

Return to Moral Twin Earth

DAVID MERLI
Ohio State University
Columbus, OH 43210-1365
USA

Roughly a half-century ago, R.M. Hare gave us a potent argument against attempts to account for the meaning of moral language in non-normative or 'descriptive' terms. The argument relies on the idea that in order to have genuine moral disagreement, we have to be talking about the same thing. Real disagreement requires agreement in meaning: if the words we use in our disputes mean different things, then we're just talking *past*, and not *to*, one another. Using this simple observation, Hare argues that if the meaning of an evaluative word such as 'good' were primarily *descriptive*, then groups with sufficiently different standards for applying 'good' wouldn't be able to enter into a real evaluative disagreement. But these disagreements are possible. Hence, he concluded, it's the evaluative meaning of 'good' that's primary — and any descriptive account is bound to fail because it doesn't capture the crucial element of *endorsement* that's central to normative language.¹

As an illustration of Hare's point, consider this parable from *The Language of Morals*. A Christian missionary arrives on a cannibal island, and finds that, in their language, the most general adjective of commendation happens to be the word 'good.' The Christian and the cannibals have very different ideas about, say, what makes for a good man, but they're able to understand each others' uses of 'good' and they can argue,

1 This kind of argument comes up in several places in Hare's work. See *The Language of Morals* (Oxford: Clarendon 1952) and 'A *Reductio ad Absurdum* of Descriptivism,' in *Essays in Ethical Theory* (Oxford: Clarendon 1989). Similar considerations seem to be behind Allan Gibbard's noncognitivism, especially in *Wise Choices, Apt Feelings* (Cambridge: Harvard University Press 1990), ch. 1.

sensibly and coherently, about judgments of goodness. This, Hare thinks, would be impossible if we understood 'good' descriptively:

If this were so [that is, if "good" were understood in terms of its descriptive criteria of application], 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. (*The Language of Morals*, 148-9)

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 univocality, 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.

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?² What's more, we've learned

2 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 Richard Boyd, 'How to be a Moral Realist,' in *Essays in Moral Realism*, G. Sayre-McCord, ed. (Ithaca: Cornell University Press 1988) for discussion.

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.³ 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.

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.⁴ In this paper, 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 make some suggestions concerning what the realist must do in order to resolve the underlying problem.

3 Terence Horgan and Mark Timmons, 'New Wave Moral Realism Meets Moral Twin Earth,' *Journal of Philosophic Research* 16 (1991) 447-65; 'Troubles for New Wave Moral Semantics: The Open Question Argument Revived,' *Philosophical Papers* 21 (1992) 153-75; 'Troubles on Moral Twin Earth: Moral Queerness Revived,' *Synthese* 92 (1992) 221-60; 'Copping Out on Moral Twin Earth,' *Synthese* 124 (2000) 139-52

4 For other responses, see Geoffrey Sayre-McCord, "'Good'" on Twin Earth,' in *Philosophical Issues 8: Truth*, E. Villaneuva, ed. (Atascadero: Ridgeview Press 1997) and David Copp, 'Milk, Honey, and the Good Life on Moral Twin Earth,' *Synthese* 124 (2000) 113-37.

I 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, *Language* and 'A *Reductio...*').
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 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.⁵ 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,' as '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

5 See David Brink, 'Moral Realism and Skeptical Arguments from Disagreement and Queerness,' *Australasian Journal of Philosophy* 62 (1984) 111-25 and *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press 1989).

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.⁶ 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 that 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.)⁷

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. ('Troubles on Moral Twin Earth,' 245-6)

6 See David Brink, 'Moral Realism and Skeptical Arguments from Disagreement and Queerness' and *Moral Realism and the Foundations of Ethics*; and Richard Boyd, 'How to Be a Moral Realist.'

7 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.⁸ 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’ (‘Troubles on Moral Twin Earth,’ 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.⁹

8 See Hilary Putnam, ‘The Meaning of “Meaning,”’ in *Mind, Language and Reality* (Cambridge: Cambridge University Press 1975).

9 I’m taking the central issue to be the preservation of disagreement, which may seem odd, given that the focus of Horgan and Timmons (‘Troubles on Moral Twin Earth’) 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.

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 *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 *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 as about morally substantive matters. Hence, *pace* the realist, it can't be those proposed natural properties that are essential to our moral terms.

Or so it seems. In the next section of the paper, 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.

II 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 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 why, 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.¹⁰ 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

10 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, 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.

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 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.¹¹ 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 grant-

11 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 ('Troubles on Moral Twin Earth,' 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.

ing, 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, 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 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.¹²

-
- 12 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 Earthian moral practice and some variant, we’d need to hear why *this* small change is enough to force the difference in reference, when, 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

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 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

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 III) 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.)

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 legitimize (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.¹³

13 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

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.¹⁴ 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

how reference is fixed are relevant to the practical or evaluative importance of these terms.

- 14 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, 'Moral Realism' and *Moral Realism*; Sayre-McCord, "'Good'" on Twin Earth'; and Nicholas Sturgeon, 'Moral Explanations,' in *Morality, Reason, and Truth*, D. Copp and D. Zimmerman, eds. (Totowa, NJ: Rowman and Allanheld 1985). See also Georges Rey, 'Concepts and Stereotypes,' *Cognition* 15 (1983) 237-62 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).

doesn't ensure that all agents will be motivated to act in accord with their moral 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 is 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, I'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.¹⁵ 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.¹⁶

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.¹⁷ If these

15 This point is made in Allan Gewirth, 'Positive "Ethics" and Normative "Science,"' *Philosophical Review* 69 (1960) 311-30.

16 See Nicholas Sturgeon, 'Moral Explanations'; Peter Railton, 'Moral Realism,' *Philosophical Review* 95 (1986) 163-207; and Brink, *Moral Realism* for discussions of this point. Michael Smith in *The Moral Problem* (Cambridge, MA: Blackwell 1995) also discusses convergence and its role in normative inquiry.

17 Think, for example, of Shelly Kagan's arguments in *The Limits of Morality* (Oxford: Clarendon 1989) 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

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 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 *cannot* 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.

III 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 T_c is the correct moral theory for us, we're supposing that it 'is discoverable through moral inquiry employing

between these starting points, which may look quite significant, are irrelevant to the question of shared reference that occupies us here.

coherentist methodology' (Horgan and Timmons, 'Troubles on Moral Twin Earth,' 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 T_d 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 H_2O ' 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.¹⁸

18 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

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.

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.¹⁹ 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,

ordinary thought is confused — though I don't follow him past that point. See Peter Unger, *Living High and Letting Die* (New York: Oxford University Press 1996) in particular.

19 Thanks to an anonymous reviewer for insisting on the importance of this point.

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.²⁰ A change in view just shows that we weren't *really* at the end of the day after all.²¹

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

20 Keep in mind, too, that it is essential to Horgan and Timmons' argument that Moral Twin Earth and Earth have *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.

21 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.

isn't attractive, because our idealized twin-moralists certainly *appear* to be rational and coherent. It's more sensible — and more 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.²²

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;

22 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 and Td 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.*

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.²³

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 reflec-

23 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*.

tion, 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.

IV 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.*²⁴ 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.

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.'²⁵ 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

24 This locution is suggested by Gibbard's attempts to describe a sense of 'bland, flavorless' endorsement that captures 'the last ought before action.'

25 This example, and a similar point, appear in Copp, 'Milk, Honey, and the Good Life.'

'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.²⁶

Reflecting on this phenomenon raises a challenge to our third argument. However it accounts for our disagreements with the twin-moral-

26 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.

ists, 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?

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.²⁷ 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. 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 expres-

27 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?'

sivism 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.²⁸ 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

28 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.

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 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 do 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.²⁹ 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.³⁰

We might think, with Hare and Gibbard, that some of the uses of language — even some that appear, on the surface, to be assertoric —

29 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.

30 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 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.

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.³¹ 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 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. For various reasons, I think this is unlikely, but these arguments 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.

V 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

31 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.

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.³²

Received: May, 2001

Revised: October, 2001

Revised: March, 2002

32 I would like to thank Justin D'Arms, Don Hubin, Nick Sturgeon, and William Taschek for comments on earlier drafts, and for many helpful conversations. This paper has also benefitted from the comments of three anonymous reviewers, and I am grateful for their suggestions. I'm also indebted to Pamela Hieronymi, Tyler Hower, and Julie Tannenbaum for insightful discussions and for their charity to alien interlocutors. Research on this paper was made possible by support from the Ohio State University and the Charlotte Newcombe Foundation.