This suggestion gives us a way to say how it is that certain speakers (in this case, our idealized Kantians) are outside the realist’s semantic reach, while, at the same time, maintaining that the bulk of moral disagreement can still be preserved on the realist’s picture. The advantage of this concession is clear: we no longer have to say that these recalcitrant speakers are tied to the same properties as the rest of us. At the same time, we don’t give up on disagreement altogether. There is, however, an interesting problem here. Suppose that our suggested explanation is correct: the idealized Kantians, unlike the everyday moralists who inspire them, are not referring to our shared moral properties, because they’ve made some initial errors in identifying the core intuitions of our practice. This suggestion relies heavily on the idea that there’s an error involved: the Kantians have seized on marginal intuitions, they’ve mistaken accidental features of our moralizing for its core, and so on. What’s more, it’s important to couch the suggestion in this language, because, without it, we’d be allowing them to say, in response, that their theoretical refinement is as good (as defensible) as ours, and that lands us back where we started. But then we’re committed to accusing our interlocutors of a serious error: we’re saying that their initial gloss on moral discourse is flawed. This is problematic, because it looks like just the kind of mudslinging we wanted to avoid. In order to vindicate the charge of error, we need to give some kind of appropriately neutral demonstration that our reading of moral discourse is better than theirs. Unless we can ground that accusation, we’re just begging the question against our imagined opponents.

Furthermore, if we’re serious about charging the Kantians with error, then we’ll also have to say that the cognitively blameless among them can be brought to agree to our complaint. Essentially, this is to say that a really idealized Kantian could be brought into
agreement with us, and so we’re back at the realist’s optimistic strategy—though, of course, this leaves open whether the erstwhile Kantian agrees with us only about the semantic question, or also assents to our claims about motivations and reasons. Recall that our motivation in discussing the concessive strategy is to respond to the apparent phenomena of perceptive, dialectically virtuosic interlocutors who find our ex hypothesi correct moral theory to be anathema. We might be leery of saying that these speakers are implicitly committed to the right view, for the reasons glossed earlier. But the alternative is, it seems, to charge them with a kind of cognitive failure. In explaining how it is they found themselves outside of the semantic mainstream of moral conversation, we’re suggesting that they made errors in fixing on the correct core of intuitions and commitments at the heart of our moral practice; then their careful inquiries after this error just led them further astray. If this is true, we can see why they no longer have the sorts of intuitions that the ordinary speaker has, that is, the intuitions that will allow that speaker to be led, by his own lights, to what we’re supposing is the truth of consequentialism. But this doesn’t explain why the idealized Kantians can’t see the original error, viz. the mistake of selecting the wrong body of intuitions for refinement.

If there is some fact of the matter about which intuitions serve this focal role, then blameless interlocutors should be able to be brought to see it. If there is no such error, then we have a difference in temperament, and nothing more.

The concessive strategy, as we’ve imagined it, agrees with the ambitious strategy insofar as both view everyday moralists are being within the realist’s semantic grasp. It then moderates its optimism by conceding that our Kantian theorists are beyond that grasp. We’ve taken issue with the second half of this approach, but it’s important to see
that the first part of the project is also open to skeptical worries. We might think that the range of opinions found in ordinary moral talk is simply too expansive to allow for a univocal treatment, as the realist has been suggesting. After all, our moralizing contains, and has contained, over its history, a vast diversity of judgments, principles, commitments, and so on. The realist says, or is tempted to say, that speakers from all the dark corners of our moralizing are referring to a shared set of properties, but we might suspect that there’s a limit to how much error can be tolerated consistent with shared reference. Furthermore, we might think, the ordinary speaker’s commitments are simply insufficient to warrant claiming that he’s been referring to consequentialist properties all along. If we’re worried by these issues, then we might well think that the realist’s concessive strategy is still too optimistic, because, in trying to preserve the univocity of ordinary moral talk, it casts too wide a semantic net.

These concerns motivate the reforming strategy. According to this line of thought, our ordinary moral talk is simply confused: it’s full of competing judgments and principles, it’s littered with the artifacts of suspicious theistic metaphysics, it’s too susceptible to the influence of psychologically salient details. There’s so much confusion, in fact, that it doesn’t make sense to identify one coherent view as what we’re getting at, when we engage in this talk; getting rid of the errors gets rid of most of the practice.\footnote{Compare this to Mackie’s reforming proposals in Mackie (1977).} What’s needed is a reform—we should abandon our confused way of doing things and go over to something new, which does the same practical work while avoiding the problems we have now. (Such a reform might be intended as semantically discontinuous with all of our past practice; on this reading, the reform simply introduces
terms that take the place of our moral language. Or it might be intended as a reform that’s semantically continuous with some of our past uses—for example, the reformer might think that the regimented language he proposes is of a piece with a particular bit of our current moralizing.)

Someone taking this view would then be faced with the task of coming up with a particular reform—what exactly would the reform look like? And why favor that particular replacement? In answering this question, the reformer faces a danger, because, if he offers the wrong sort of reasons, his account threatens to collapse into a version of the ambitious or concessive strategies. Suppose, by way of illustration, that he supports his proposal by suggesting that it fits neatly with our current way of doing things; that is, it lets us go on as before in ways that we think are central and important. Then, we might wonder, why don’t we see the substance of the proposed reform as what we’ve been getting at all along?189 If the reform is too close to what we’re doing now, we’ve simply ended up back at the ambitious or concessive strategies. Similarly, the reform can’t be supported by considerations that would be compelling reasons, within moral discourse, to revise our current practice so that it agrees with the reformer’s favored replacement. This suggests that the choice of reform must be guided by ‘extra-moral’ considerations if the reforming strategy is to avoid collapsing into the ambitious or concessive alternatives.

189 This discussion owes a debt to J.L. Mackie and his critics. Mackie, of course, argues that moral thought is rife with error, because of its false metaphysical presumptions, but he then goes on to suggest that we can still engage in something very much like our current moralizing in a metaphysically respectable way. One plausible response to this—one suggested by Simon Blackburn (1993b) and Crispin Wright (1988)—is to wonder why we don’t interpret our current moral practice in a way that lets at least some of its claims come out true.

As an aside, it’s interesting to note that error theory also faces a problem accounting for the univocity of moral conversation.
We’ve been considering a serious problem that’s based on longstanding, seemingly intractable debate between sophisticated interlocutors. The realist’s response to the semantic puzzle seems to require strong claims about theoretical convergence, but these claims are hard to make in light of this sort of disagreement. The ambitious way to respond to this problem is to say that at least one party in these disputes is in error, by her own lights; her commitments are such that she can be led to the right view. (I’ve been assuming, just for the sake of illustration, that consequentialism is correct, but this isn’t crucial). This is a bold claim, because it involves the idea that we can (or should be able to) talk our smart Kantian friends, say, into a moral view that they find seriously objectionable. The less ambitious approaches, though, run into serious trouble, as I’ve argued here. Judging our arguments to be equivocal might still involve an error-attribution, as the concession strategy demonstrates. And the reforming suggestion seems more like a change of subject than a response to our original worry. So, for better or for worse, the ambitious answer seems to be one the realist has to give. Let’s look now at how plausible an answer that is.

4.8 Ambitious realism: arguments inside and outside moral theory

The realist’s ambitious response is to argue that speakers who appear to be, for example, paradigms of Kantian commitment are ripe for conversion. By this, I mean that their moral beliefs and attitudes provide a toehold for the possibility of rational theory-change. What’s really at stake here is whether or not the disputes between, say, the Kantian non-naturalist and the naturalistic realist are resolvable in a reasonable and fair way. If they are, then there are reasons sufficient to change the mind of a sufficiently fair-minded interlocutor. Showing that there are such reasons is difficult, because, of
course, deontological and consequentialist views pull in exactly opposite directions when it comes to a variety of important moral issues—not only particular cases, but with regard to justifications and broader metaphysical questions as well. But suppose that we did want to contend, first, that even recalcitrant theorists such as the idealized Kantians are talking our moral talk, and second, that this semantic claim requires them to candidates for conversion in the sense we’ve just specified. This is, of course, a very strong claim, because it suggests that failures to come to theoretical agreement, even on the part of sophisticated thinkers, can be attributed to errors of one form or other.

There are two broad strategies we might invoke in support of this strong claim. First, we could deny that views often thought to be at loggerheads, such as consequentialism and deontology, are genuinely counterbalances within our moral theorizing. Making this case involves arguing for some particular normative theory using distinctly moral considerations. Second, we might rely on extra-moral considerations to provide an advantage to one side or other. Here, the motivating idea is that one view or other fits best with holistic considerations concerning our other (non-moral) theoretical commitments. We need to examine each in order to assess their plausibility.

Let’s start with the second tactic. (I won’t develop this theme fully, but it will be useful to give it a rough outline for the sake of discussion). It’s not entirely by accident that all of the prominent naturalistic realists are also committed to consequentialism: Brink, Boyd, Railton, and Sturgeon not only agree that moral properties are (explanatorily potent) naturalistic properties, but they agree, more or less, on the nature of
those properties and hence on the shape of the correct moral theory. One story about why this is so is fueled not so much by realism as by naturalism. Suppose metaphysical respectability is your life’s ambition, and you think of the entities of natural science as guides in thinking about what that means. You might then be drawn to Humean theories of practical reasoning and agency, because such views rest on simple and psychologically respectable mental states. Desires don’t seem to be at all spooky; they fit neatly into our understanding of the mind. And, for the same methodological reasons, you might think that the best account of human good and flourishing will be given in purely natural terms. You get the idea. The point of all this is that the realist’s commitment to naturalistic strictures shapes the sorts of positions available both in and outside moral theorizing. The naturalist has things to say about what reasons are and how moral agency works, and those commitments influence the dialectical terrain. In particular, they influence choices in the realm of normative theory, by, for example, adding to the costs of holding onto a view that requires non-natural entities.

This is not to say that naturalism is the right commitment to make, and it’s not to say that naturalism entails consequentialism; for now I’m noting only that debates within normative ethics don’t take place in isolation. Instead, they’re located against a backdrop of other philosophical puzzles. In this case, these are particular questions about practical rationality and its relationship with desire, as well as broader issues involving naturalistic methodological commitments. This, in turn, gives us a new source of data, since our best

---

190 We might also explain some of this agreement by other means: perhaps there are contaminants in Cayuga’s waters. And I don’t mean to suggest that consequentialism is a dividing point between realists and their opponents, since Blackburn also favors it. If there’s a clear divide correlated with the consequentialism issue, it’s the one between naturalists and non-naturalists.

191 See for example Railton (1986a).

192 Such as reasons, for example. See Scanlon (1998) ch. 1.
moral theory needs to cohere with both our moral intuitions and with the rest of our best judgments about the world. It does us no good to make sense of our moralizing in a way that strains our other (nonmoral) theoretical commitments. Thus, if we want to argue that we face a genuine impasse in moral theory, it’s not enough to say that, for example, consequentialist and deontological moral theories are equally adept at making sense of our commonsense moral judgments and intuitions. In order to make the pessimistic case, we’d need to go further by arguing that these extra-moral considerations are also neutral between the theories.

Now let’s look at how this helps the naturalistic realist, who’s trying to show that ‘deviant’ Kantians are referring to the same consequentialist moral properties as the rest of us, simply in virtue of their own commitments, broadly construed. Since we’re supposing that our deviant moralists are ideally reasonable, we need considerations to boost our confidence in the idea that the balance of reasons favors our view. In our narrowly moral reasoning, we often try to make the case for a particular theory by showing that it coheres well with some central intuitions and clashes only with judgments that are susceptible to revision. The same procedure can be applied here. What’s needed, on this view, is support for the realist’s broadly naturalistic commitments, because this will tip the balance in favor of, say, a consequentialist view over one that is less metaphysically austere. Once this is provided, the realist will have a way of arguing

193 Compare Mackie’s arguments for error theory: Mackie needs to establish both (a) that our moral discourse presupposes the existence of objectively prescriptive moral properties and (b) that there are no such properties. The success of the argument for (b) seems to put more pressure on (a), since there are other, more charitable readings of our moralizing. See Blackburn (1993b) and Wright (1988). On the general point that our inquiry must be global, and not confined to considerations within moral discourse itself, see David Brink (1989).

194 See Brink (1989) for discussion of coherence-maximizing reasoning. My point is simply that what needs to be made coherent are both moral and non-moral commitments.
that the balance of considerations (not just within moralizing, but globally) supports our view over that of our deviant colleagues. (We might think of this as a natural extension of the maxim that ‘only a theory beats a theory.’) What’s needed is a well-worked-out and comprehensive theoretical framework for making sense of our moral commitments and their place beside our other commitments, for example, about what kinds of things exist, about the nature of intentions, and about what sorts of considerations provide reasons for acting. The conflict between the Millian naturalists and the Kantian non-naturalists, then, should be understood not only as a conflict between two ways of grounding our moral commitments, but also as a disagreement about how to approach our moralizing and where to locate it.

The naturalistic realist might also embark on the other half of this approach to preserving univocity. Now that we’ve suggested a broader theoretical bedrock on which to ground our arguments with opposing moralists, we can also begin to work within moral theory to increase the attractiveness of the naturalistic realist’s position. For example, we might take on familiar deontological objections to, say, sacrificing individual well-being for the sake of overall utility. We might argue that this rests on misguided intuitions that stem from contexts where we’re technologically unable to sacrifice one to save five, and where the sacrifice, but not the benefit, are vividly imaginable. Increased sophistication in our deliberations, and changes in our scientific...

---

195 I should note that these remarks about the content of our normative inquiries are strictly provisional: it might be revealed, in the fullness of time, that the consequentialism to which various naturalistic realists are committed is not the right moral theory. On one hand, this is all well and good, since the realist per se is committed only to particular story about the reality of moral properties and the reference of moral terms. On the other hand, this might be troubling, depending on what the end-of-the-day story turns out to be. If the content of our moral inquiry turns out to favor a view that’s incompatible with naturalism, then, it seems, something’s got to give (as long as the considerations that favor methodological naturalism are insufficient to have some effect on the content of moral inquiry).
prowess, may undermine this commitment, as we’re made more vividly aware of the situation of the five, as well as the one, and we’re increasingly confident of our epistemic position vis-à-vis the outcome matrix. Psychological investigation may reveal that creatures like us do not suffer serious tension or ‘moral schizophrenia’ when we adopt the standpoint of a sophisticated consequentialist; instead, we can maintain the sorts of partisan commitments that make life worth living while accepting a teleological moral theory. And similar investigation might reveal that some robust but hard-to-justify moral intuitions, such as faith in a distinction between doing and allowing, are, like faith in basketball’s ‘hot hands,’ unshakeable though rationally unsustainable. The general thought here is that considerations from outside of moral inquiry will be relevant to settling disputes within it. Those naturalistic realists with normative axes to grind have these extra-moral facts as resources to use in their arguments for consequentialism.

The naturalist will insist that the full story about moral facts must be developed alongside other sorts of empirical inquiry, and will use the conclusions of non-moral investigation to suggest that (a) within the realm of normative ethics, the intuitions that fuel the anti-consequentialist position are those that are most likely to be revised in light

---

196 We might also think that exposure to other points of view is relevant here. Moral qualms about maximizing seem to vary with cultural differences. Recognition that certain moral principles are culturally specific (to some extent) forces reflection on the justification for these principles, which were once thought to be so obvious as to scarcely need support, and this leads us to see that justification is both required and hard to find.

197 For a discussion of this sort of sophisticated consequentialism, see Railton (1984). I should note some reservations about this use of the term ‘schizophrenia,’ which seems to trivialize a serious medical condition.

198 This is a bad example because the hot hands question is statistical, if complex, and the doing-allowing distinction is an issue within ethics and metaphysics (or the philosophy of mind; cf. intention debates). The point is that certain flawed heuristics are nonetheless irresistible.

199 Again, the first point is more important than the second for the realist’s purposes. If there’s a naturalistically respectable moral position that coheres well with these and other non-moral facts, and it isn’t consequentialism, the realist can happily admit defeat on the normative question and declare victory in the larger meta-ethical debate.
of a more complete social-scientific understanding of humans and their interactions; (b) competing normative theories conflict with our best views in the social and natural sciences;\(^{200}\) and (c) the non-naturalist’s commitments to spooky entities makes for a less coherent picture of how normative theorizing fits in with our best picture of the world.

These considerations put pressure on the Kantian non-naturalist, among others, by giving us reason to think that his position will not cohere well with the rest of our inquiries. But this is a problem for the Kantian, of course, only if we think that our moral theorizing should be of a piece with other subjects, and assuming this seems to beg an important question: it’s no surprise that a non-naturalistic view of ethics will stick out like a sore thumb against a backdrop of naturalistic endeavors like the natural and social sciences. Presumably, the Kantian thinks that moral investigation is different, perhaps by being immune from certain metaphysical worries. But, as I’ve suggested, this commitment has to be evaluated, and it seems to me that the way to do this is, in part, by looking at what it provides (freedom from naturalistic strictures) and what it costs (among other things, vulnerability to a certain kind of challenge— if moralizing is *sui generis*, then it’s also suspicious). If we think of the divide between naturalists and non-naturalists as one between different (truth-evaluable?) answers to a reasonable question, and, furthermore, if we think that this question can be adjudicated by appeals to reasons that are themselves neutral between the two positions, then we’ll have confidence that this question can be resolved. This, in turn, gives us reason to think that the answer to

\(^{200}\) John Doris (2002) argues that evidence from the situationist tradition in social psychology counts against the claims of virtue ethics. Roughly, the idea is that people just don’t have the kind of enduring, context-invariant personality traits that they would need to have if virtue ethics is to work. I offer this as an example of the sort of argument I have in mind.
this question will be one of the resources to invoke in discussions of conflicting
normative theories.

All of this discussion has been aimed at undermining the idea that normative
moral theorizing is at a kind of insurmountable impasse, without the raw materials
needed to decide between various theoretical options. The strategy has been to suggest,
in a quick and perfunctory way, that we have resources beyond those contained in our
moralizing. Debates in one area (metaphysics, or the philosophy of mind, or…) ramify in
others: as a result, we have to consider the overall appeal of various views measured in
terms of their fit with data points both moral and non-moral. If this expansive notion of
relevant considerations is correct, then we’re at least in a better position to hold on to
hope regarding the results of our moral theorizing. That optimism, recall, amounts to the
belief that theoretical disputes within normative ethics can be resolved in a reasonable
way, which means that reasonable interlocutors would be led to agreement under the right
sorts of investigative conditions.201 And that hope about our moralizing allows us to
preserve univocality across diverse moral conversations. Our original question concerned
how the committed proponent of some rival view—Kantian deontology, for
example—might nonetheless be connected with (and referring to, and so on) one set of
moral properties, in a manner consistent with our earlier worry about ‘fortuitous locking.’
And the answer is that there’s some theoretical continuity between the views of our

201 Just what it is to be reasonable is a hard question, and I won’t say anything about that here, except to
note that it would be all right, by my lights, if the notion turned out to be a bit thicker than simple
rationality. See Scanlon (1998), who discusses a descriptively rich notion of reasonableness. We might
also build thickness in as part of our account of the right investigative conditions, as these might include
having, say, particular emotional capacities. The obvious rejoinder is “mightn’t some interlocutor be ruled
out by these conditions?” and, of course, that’s a possibility. How interesting a possibility it is will depend
on just what substantive views are ruled out by the descriptive components of these notions.
imaginary Kantians and our hypothetical end-of-the-day views, such that we can understand the move from one to the other as a reasonable revision with a shared subject matter. Whatever the details of the correct theory of reference, it should allow us to say that speakers who are committed to norms of inquiry and theory-revision that lead them to agreement, under the right conditions, are in fact referring to the same entities, despite the differences in their current beliefs.  

4.9 Two kinds of convergence.

The realist needs the convergence claim if she’s to provide a satisfying answer to the semantic problem. But this isn’t enough, as we’ll see when we look more closely at the work that convergence promises to do for the realist. Ultimately, we’re looking for a way to justify the realist’s claim that diverse speakers, here and now, share terms in spite of their disagreements, and that they share terms in virtue of the right sorts of commitments. We need to say that there’s enough raw material in speakers’ body of commitments, norms, beliefs…to pick out one unique set of properties—if only implicitly. If moralists’ current commitments underdetermine which property best captures what they’re getting at, it’s hard to say that there is, nonetheless, a fact of the matter about what property they’re referring to, because this fact seems to be cut off from their cognitive lives. The question is whether or not these commitments are a rich enough

---

Note that Richard Miller, for one, thinks that convergence is not necessary for semantic continuity: we share moral terms and concepts with remote peoples despite the fact that our epistemic commitments are different enough to prevent this kind of resolution. Miller’s view is puzzling because, as far as I can see, he doesn’t provide a compelling case for interpreting these speakers as using our moral terms incorrectly rather than as using some different set of terms to do the same work. I suspect that it’s at least close to correct to say that moral terms and concepts are identified in part by the epistemic apparatus surrounding their use (that is, what kinds of reasons go into moral judgments, what facts are taken as relevant evidence, and so on). If so, then Miller’s argument for semantic continuity is flawed, since his view claims that speakers share terms when, in part, use of the correct methodology or ‘way of learning’ would lead to convergence. But the worry is that a change in methodology would be a change in subject matter.
resource to underwrite the kind of convergence claim we need. The realist might be out of luck even if some varieties of the convergence claim are true.

To see the problem more clearly, let’s distinguish between two different views of convergence. Both assume that idealized, fully-informed reasoners would end up in agreement at the end of the day, but they give different explanations of why this is so. According to optimistic convergence, our conceptions are sufficient, when refined under the light of idealization, to identify unique referents for our moral terms. On this view, speakers are, at some level, committed to norms of living and norms of view-revision that will, when applied and revised correctly, deliver up the property of rightness. Each step in the process of revision is governed by reasons, and each change in view is made because it is mandated by those reasons. Our inquiry merely clears away the impediments, such as self-interest, failures of reasoning, and so on, that prevent us from seeing these properties clearly. (Notice that this seems to involve the idea that there is a more-or-less uniquely correct way to revise our commitments; other reforms that might deliver up a different limit theory must be dismissed as mistaken. If they weren’t, then it’s hard to see how we would maintain that we’re connected to one set of properties rather than another.) Thus, there’s a fact of the matter, here and now, about the properties to which we refer, and a fact about the content of our end-of-the-day moral theory. Our

---

203 And, of course, it’s important that the light of idealization reveals only what’s already there, at least implicitly. The point is that we aren’t constructing anything new when we deliberate under idealization; we’re only getting clear about what we think already.

204 Correctly, that is, by their lights i.e. they reason correctly by their standards of what counts as a reason, etc.
inability to say just which property is *rightness* is a purely epistemic failure, because the facts are there waiting to be discovered.\footnote{It’s not clear that NMR needs to be committed to saying that rightness is identical to some other property.}

Contrast this with *pessimistic convergence*. Like optimistic convergence, this position holds that we will reach agreement at the end of the idealized day. But it goes on to say that the agreement we’ll reach is *not* the result of simply clarifying just what our inchoate commitments say when their inconsistencies and theoretical pressures are weeded out. This might be so for a variety of reasons. First, it might be that our convergence on a particular theory is fixed by the sorts of facts that don’t matter.\footnote{For example, suppose some sort of determinism is true. We might think that, in this case, there *is* convergence but that it doesn’t help the realist.} That is, our commitments themselves aren’t what’s doing the work of fixing on a particular end-of-the-day theory. Second, it might be that our moral inquiry simply underdetermines the end result of our investigation. Perhaps our idealized agreement is the result of, say, a willingness to defer to others simply for the sake of uniformity; a commitment to sharing *some* set of behavior-regulating norms, regardless of their exact content; and various accidents of inquiry, among other things.\footnote{On this view, the process of deliberation and idealization *doesn’t* merely reveal what’s implicit in our practice; it changes things along the way.} Our revisions in moral view are *not* all mandated by reasons, and they are not simply making our commitments explicit. Instead, our inchoate commitments underdetermine the course of our moral inquiries, and the path our investigation takes is affected by all sorts of contingencies that are not reflections of our views so much as they are accidents of the way our investigations unfold.
We might develop the contrast between optimistic and pessimistic convergence by considering what the two readings say about the influence of others. Optimistic convergence sees the force of testimony as *epistemic*: a dissident might change her view because she thinks her colleagues just see something she doesn’t about the matter at hand. Pessimistic convergence, on the other hand, suggests that the force of others’ views in a case like this stems not from the correctness of the revision, by the individual’s lights; instead, it’s explained by, say, the psychological influence of others, or the desirability of shared norms as such. Though we can, from the perspective of the conclusion of inquiry, read our investigation as having a continuous subject matter, it’s not possible to justify every revision in practice *at the time*. Or, more accurately, it’s not possible to say that every revision is forced on us by the balance of considerations then available.

This way of drawing the distinction makes it clear that the realist must be committed to the optimistic convergence claim. If pessimistic convergence is true, then there are insufficient grounds for the realist’s claim that speakers share reference to a common set of moral properties despite their theoretical differences. After all, pessimistic convergence suggests that equally reasonable, equally defensible lines of inquiry might move from our starting points and end up with very different end theories. If so, then the resources available within our practice, the ones the realist *must* draw upon in order to satisfy the constraints suggested by our discussion of Rey, are simply not rich

---

208 Also notice that it’s more or less the same point to say that under idealized conditions we *wouldn’t* reach convergence, if the idealization includes the stipulation that speakers revise their views only when there’s sufficient reason.
enough to solve the semantic puzzle. Resolution at the end of the day does us no good unless it’s the result of the right sort of facts about us here and now.\textsuperscript{209}

4.10 Some considerations in favor of pessimistic convergence

This distinction between optimistic and pessimistic convergence puts more pressure on the realist. So far, our argument has gone along these lines: the semantic puzzle cannot be dismissed merely with an appeal to, for example, causal theories of reference and content. The convergence claim is needed because it allows a way to connect diverse speakers with shared properties by the right sorts of ties, and so it allows the realist to avoid concerns about the fortuitous locking problem. This is crucial, because it clarifies an important difficulty: convergence might occur in a way that \textit{doesn’t} solve these problems. The convergence claim is, in effect, doing two things: it’s connecting speakers who have (what appear to be) substantive moral disagreements to a single set of moral properties, and it’s making this connection in a way that respects worries illustrated by the fortuitous locking problem. Hence, the appeal to convergence faces two sorts of objections. First, we might claim that we won’t, in fact, converge on a single set of properties, at least not under the sorts of idealizations that matter. Second, we might object that whatever convergence \textit{does} occur fails to connect speakers and properties in the right ways. That is, the link between them isn’t one that matters. The second objection is particularly potent, since it allows the critic to grant a contentious

\textsuperscript{209} There’s an interesting connection here to the Moral Twin Earth argument. Horgan and Timmons contend that twin-moralists would, after engaging in the same process of inquiry as Earth moralists, come to significantly different normative conclusions. They don’t address the possibility of underdetermination, but, I think, this is at the heart of the problem. If our starting point and methodologies were compatible with a variety of different theories, we’d have no basis for saying which of those theories describes the properties we were referring to all along. Thus there’s no sufficiently robust fact of the matter about our reference here and now.
claim (that idealized reasoners would share an end-of-the-day moral theory) while denying that the semantic puzzle has been solved.

Let’s turn now to the case for pessimistic convergence, which grants that we’d agree at the end of the idealized day and then denies that this matters. One interesting question is, of course, whether the claim is true—that is, whether our current moral discourse is too impoverished to underwrite the claim that we’re referring to a single set of properties, even now, despite our confusions. A second issue concerns how the answer to this question locates moral discourse in relation to other sorts of talk. It might turn out that normative and non-normative discourse are in the same boat when it comes to the relationship between our current views and our idealized limit theory, or it might be that this underdetermination worry is more troubling in one case than in the other. In order to see how worries about convergence and the semantic puzzle play out, we’ll need to give answers to these questions and then assess how our answers square with the realist’s proposals.

Start with a familiar worry about moral inquiry. We need to use our moral intuitions as data, if only because it’s hard to see where else to begin. But these intuitions look to be nothing more than the products of acculturation, ideological pressure, and so on. If our intuitions aren’t truth-trackers so much as they’re the detritus left behind by various social forces, then how can we use them as the starting point for an investigation that claims to get things right? This concern is what gives force to one of the skeptic’s favorite counterfactuals: if you came from different circumstances, you’d have different

---

20. I take this to include both the worry about convergence *simpliciter*, and the concern about the convergence occurring in the right way.

21. The appeal of (a certain kind of) constructivism here is that the intuitions are part of the apparatus that *makes* the claims true.
views (and different intuitions, emotional responses, and so on). The familiar worry is useful here because it focuses our attention on the *malleability* of our theories, intuitions, and other possible data-points used in moral theorizing. If our data are infected by various sorts of contamination, then it’s hard to see how we have a reliable means of getting at a robust, suitably mind-independent moral truth.

Concern about the reliability of our intuitions becomes more serious when we consider it in light of some common and attractive views about moral methodology. Consider some familiar Rawlsian points about coherentism in our moral investigations. According to this picture, neither broad principles nor judgments about particular cases are immune from revision. Instead, all of our commitments are subject to correction in light of the balance of considerations, and we always have the option of rejecting claims at any level of generality in order to increase the overall coherence of our moral (and non-moral) views. The important point, for our purposes here, is that the process is holistic: there is no fixed point that is safe from rejection, come what may. If so, of course, then the data we use in making revisions—the evidence that convinces us that a particular commitment has to go, for example—is itself subject to rejection or revision. There’s no difference in kind between our data and our theory: we use our intuitions about cases and principles to revise our intuitions about cases and principles.

The problem, then, concerns the extent to which our theorizing *infects* our data. If our moral inquiry goes awry, it changes the very evidence we’d use in order to get it back on track. It’s too strong to say that the difference between moral and scientific investigation is that only the latter offers the possibility of theory-neutral observations. Some interesting scientific observations may well depend on theoretical commitments.
That said, there might be an important difference of degree here that’s sufficient to cause problems for the realist’s proposal. The general suspicion, put metaphorically, is that, in the case of scientific theorizing, our errant speculations will eventually run into the bedrock of facts, or of indisputable observations, or of something that’s more or less immune to the infection of bad theoretical commitments. Our moralizing, on the other hand, is free to run unchecked because the judgments that serve as the basis for revision are more malleable and because there’s not some kind of theory-independent constraint to correct us when we go astray.

This is particularly troubling if we think that our inquiry is subject to various accidents and contingencies. That is, it might be that the course of our moral investigations is not entirely reason-governed, in the sense that its outcome is affected by facts that we think, intuitively, shouldn’t affect it. The way we treat novel cases might depend on the context in which we encounter them, for example, or on which of many competing action-descriptions we consider most salient. One particularly interesting sort of contingency concerns our interpretation of various interlocutors as participants in moral discourse, that is, whether we think of them as making moral claims or as giving some other form of evaluation. We might imagine meetings between moralists and some alien speakers in which the decision to interpret these novel agents one way rather than another is determined by, say, the order in which particular claims are made, or the ways in which various topics are broached. If we hear them make claims that are of a piece with our own moralizing—because, for example, their evaluations agree with our own, or

---

212 For example, an action is an instance of lying, of lying under duress, of lying under duress for a commendable purpose, and so on.
because their justifications are similar to those we ourselves give in moral
discussion—then we might think that their claims are semantically of a piece with our
own. On the other hand, if we begin with the points of difference between our practices,
we might take them to be talking about something else. That is, we might take the other
hermeneutical path, and see these alien speakers as making a different (i.e., nonmoral)
sort of evaluation, one that has an interesting relationship with our own moral judgments,
but which is nevertheless orthogonal to them.

Let’s now make the problem sharper by considering how these worries about the
contingencies of interpretation fit with our broader concerns about Rawlsian reflective
equilibrium. I’ve suggested, briefly, that our decisions to hearing other speakers as
making moral claims, rather than, say, claims about fierceness, or honor, or some other
evaluative property, are not determined solely by our linguistic disposition, but, at least in
part, are settled by seemingly irrelevant contingencies. Whether or not we hear the
explorers from Moral Twin Earth, or the Tiv, or whomever, as making moral claims
depends on accidents about, for example, our first encounters with them. Suppose this is
ture. The difference in interpretation is relevant to the outcome of Rawlsian coherentist
inquiry, because interpreting speakers as (perhaps deviant) moralists makes them, in
effect, members of our moral community, with intuitions, principles, and methodological
commitments (among other things) that must be taken seriously when we attempt to
move toward coherence. When we hear them as moral speakers, we take on the burden
of making sense of their views. This, in turn, changes the data that serve as the starting
point for our investigations. On the other hand, if we don’t interpret them as semantic
fellow-travelers, our moral inquiry isn’t affected by their views, because we see their
views as non-moral. Hence, an accident of our inquiry can have a significant impact on the results of that inquiry.

There are two conclusions to draw from this. The first is a point about the effects of interpretation: whether and how we converge is a function of how we interpret various marginal interlocutors.\footnote{Frank Jackson (1998: chs. 5-6) seems to think that we mean the same thing by our terms only if we converge on ‘mature folk morality’ when we reason correctly. If we don’t converge, then we—or those of us who don’t converge, at any rate—don’t share meanings. The problem with this is that we need to make judgments about univocity in order to get on with our inquiry. Or, to put it another way, our assessments and initial judgments about shared meaning determine the content of our mature folk morality. There’s no truth about the content of the mature folk morality (and no truth about assent to this by various interlocutors) that’s independent of (prior to?) our initial interpretations, i.e., our tentative judgments about other speakers’ meanings. Apparently, Jackson thinks that it’s possible for our moral discourse to converge in different directions—that is, we might divide our semantic community along the way, and decide that we’re talking about different properties. It’s not clear what would prompt us to say this rather then decide that we share essentially contestable concepts, or that agreement is impossible, or some other more pessimistic conclusion.} The second, broader point is that this concern with interpretation is only one instance of a general phenomenon, namely, contingencies of our inquiry that affect the outcome of our moral theorizing. If these contingencies exist, the result of our theorizing is underdetermined by the commitments we have before the end of the day. There’s simply no fact of the matter about where we’ll end up, even under idealization, because our path to the end of the day is filled with various contingencies that shape the direction of our moral theorizing. These considerations pose a challenge to the realist because they suggest that there’s no defensible way to connect present-day moralists with the properties picked out by a limit theory. If these underdetermination worries are correct, then present-day speakers are not connected, in the here and now, and by the sorts of bonds that matter, to the properties picked out by our limit moral theory. Since the realist’s solution to the semantic worry requires this connection, this is a serious problem.
CHAPTER 5

CONTINUITY AND INNOCENCE BY ASSOCIATION

5.1 Introduction

We’ve considered two attempts to make sense of the univocity of moral conversation and thus to preserve genuine agreement and disagreement. Both the naturalistic realist and the expressivist are unsuccessful in their efforts to solve the problem, for roughly similar reasons. Here, I’ll look at an important rejoinder, the reasons for its failure, and the light this sheds on some broader questions in meta-ethics and the philosophy of language.

5.2 What’s wrong with expressivism and realism?

Let’s begin with a quick look at the problems facing the two views we’ve been considering. Naturalistic realism’s alleged failure to make sense of shared meaning is an old theme and, we might have thought, one that’s easily dismissed: the early noncognitivists objected to ‘descriptive’ accounts of moral language on the grounds that these failed to capture what’s at issue between speakers with sufficiently different views, but these objections seem outdated and without force. Realists reassure us that our
substantive disputes are no impediment to shared meaning, just as, say, scientific disputes don’t place undue semantic strain on interlocutors, who manage to refer to the same electrons and tigers in spite of their conflicting theories. The old noncognitivist argument is further undermined by familiar appeals to the (once) New Theory of Reference, which promises to explain how these disputes are possible by explaining how to make sense of belief- and theory-revision over time, as any plausible account of meaning must. Clearly, disagreement is possible, in moral talk as elsewhere, and the noncognitivists gave us no compelling reasons to think that the moral case is especially problematic.

These familiar points are right: the traditional arguments against realism are faulty. But there’s a serious problem here that’s been obscured by the old argument’s reliance on implausible views about language. Even armed with the resources of the post-Kripkean era, the realist is unable to account for agreement in meaning. Quickly put, the problem is that the realist’s account of language must connect diverse speakers with a shared set of moral properties. This is possible when speakers disagree, of course, as familiar examples demonstrate, but this point shouldn’t obscure the important requirement that there must be something about the speaker that ensures the relevant ties between term and property. That is, the realist’s attribution of reference must be justified by some sort of appeal to facts about the speaker. While it’s true that this appeal can be made when speakers have serious differences in theory, noting this does not, by itself, give us sufficient reason to accept the realist’s semantic claim.

As I’ve argued, the core of the problem has to do with characteristic differences between moral discourse and, say, the language of empirical inquiry. Naturalistic realism

---

214 Including, in the right circumstances, social facts, etc.
is committed to a particular kind of story about how disputing moralists (at different times and places) are all talking about the same thing. But moral language cannot provide what’s demanded: our moralizing, unlike other parts of language, lacks the resources that are needed to make good on the realist’s debt. An important part of the argument for this claim—though, as we’ll see, not the only crucial part—is the idea that our moral discourse underdetermines the future course of moral inquiry. Where we end up, at the end of the day, simply isn’t fixed by our current fund of commitments, judgments, principles, and so on. The importance of this claim, and how it matters more to the moral case than to others, will be examined later in this chapter. For now, this point is only a reminder of the realist’s problem.

The realist’s difficulty isn’t surprising, given some commonplace reflections on the point of moralizing. The role of our practice has something to do with evaluating actions, agents, and institutions, and this, in turn, is useful because it allows us to coordinate activity and achieve important collective goods. Given that it has this function, rather than some other, we shouldn’t be surprised to find that it lacks the sort of specificity or determinacy that the realist’s semantic project requires. Moral talk can do its job without committing us to a specific cluster of properties from the start. If so, the realist’s account of disagreement rests on a picture of moral language that asks for too much, and it runs into trouble because we simply don’t have the resources to accommodate this demand.

Oddly, expressivism fails for roughly the same reasons. This is something of a surprise, not least of all because early expressivist accounts were motivated by the sort of platitudes about the function of moral language that we’ve just considered. By the
noncognitivist’s lights, the purpose of moral language is to commend, condemn, and influence; it’s not in the business of describing the world, and we should take this into account when trying to make sense of its workings. Yet the expressivist’s positive story is unable make good on this important observation. The realist’s problem is that his picture of moral language requires speakers to share a feature that just isn’t there, namely, links of the appropriate sort to a common cluster of moral properties. The expressivist, on the other hand, requires a particular relationship to common conative states. Regardless of just which state forms the foundation of the view (pro-stances, all-in preferences, norms for guilt and anger), it’s not clear why we’d need for all speakers to share that attitude in the relevant way, as the expressivist requires. The possibility of conative variation is consistent with the univocity of moral discourse.215

There are two important points to be made about this argument, both of which parallel remarks made about naturalistic realism. The first concerns the possibility of revision and rejection. In the realist’s case, the problem is that the ongoing changes in our moral thought leave its content so varied as to strain the proposed semantic link to various properties. In the expressivist’s case, the related problem is that the same sort of flexibility and revisability are true of the connections between moral judgments and various sentiments. That is, it’s possible to question, and perhaps reject, the proposed conceptual link between a particular conative state and moral judgment.216 The second

215 There is, of course, an additional problem: showing that the state in question is tied in the right way to moral evaluation, as opposed to, say, other kinds of normative judgment. (This is clearest when considering all-in endorsement as a candidate, though the argument can be made in other cases as well.) Here, though, the central point concerns interpersonal variation, which, if we take examples of amoralists and the like seriously, appears to be possible even with quite expansive states of endorsement.

216 This sounds too much like a Moorean argument to carry much weight in isolation, but combined with the constraints developed in the noncognitivism chapter it’s more serious. The point is not that we think
concerns conative variation within a community of moralists. As others have pointed out, we seem to interpret others as making moral claims even when they display nonstandard attitudes. Thrasymachus scorns justice but still talks about it: how is this possible if judgments about what’s just are constituted by pro-attitudes? Similarly, some communities might lack the parochial sentiments intimately associated with our own evaluations, if, for example, guilt is culturally specific, and yet we tend to think that theirs are moral judgments nonetheless.

The purported advantage of expressivism is its ability to make sense of shared meaning despite intersubjective differences in substantive views and in theory. The positive proposal comes at a cost, though, since genuine agreement and disagreement (or the expressivist’s variants of these) are available only when we have agreement in conative states. This, as I’ve suggested, is asking for too much, since it seems that we can and do get by without this sort of shared conations. Our ordinary patterns of interpretation seem indifferent to variations in conative attitudes, and for good reasons: this allows us to question whether our typical responses are the right way to feel, and it allows us to engage in conversation with those who, like Thrasymachus, have different attitudes.

So expressivism and naturalistic realism fail for similar reasons: they set the bar of competence too high by requiring too much of speakers who we ordinarily take to be part of moral conversation. In different ways, they each make moral discourse too parochial. Realism requires individual speakers to be connected to a common set of

---

217 This point is developed in Sturgeon (1986) and (1991).
properties, while expressivism requires agreement in conative response. Both of these features may be true of typical speakers, but they aren’t constitutive of participation in moral conversation. Part of the problem, of course, is that moral discourse seems to allow all sorts of atypical or deviant speakers into the discussion. But this permissiveness is threatened by the two proposals we’ve been considering.

5.3 A reaction

Both the realist and the expressivist find themselves in the same boat, broadly speaking, when it comes to the general problem, but they face different dialectical options when responding. This difference stems from each theory’s implicit view of the relationship between moral discourse and other sorts of language, in particular, straightforward empirical or descriptive talk. Expressivists, of course, think that the surface similarities between normative and non-normative language are misleading: grammar notwithstanding, these have different functions and work by means of different mechanisms. Moral talk is not of a piece with everyday empirical talk. Naturalistic realists, on the other hand, are impressed with the similarities between moral and nonmoral language and, accordingly, build their view around the observation that these discourses aren’t as different as they’re commonly thought to be. Call the thought that moral language works just the way other sorts of language do the continuity thesis.²¹⁸ I highlight this issue to emphasize that the differences between expressivism and realism are deeper than the incompatibilities between their positive proposals. At the heart of the disagreement is a more basic difference about where to place moral discourse in relation

²¹⁸ Blackburn tries, rather oddly, to adopt the continuity thesis while retaining his brand of expressivism in (1991). The problem, as Sturgeon points out, is that Blackburn’s way of making sense of conative variation via the continuity thesis is at odds with his own positive proposal.
to others—that is, whether moralizing is fundamentally akin to familiar descriptive talk or whether it’s quite different.\textsuperscript{219}

This is important because these differences over the continuity thesis affect the sorts of options expressivists and realists face when responding to our earlier arguments. The expressivist, we might think, is in a more difficult position, because his reply must be a denial that such conative variation exists within the scope of moral conversation. The expressivist’s rejection of the continuity thesis—a rejection that’s required by his positive proposal—rules out an appeal to, say, the mechanisms that preserve the univocity of empirical terms through changes in theory. In other words, the expressivist’s insistence that normative language is semantically distinct means that he has fewer resources on which to draw in answering the present objection.

The realist has a few more options. Like the expressivist, he can, of course, meet the objection head-on and deny that moral inquiry is underdetermined in the way that we’ve discussed. But there’s another, more intriguing line of response available, though it does come at a significant cost. Naturalistic realists have, of course, made various proposals about how to think about language generally, or at least such proposals are implicit in their discussions of moral discourse. These, however, are commitments above and beyond the continuity claim, and the realist, or someone with realistic leanings, might respond to our worries by jettisoning those semantic proposals while holding on to the continuity thesis. Instead of denying that moral reasoning is impoverished in the ways

\textsuperscript{219} This way of putting it emphasizes the semantic agenda behind realism and expressivism, and so highlights the extent to which error theory is like naturalistic realism—i.e., it sees the workings of moral \textit{language} as akin to empirical talk, then distinguishes it on \textit{metaphysical} grounds.
we’ve discussed, the defender of the continuity thesis might argue that all inquiry is underdetermined in just this way. This gives up the idea that moral terms are locked on to particular properties over time but retains the central thought that moral terms work just as others do. This is, in effect, a new instance of the familiar ‘innocence by association’ strategy that denies there’s any special problem for moral discourse. Just how serious a revision this is will depend on the details of the argument. Whether the resulting position should be called naturalistic realism, and just how far it deviates from the paradigmatic instances of the view, is less interesting than the more general issue brought up by this sort of response—namely, the relationship between moral and empirical inquiry.

Before examining these issues directly, we should take a look at the proposed reply. Here’s the sort of rejoinder I have in mind. Our argument against realism so far has relied on the idea that our moral commitments underdetermine the course of our inquiry. Our first-order norms and judgments, combined with the epistemic commitments governing our revisions, simply don’t pick out a unique path leading toward the idealized limit theory. If this is right, I’ve argued, the realist can’t make the critical claim that diverse speakers share reference to a single set of moral properties, because there’s simply no fact of the matter about which properties these are. What’s more, if this is to be an argument against a particular sort of moral realism, this feature must be distinctive of moral discourse, because we’re trying to show that moralizing just doesn’t work the way other sorts of language do. Moral discourse might not uniquely determine the course of future inquiry, the reply might concede, but this makes it just like empirical discourse, because the same underdetermination point holds there as well. To
use our earlier terminology, whatever convergence there is, within or outside moral inquiry, is pessimistic convergence. Thus we might give up the traditional realist’s claim that we’re locked on to particular, clearly defined properties over the history of our moral investigation, while still holding on to the continuity claim.

Before examining the details of this sort of reply, we should pause to note the extent to which it counts as a serious concession on the realist’s part. The naturalistic realist who takes this line is giving up something significant, namely, a picture of how language works that’s common to both normative and non-normative (particularly scientific) discourse. So this revision will require reworking that positive proposal and replacing it with something else. This is important, simply because so much of the discussion of the naturalistic realist’s position involves an analogy to scientific discourse that rests on the familiar view of language. At the same time, this ‘innocence by association’ strategy might be able to save several important commitments of the realist’s original position. In order to determine just what sort of concessions this strategy involves, we’ll have to look in more detail at the sort of position that comes out of such a maneuver.

5.4 The detour

The reason that the ‘innocence by association’ rejoinder is more than an idle threat is because ideas like optimistic convergence have come under attack even in cases of empirical discovery. We have reason to think that, even in some paradigmatic cases of scientific inquiry, we face the same sorts of underdetermination worries that are so troubling in the moral case. If so, it appears that what we thought was the distinguishing

---

20 See Boyd in particular.
feature of moralizing—its undetermination, vis-à-vis the limit theory—turns out to be true of all sorts of non-moral inquiry as well. This, of course, allows room for a defender of the continuity thesis to argue that moral inquiry is of a piece with empirical inquiry, which would show there’s no *special* moral problem. This requires rejecting some of the semantic claims that realists have found appealing, but it preserves one of NMR’s fundamental commitments. If this argument works, we’ll be forced either to accept the continuity thesis or provide additional arguments in favor of a distinct moral difficulty.

The argument for underdetermination in the straightforward empirical case runs as follows. In a variety of instances, we can maintain, plausibly, that nothing in (for example) a past community’s use of a particular term uniquely determines its extension. Even construing ‘use’ broadly, to include the mental lives of community members, linguistic dispositions, relations to environmental facts, and so on, we’re forced to admit that the community’s use of the term leaves open different views of the term’s extension. So, for example, we think that “water” or “gold” has the same extension now and in 1650, i.e., before anything like contemporary chemistry, but we’re hard-pressed to point to features of the 1650-speakers’ use of the terms that commit them to thinking that microstructure, rather than, say, phenomenal qualities, are essential to something’s falling under “water” or “gold.”

This point will be clearer after considering some examples. Here are two. In both cases, the structure is the same: the totality of facts about the community’s use of a term leave open various ways to proceed in novel or unanticipated cases. The first example is from Mark Wilson’s “Predicate meets property.” Long ago, let’s suppose, a colony of
Druids sets up shop on a remote South Seas island, where their descendents live, more or less undisturbed, until slightly before the present day. Then,

a B-52 of regular American types landed on their uncharted island and the Druids exclaimed, “Lo, a great silver bird falleth from the sky.” Aside from the local avian population, this bomber was the first flying device the Druids had ever seen….Airplanes, helicopters, and dirigibles are now called “birds” in the Druidic dialect and true *aves* are specially denoted by compound phrases like “feathered birds.” (M. Wilson 1982: 549-550)

This classification seems sensible enough. But consider an equally plausible counterfactual history—what would have happened if things had gone just a little bit differently:

If the hapless aviators had crashed in the jungle unseen and were discovered by the Druids six months later as they camped discontentedly around the bomber’s hulk, their Druid rescuers would have proclaimed, “Lo, a great silver house lieth in the jungle.” This appellation might stick and the later, globe-trotting Druids would persist in calling all sorts of aviation “flying houses,” to be distinguished from “stationary houses” and “earthbound but mobile houses.” Here airplanes are no longer held to be “birds.” (M. Wilson 1982: 550)

The point of all of this is that there’s nothing in the Druids’ linguistic dispositions, before their encounter with the plane, to determine whether or not this unexpected object is in the extension of “bird” or “house.” How they classify the object depends on an accident: whether they encounter it first in the sky or in the air. And both the “bird” and “house” classifiers insist (with good reason) that their taxonomies do not represent some change in meaning or concept—that is, both think they’re just extending their conceptual repertoire to cover new cases.

Consider a second example, from Gary Ebbs’s “The very idea of sameness of extension across time.” First, some chemical and historical stage-setting. Platinum (discovered in the mid-18th century) and gold are different elements: platinum is the
element with atomic number 78, and gold is the element with atomic number 79. Yet platinum and gold share many features—in particular, they both dissolve in aqua regia, which was named for its ability to dissolve gold. Now for the thought-experiment:

Suppose that there is a Twin Earth that is indistinguishable from Earth up until 1650, when large deposits of platinum are uncovered in Twin South Africa, and that once it is established by Twin Earth chemists that the newly uncovered metal dissolves in aqua regia, members of the Twin English-speaking community call it ‘gold,’ treating it in the same way we treat gold…Suppose also that on Twin Earth chemistry develops in almost exactly the same way in which it develops in our community, except that when chemists in the Twin Earth community investigate what they call ‘gold,’ they conclude that there are two kinds of ‘gold:’ one of these kinds of ‘gold’ is gold, which they know only as the element with atomic number 79, and the other is platinum, which they know only as the element with atomic number 78. We can say that their word ‘gold’ is true of an object \( x \) just in case \( x \) is gold, the element with atomic number 79, or platinum, the element with atomic number 78. (Ebbs 2000: 248-9)

The point of this example, for our purposes, is that, before 1650, the two communities have the same dispositions, mental states, and so on, and that all of these things leave open whether to think that a new bit of stuff (which will turn out to have atomic number 78) is a new kind of gold or a new metal that’s a lot like gold. Both of the classifications are completely compatible with the pre-1650 dispositions, beliefs, and so on, of Earth and Twin-Earth speakers, just as the Druids’ use of their terms is compatible with either way of extending them to new cases.

These examples suggest that our use of terms (including our dispositions, first-order theories, higher-order norms, and so on) underdetermine the way that terms should be applied in at least some new cases. This, in turn, suggests that we’re not ‘locked on’ to a single (determinate) property throughout the term’s history, because there’s no fact of the matter about whether we’re hooked up to, say, one microstructure or several.

While Ebbs uses these sorts of cases to argue against a popular thesis about
language—that the use of a term at a given time determines its extension at that time—we don’t need to draw his conclusions in order to see the problem for our argument against realism. That argument suggested that the indeterminacy of our moral inquiry poses a special threat. These examples suggest a natural ‘innocence by association’ reply: since *other* sorts of inquiry suffer from the same indeterminacy, moral discourse has no special problem, and thus there’s no special threat, at least to the continuity thesis. While these underdetermination cases might pose a threat to particular views of the workings of language, both moral and nonmoral, there’s nothing here that singles out moral discourse as uniquely problematic. Thus, the realist’s central claim, that moral language is not especially problematic, is preserved.

5.5 The significance of underdetermination

As we’ve noted, realists might reply to the argument from underdetermination by denying that moral reasoning is as impoverished as that argument claims. If this can be made good, there’s no reason to think that there’s any *special* problem with univocity. This requires rejecting Ebbs’ argument, of course, and retaining a more traditional picture of language. It also involves showing pessimistic convergence about moralizing in particular isn’t as plausible as it looks. But the Ebbs-Wilson arguments suggest a different tactic. Since they cut against both the discontinuity thesis and the traditional realist’s semantic claims, someone whose primary commitment is to the continuity claim might see these arguments as powerful ammunition. She might reject the realist’s picture of language, that is, while holding on to the idea that moral talk is of a piece with other

---

221 This will be a difficult task, given the various features of moral discussion that prompted us to be suspicious of moralizing in the first place. One reason the ‘innocence by association’ strategy is appealing is that it doesn’t require the realist to make implausibly strong claims about the course of moral inquiry.
sorts of language. If the same underdetermination worries affect “gold” as well as “right,” she will point out, we haven’t yet found good reasons for thinking that univocity poses a special problem for moral discourse.

As we’ve noted, this isn’t a maneuver the realist, as we’ve imagined him, can make easily, because it involves a significant concession about the workings of language. The sorts of mechanisms that preserve semantic continuity across scientific and moral inquiry, both synchronically and diachronically, are no longer part of the picture. This leaves the defender of the continuity claim in something of an awkward position, because we’re left with the insistence that the continuity thesis is correct, but no explanation of why it’s correct. The realist’s old picture, which involved speakers being connected to natural properties through various kinds of causal (or, in Boyd’s case, epistemological) links, and which at least attempted to explain how shared meaning is possible, has to be jettisoned once we see the force of the underdetermination arguments.

At the same time, the underdetermination point puts pressure on our initial argument against realism, because it provides a way to deny the objection’s central claim: that there’s a special problem with the univocity of moral conversation. The defenders of the continuity thesis can help themselves to many of the traditional realist’s arguments against the distinctness of moral discourse, since these often don’t rely on the details of the linguistic view that’s under attack. These advocates for continuity can admit that they don’t have a wholly satisfying view about reference and meaning while providing reasons to think that our best views of these issues, whatever they turn out to be, will

222 For example, Sturgeon, Brink, Railton, etc.
yield no interesting difference between moral and non-moral talk. In response, we have to do more to explain why there’s a special problem in the moral case. What’s really at issue, then, is the continuity thesis itself, and whether we should think that our puzzle about moral disagreement is simply an instance of a more general philosophical worry. What is it about our moralizing that puts unusual strain on univocity—or, to put it another way, why isn’t innocence by association enough?

It’s easy to see what sort of reply is needed, at least in abstract and general terms. We need to identify differences between moral and non-normative discourse that cast doubt on the continuity thesis. If the characteristic features of moral discourse make the univocity problem particularly difficult, or require a different sort of solution, the innocence by association reply will be unconvincing. Here I’ll try to provide some reasons for thinking that there is a problem for moral discourse that goes beyond the general underdetermination worries raised by our thought experiments. This argument will involve two stages. In the first, I’ll identify some features of our moralizing that undermine the analogy between it and, say, the development of chemistry or the Druids’ investigations into birds and houses. This will put some initial pressure on the continuity thesis, because these differences seem to make the univocity problem more pressing in the moral case. After looking at a view of language that’s suggested by the Ebbs-Wilson phenomenon and the underdetermination worries more generally, I’ll provide a few

---

223 We might think of this as a meta-ethical view that says there’s no distinct meta-ethical problem: the story about univocity in the moral case will be the same as it is for the natural sciences.
224 If it’s just a puzzle about language generally—how can we share meanings despite our different beliefs?—moral theory itself has little to offer.
225 One question that’s being glossed over here is whether or not the second realist response (accept Ebbs/Wilson and use this to motivate innocence by association) really counts as a realist response.
arguments that suggest there is a special problem in the moral case and thus that the continuity thesis is in danger.

Let’s start with some fairly obvious but important differences between the moral case and worries about gold and platinum. The crucial feature of the Wilson-Ebbs thought experiments is that the totality of the community’s dispositions, mental states, and so on leaves open what to do in an unexpected case. The Druids hadn’t thought about a flying device that might double as a shelter, and we, at one point in our chemical investigation, hadn’t thought that there might be two microstructures that behave in such similar ways, e.g., dissolving in aqua regia. That’s why the cases are interesting: there seem to be (at least) two ways of going on, and neither seems to fit past practice better than the other. Of course, there are examples of moral cases like this as well; we might think, for example, that hard cases at the margins of life present us with new and unexpected situations that mirror the gold/platinum case. But this shouldn’t lead us to overlook an important difference. In the examples, there’s fairly stable community-wide agreement on what gold is, or what a bird is, and that agreed-upon account doesn’t say anything about the new case. (Not all speakers will be in on the theory, of course, but we might imagine the right sorts of deference and expertise would hold the community together nonetheless.) As a result, progress in theory appears, in many cases, like refinement: we learn, to our surprise, that there are two microstructures in our lump of (what we judge to be) gold, and then decide whether our classifications are mistaken, or if “gold” is disjunctive. In moral discourse, on the other hand, there is no well-

---

226 Ebbs’ Twin Earth cases are meant to reveal that a counterfactual history in which the community takes the other path has just as much claim to getting it right as we do, and so there’s no way of saying which option is the correct one.
developed, stable, and accepted account of what makes an action right. The problem here is not that there’s received wisdom that leaves gaps; it’s that there are competing accounts that pull in different directions when it comes to the disputed cases. And, of course, there are a lot of disputed cases, and the disputes have staying power.

This difference is relevant because it’s this disagreement, not the threat of underdetermination by itself, that puts pressure on univocal interpretation. There’s no reason to suspect, in the case of “gold” or “house,” that speakers are talking past one another—at least not speakers at a single time. For one, there’s widespread agreement in belief and theory, and disagreements about the extensions of the terms are, for the most part, explained by differences in other beliefs (e.g., that a substance would dissolve in aqua regia). For another, speakers recognize the same authorities and, for the most part, the same standards of evidence concerning the stuff in question. The sorts of disagreements that mark moral discourse, where speakers have seriously different theories, don’t recognize the same methods for resolving disagreements, and so on, put real pressure on our interpretations, and raise the possibility that we are talking past one another. It’s this pressure, I suspect, that makes it more important in moral discourse than elsewhere that we stand in some kind of interesting relationship with a common set of properties. That is, this metaphysical support (of the sort provided by the realist’s traditional view of language) is needed so badly here because our ordinary conversations seem to strain the univocal interpretation that’s so natural, and unopposed, in the empirical cases. The support might be provided by an argument showing that, for

---

227 This isn’t to say that there aren’t gaps of the same sort in moral thinking, of course, just that here there’s an additional problem.
example, we’re all locked on to the same cluster of properties, despite differences of opinion. But if this is impossible or implausible, as I’ve suggested that it is, the moral realist is *not* simply in the same bad position as the gold-realist, or house-realist, or whatever. He’s in a significantly *worse* position because of the special nature of moral discourse and, in particular, moral disagreement.

This point reveals what’s flawed, or at least fundamentally incomplete, about the innocence by association strategy. The objection we’ve been developing rests, in part, on the idea that our moral discourse underdetermines the correct path of future inquiry. This might have force against the *traditional* naturalistic realist, but defenders of the continuity thesis will be quick to argue that everything else works this way too, and so there’s no special cause for concern in the moral case. This reply, it should be said, has merit, because it requires us to say more about why underdetermination is a special problem. But the reply is incomplete, for the following reason. If we find the Ebbs-Wilson examples compelling, and if we accept the associated arguments, we’ll have to reject some familiar and popular views about language. We’ll have to do more than that, however: we’ll also need to say how we can go on making judgments of univocity (or, in Ebbs’ terminology, how we can make practical judgments of sameness of extension at a given time), or we’ll have to deal with the consequences of rejecting these judgments. The first alternative is obviously more appealing, simply because of the role and importance of the judgments that are under attack. Suppose such a story is possible. We’ll *also* need an argument that *this* proposal allows for moral conversation to end up all right, just as, presumably, it ends up vindicating talk about houses and gold. It’s not at
all clear that this is possible, given the characteristic differences between moral and non-normative conversation that we’ve been discussing. In other words, the defenders of the continuity claim are committed to arguing that moral talk will come out all right when seen through our replacement view of how univocity is possible. But not all discourses are created equal, and it remains to be seen whether the problematic features of moralizing will threaten the realist’s strategy.

5.6 A minimalist proposal

Taken by itself, this isn’t much of an argument. It’s more of a suspicion that the unsettling features of our moral conversation—the ways, that is, in which moral conversation looks different from, say, the Druid’s house-talk or our past gold-talk—will throw a wrench into the attempt to make sense of univocity. Underdetermination itself isn’t enough to put the kibosh on the continuity thesis, as we’ve seen, but this doesn’t mean that moralizing is in the same boat as empirical talk. In order to add some force to the suspicion, we should look at what general lessons might be drawn from our Druids and Twin Earth speakers, and then see how those lessons square with the characteristic oddity of moral talk. This can’t be a conclusive argument, because the general lessons (and positive proposal) that I’ll consider aren’t the only ways of responding to these Wilson-Ebbs worries. It will, however, put some pressure on the innocence by association strategy. The lesson to be learned from all this, I suspect, is that moral discourse is seriously problematic no matter what underlying view of language we adopt.

Now let’s consider some general responses to our predicament, as revealed by considering the lessons taught by Druids and Twin Earth chemists. The unsettling result is that our terms aren’t tightly bound up with a fully determined extension at all points of
their use, or, if they are so bound, it’s not by our patterns of use at a given time. The puzzle arises this way. We think that the extension of ‘gold’ as we use it today is the same as the extension of ‘gold’ as used in 1650—that is, we think our term and theirs are the same, have the same extension, and so on. But, as George Wilson puts it,

strong ‘realist’ convictions about the grounding relations between language use and truth conditions incline us to the view that the extension of a term, as it is used at a time t, is determined by facts about the use of the term in the language at or before t, together with pertinent facts about the various items to which the term prospectively applies. (G. Wilson 2000: 90)

This suggests, in turn, that the use of ‘gold’ in 1650 determines that it applies to something if and only if it’s composed of the element with atomic number 79. But this seems false, as the Twin Earth thought experiment suggests, so we must reject either our strong realist convictions or our judgments of sameness of extension across time. Ebbs suggests we reject the former and take our judgments of sameness of extension across time as basic or primitive. A view of this sort (that is, taking these judgments to be basic, instead of looking for some deeper sort of foundation) is suggested in this passage of Mark Wilson’s:

[The traditional view] conceives of language rather as a railway line designed by a corps of engineers (the advance party sent out at the “setting up” of the language) and the process of employing it consists in sitting in the observation car and remarking upon the scenery passed en route. I advocate thinking instead of those comical locomotives in children’s literature which unroll their tracks before them as they move through a terrain. In the short run, the path of the train

Oddly, Ebbs doesn’t spell this out in detail, but it seems to be the only option left. Consider, by way of exegetical support, his remark that thesis (M)—the thesis we’re to reject, which claims that use fixes extension—“encompasses all substantive theories of, or conjectures about, what determines the extension of a word—theories or conjectures that do not just trivially assert, for instance, that what determines the extension of a word is the extension itself” (Ebbs 2000: 246). Since we’re rejecting all substantive accounts, we’re left with the sort of minimalism I’m describing here. It’s worth noting that Simon Blackburn’s tentative and sketchy proposal, in “Just causes,” seems to be along the same lines: we make interpretive judgments, and that’s all there is to it. In particular, there are no prior facts about univocity, or sameness of extension over time, or whatever, that our judgments must answer to. The view I’ll sketch below is one that I take to be a development of Blackburn’s proposal.
responds to the hills and gullies of the landscape virtually as well as the better designed line and the view from the observation car may seem practically identical. (M Wilson 1983: 586)

Clearly, this suggestion has strong affinities with our earlier discussion of optimistic and pessimistic convergence. If Wilson is right, we develop our language as we go along, and we needn’t be bothered by the absence of the rigid constraints that are present in the ‘optimistic’ view.

This suggests that our worries about univocity are, in a sense, misplaced. Both the naturalistic realist and the expressivist looked, in vain, for some firm ground on which to rest shared meaning and hence real disagreement. We count as participants in moral conversation in virtue of shared conation, or links to properties, or some feature of this sort. Ebbs and Wilson urge us to reject this demand and reconcile ourselves to the idea that no such foundation is required. It’s norms all the way down, we might say: here there are nothing but interpretations, judgments, and commitments to treat other speakers as semantic fellow-travelers, at least up to a point. We should take our interpretations and commonplace judgments of univocity as basic, rather than looking for some other sort of foundation.

What would such a view of moral discourse look like? In many ways it would be similar to the traditional realist view. Moral terms express or refer to properties, they’re used to make (objective) claims with truth values, and so on. But, unlike the usual realistic view, which tries to augment these claims with some further story about how these terms refer to these properties—a story often couched in causal terms, for example—this minimalist suggestion rejects the need for this account. What’s basic, here, is not some view about ‘causal regulation’ or the like, but our judgments of
univocity themselves. There’s nothing beyond or behind our best judgments of shared meaning, hence there’s no deeper feature to which we can appeal to shore up our confidence that moral agreement and disagreement is in good standing. Consider this suggestive remark of Ronald Dworkin’s, from another context, concerning the continuity of concepts and meaning:

…the institution has the continuity—to use the familiar Wittgensteinian figure—of a rope composed of many strands no one of which runs for its entire length or across its entire width. It is only a historical fact that the present institution is the descendent, through interpretive adaptations of the sort we noticed, of earlier ones…Which changes are great enough to cut the thread of continuity? That itself is an interpretive question, and the answer would depend on why the question of continuity arises. (Dworkin 1986: 69-70)

The suggestion is something like this: doubts and reassurances about univocity can be answered only within the practice, not by appeal to some causal link or reference-preserving chain. Our norms of interpretation have the last word. So there’s nothing more to say in the moral case either—that is, moral talk really is of a piece with other sorts of language—unless, of course, there’s something within the practice that places undue strain on our interpretive norms. Dworkin’s parting shot is not, I take it, a rhetorical question; it brings up the possibility that various features of a discourse might raise the threat of equivocation from within. So let’s turn back to the moral case and, in particular, to the features of moralizing that seem to pose a threat to univocity.

5.7 Against continuity

Assessing the minimalist suggestion is a delicate matter. Since there’s no background story about the preservation of univocal disagreement, as there is in

---

Note that this is not to say that our judgments are indefeasible.

Contrast the anti-reductivist commitments of Dennett and Davidson.
traditional realist views, the only way to argue that moral discourse faces a particularly pressing challenge on this front is to argue that our ordinary norms of interpretation (or suitably idealized norms…) are strained by moral conversation. And this is a challenge, we might think, because worries about univocity are fairly abstruse philosophical concerns. Since they don’t enter, in any straightforward way, into ordinary worries about moralizing, there’s some evidence that our interpretive stance treats moral disagreement as unproblematic. The very nature of the view in question has ruled out, as question-begging, the various arguments we’ve considered against more semantically robust views. As Dworkin’s intriguing remark suggests, we’ll have to argue that the problem arises not only from some ‘external’ vantage point, but from within the practice, as it were. This is as it should be, since, after all, we’re wondering if moral discourse raises any special worry about univocity, or if it’s of a piece with other sorts of language.

Contrary to initial appearances, it’s too quick, and perhaps false, to say that the univocity problem rests on unnecessarily baroque conceptions of language. I’ll argue that moral discourse is particularly problematic on this score, even if the minimalist proposal we’ve considered above is correct. Before looking at these arguments, though, it’s important to get clear on what’s required of them. We’re setting the bar too high, I think, if we demand evidence that ordinary speakers do or should (under minimal idealization) regard moral conversations as equivocal. This is because equivocation is something of a term of art (and so it’s fairly removed from ordinary speakers’ conversational interactions) and, in addition, because our ordinary conversations don’t often provide instances of ostensibly univocal disagreements that are later recognized as spurious. What’s needed, instead, is evidence that speakers react to moral conversations
in a way that suggests an implicit awareness of something deeply amiss. This might take a form of a sort of implicit judgment of equivocation, or it might not: what matters is that speakers’ responses and interpretations are sensible or understandable responses to *equivocal disputes*, not that their responses are explicit or implicit *judgments* or *attributions* of equivocation.\(^{231}\)

That said, there are a few different sorts of considerations suggesting that moral discourse places undue strain on our norms of interpretation. The first is an argument from the sociology of philosophy, and it starts with the observation that worries about univocity in moral discourse became an issue long before those same problems were seen to infect *other* parts of language. Ayer’s (1936) confused musing about disagreements (including his doubts about their possibility, their dependence on shared values, and so on), and Stevenson’s (1937, 1944) demand that any view of moral judgment account for genuine incompatibility, suggest that something about moral disagreement raises the possibility of equivocation. As our brief survey of early noncognitivists reveals, worries about this issue were in the air decades before various post-Kripkean developments focused attention on concerns about shared meaning and nonstandard theory.

This observation doesn’t carry a lot of weight by itself, but it does suggest that there’s something about moral conversation that raises the specter of equivocation in a way that, say, disagreement about gold and water do not. The threat to univocity is at least more *salient* in the moral case than in these others. And, what’s more, the features of moral conversation that make this threat more salient, in particular, the way

\(^{231}\) If “implicit judgments of equivocation” come to nothing more than responses that are sensible reactions to equivocation, then this distinction comes to nothing. We might think, though, that an implicit judgment is something more robust.
disagreements work, are, in fact, features that pose genuine problems for the continuity thesis (as we’ve seen in Chapter 4). Early noncognitivists can be understood as struggling to provide an account that made sense of univocity, and this suggests that they saw the unusual or distinctive features of moralizing as threats to genuine disagreement. Thus the need for a radical reinterpretation. These considerations suggest that sensitive observers thought, perhaps implicitly, that moral conversation strains univocal interpretation. This is some evidence that it is more problematic than, say, straightforward empirical discourse.

The second argument is motivated by the special character of moral disagreement, and, in particular, to speakers’ responses to this disagreement. Everyday speakers often think that (at least some) moral disputes cannot be resolved, that certain issues aren’t worth talking about, and that fundamental ‘differences of opinion’ are an unavoidable part of our moral practice. As a result, many of these speakers simply give up on certain moral conversations. I suggest that this reaction is some evidence that our interpretive norms are stretched a bit thin by moral discourse. One natural explanation for why speakers would avoid these disputes, after all, is that they’re ill-founded; that is, the reaction is a natural and sensible response to a problematic conversation. In the moral case, disagreements without clear and neutral means of resolution strike us as pointless because, at least for practical purposes, it’s as if disputants are talking about different things. Whether or not speakers’ responses count as implicit judgments of equivocation, their reaction to moral argument can be understood, plausibly, as a response that indicates a sense that something’s amiss. Furthermore, if moral conversation were equivocal, we’d
think that this response is a sensible one.\textsuperscript{232} This counts as some evidence that speakers implicitly view moral discourse as problematic and respond to it in a way that makes sense if the conversation is equivocal.

There’s a particular application of this second argument that deserves special mention. This concerns relativism and its attractions to ordinary participants in moral discourse. Speakers frequently make conversational moves that are often interpreted as manifesting a commitment to a ‘thinking makes it so’ variety of moral relativism. Yet this interpretation isn’t required by a charitable reading. Their remarks seem to indicate, instead, a deep skepticism about any kind of meaningful discussion between cultures or individuals with radically different commitments. This skepticism needn’t be glossed as an implausible relativistic thesis, though it often is; it’s equally compatible, and perhaps better explained by, the idea that moral discourse is equivocal, or at least diverse and chaotic enough to tempt us toward that interpretation. This, in turn, is further support for the claim that moral discourse strains our interpretive resources in a way that other sorts of inquiry do not, a conclusion that casts some doubt on the continuity thesis and the ‘innocence by association’ strategy.

The third argument concerns the principle of charity, and it runs along similar lines. Suppose that we’re faced with a community of speakers who appear to be making radically different moral judgments. If we interpret them as fellow moralists, we’re committed to condemning their judgments, as well as their actions: after all, they’ve gone horribly wrong in their moral assessments. This is unpalatable, simply because we don’t

\textsuperscript{232} As our earlier examination of the Moral Twin Earth argument demonstrated, there can be real incompatibility without moral disagreement, narrowly construed. One thing that remains at issue between disputants in a deadlocked conversation is, of course, what to do.
see any reasonable explanation of how they’ve gone wrong in such dramatic fashion. (We can say that they’re insufficiently sensitive, or rotten to the core, or…, but in many cases this is just name-calling.) We can avoid this unappealing conclusion by seeing them instead as making a different sort of evaluation and, furthermore, as getting things right. (This interpretation isn’t plausible in cases where they’ve gone wrong in straightforward ways, but it becomes increasingly attractive when our interlocutors’ errors become increasingly inexplicable on attribution of our moral concepts.) So a more charitable interpretation makes use of the idea of equivocation: to the extent that we need to make our interlocutors come out all right, we’ll have them meaning something else with their terms.

Our fourth argument is based on similar charitable considerations applied to the epistemology of moral disagreement. Like the other arguments, it attempts to make good sense of parts of our practice that might appear to be misguided or confused. The underlying theme is that, once we acknowledge that our moralizing is equivocal, or at least strains univocal interpretation, we’ll be in a position to understand why ordinary responses have certain features. In this case, the phenomenon to be explained is the effect of certain kinds of moral disagreements on our own commitment to our moral positions. People seem to be (almost) completely undeterred by the existence of equally well-informed, equally thoughtful individuals with diametrically opposed views. That is, they don’t take this disagreement to be a reason to doubt that their own commitments are the right ones. Yet, we might think, they should, because, generally, the existence of such opposing doxastic attitudes about some proposition should undermine confidence in it (and its negation, for that matter). We have no convincing story about why they’re
wrong, let’s imagine, and we think they’re as insightful as we are. Their contrary conclusions should give us pause, since we see that it’s hard to say, in a neutral and fair way, why we’ve got it right and they don’t. Yet we don’t often respond this way to moral disagreements.

One way to make sense of this epistemological intransigence is to interpret our interlocutors as meaning something else with their terms. If speakers with strange ways of doing things aren’t making judgments about rightness at all, their differences are no longer a threat to our judgments, or our confidence in them. So our reactions—that is, our failure to back off from our judgments in the light of what looks like ‘reasonable disagreement’—would be sensible reactions if moral discourse is semantically ill-founded in the ways we’ve been discussing. This point, in turn, provides some evidence that participants in moral discourse are responding to it in ways that suggest some serious problems. Their responses are defensible if moral conversation puts stress on univocal interpretation, while they may not be if we interpret moralizing in a straightforward way. The sorts of disagreement we find in moral discussion, then, seem to prompt the sorts of responses we’d expect if speakers thought, at some level, that the discussion wasn’t entirely all right. This, in turn, undermines the continuity thesis from within, as it were, by appealing to nothing more than our norms of interpretation.

The epistemological argument I’ve just sketched faces an obvious and serious objection. Other sorts of disagreements work the same way—that is, they don’t, in themselves, prompt interlocutors to revise their views or withdraw their confidence. These disagreements don’t really tempt us to think of the discourse as covertly equivocal, even though the epistemological argument we’ve considered draws that conclusion about
our moralizing. We might think, for example, that theological disputes are a clear example: believers of different faiths are commonly aware that only accidents of birth stand between them and the heathens, they admit that rival faiths have equally good support for their views, and they concede that they are in no better epistemological position. Yet they go on believing as they do, in most cases, which involves at least an implicit commitment to the falsity of rival views. The situation is comparable to the one in moral discourse, but we’re not tempted to think that, say, Christians and Muslims are simply talking past one another, and most often Christians and Muslims aren’t tempted to think this either. Yet our argument claims that this epistemic phenomenon is evidence for that conclusion in the moral case.\textsuperscript{233} The objection requires an explanation of the asymmetry.

One obvious but unappealing response to this argument is to capitulate—that is, to conclude that disputes between, say, the theist and the atheist are problematic in just the way that moral disputes seem to be. What’s interesting here is \textit{why} it’s so implausible to say that these theological debates strain univocal interpretation. The crucial difference seems to be that theists and atheists agree on just what’s at issue: whether or not there exists a being with particular attributes who is responsible for various events.\textsuperscript{234} Of course, there isn’t complete agreement as to the nature of the entity in question, but people on various sides of the debate do seem to share a rough-and-ready conception of what they’re looking for. The same is not true in moral dispute, where the worrisome

\textsuperscript{233} It’s important to note that these points apply to debates over theism itself, as well as to interdenominational disagreements.
\textsuperscript{234} We’re often skeptical of speakers who argue for God’s existence by saying that God is something more banal and less metaphysically problematic than typically supposed. This seems to change the subject rather than illuminate it.
differences are really over which acts are the right acts, not over whether or not there are right acts. In moralizing, we’ve got significant disagreement about the ‘role description’ of, say, rightness, in addition to worries about whether anything can meet the demands of these descriptions. The semantic worry, as opposed to Mackie-style objections from queerness, is driven by the first issue, not the second. (For comparison, imagine that the atheist and theist disagree about whether God exists and what sort of thing God is). We agree, more or less, on just what’s at issue in disputes over God’s existence, and the disagreements about what counts as evidence are less severe than in the case of normative dispute. Given this sort of rough agreement, we tend to think that disagreement about theism is on firm ground; since it’s absent in the moral case, the univocal interpretation is undermined. And this suggests that speakers’ failure to react to disagreement is (at least to some extent) vindicated in the case of moral discourse and brought into question in the case of theological dispute.

This quick, incomplete discussion helps to illuminate the epistemological argument’s role in the larger issue we’ve been considering. The argument suggests that our responses to interlocutors with seriously different views makes sense if moral discourse merits some interpretive skepticism and is hard to justify otherwise. By itself, this isn’t a conclusive argument, but, in conjunction with the others, it adds force to the

---

235 This is less contentious than it looks. When we’re doing normative ethics, our meta-ethical views—most of them, anyway—allow us to talk moral talk, classify acts as right or wrong, and so on. We can assume for the moment that these conversations are possible whether or not there really are properties such as rightness.

236 An additional difference between moral and theological disagreement might be that, in the theological case, we feel we have a better handle on the sorts of psychological stories that might be told in order to explain the theist’s commitment without legitimizing it, whereas we lack these stories in the case of, say, the committed deontologist (or consequentialist, perhaps). Thus the intractability is real in one case, apparent in another. This is a delicate argument to make because it suggests that we ‘psychologize’ the commitments of very smart people, and those people are apt to find this extraordinarily obnoxious.
suspicion that moralizing is specially problematic—even given the minimalist view that serves as the backdrop of this discussion. The phenomena I’ve gestured toward here—various features of our everyday moral experience—are just what we’d expect to see if moral discourse were problematic in ways that suggest the equivocal interpretation. This suggests the problem with the ‘innocence by association’ argument and the underlying continuity thesis: they fail to make sense of the special problems of moralizing.

5.8 Conclusions

The most serious attempts to come to terms with the univocity problem end in failure. Expressivism cannot make sense of conative variation and struggles to provide a non-circular account of moral language. Naturalistic realism, on the other hand, relies on resources from the philosophy of language that cannot make sense of the characteristic features of moral discourse. What lesson should we draw from these failures, especially in light of our arguments against the continuity thesis?

The appropriate lesson, I suggest, is humility: the problem is a difficult one, and our best attempts do not succeed. This is a valuable lesson, partly because it opens the door to further inquiry, and partly because the failure is an instructive one. It has focused attention on subtleties and nuances of moral discourse and its relationship with other forms of evaluation, and so it teaches us more about the requirements on any successful meta-ethical view.
BIBLIOGRAPHY


D’Arms, Justin, and Daniel Jacobson. 1994. “Morality, Expressivism, and the
Emotions.” *Ethics* 104: 739-763.


273


Railton, Peter. 1998. “Moral explanations and moral objectivity.” *Philosophy and


