6

Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument

Mark van Roojen

At the beginning of the twentieth century, G. E. Moore's open question argument convinced many philosophers that moral statements were not equivalent to statements made using non-moral or descriptive terms. For any non-moral description of an object or object it seemed that competent speakers could without confusion doubt that the action or object was appropriately characterized using moral terms such as 'good' or 'right'. The question of whether the action or object so described was good or right was always open, even to competent speakers. In the absence of any systematic theory to explain the possibility of synthetic as opposed to analytic identities, many were convinced this demonstrated that moral properties could not be identified with any natural (or supernatural) properties. Thus Moore and others concluded that moral properties such as goodness were irreducible sui generis properties, not identical to natural

I owe thanks to a number of people for discussion of this paper and related topics as I was working on the first draft. They include David Chalmers, John Gibbons, Terry Horgan, Harry Ide, Mark Kalderon, Clayton Littlejohn, Jennifer McKitrick, Joe Mendola, Sarah Sawyer and Russ Shafer-Landau. I also owe the audience at the First Annual Madison Metaethics Workshop in Oct. 2004 for comments and questions. People I specifically recall talking about the paper on that occasion include Sarah Buss, David Copp, Janice Dowell, Jamie Dreier, Terry Horgan, Mike Ridge, Russ Shafer-Landau, Michael Smith, David Sobel, Sarah Stroud, Mark Timmons, Pekka Vayrynen, Ralph Wedgwood, and Eric Weiland. If there are others whose comments I don't recall it says more about my memory than their comments. I apologize to them. I owe further thanks to two anonymous referees for Oxford University Press, Jennifer Haley, Russ Shafer-Landau (again!), and Pekka Vayrynen for helpful comments on still later drafts. Finally Mark Timmons kindly suggested I come up with a less modest title than the original.

properties (Moore, 1903: 15). Noncognitivists used the same argument to support the idea that moral judgments have an expressive function rather than a representational function. Their explanation for the failure of competent speakers to recognize the equivalence of moral predicates with other predicates was that these terms, unlike other predicates, did not serve to represent properties at all (Ogden and Richards, 1923: 125).

Contemporary philosophers recognize the possibility of synthetically (as opposed to analytically) identifying objects or properties referred to using different terms. We can discover that water is the same stuff as H_2O without being able to infer it from the meanings of the terms involved (Kripke, 1972; Putnam, 1975). Descriptive naturalists with respect to ethics capitalized on this to point out that the openness of Moore's question to competent speakers does not rule out the possibility of discovering that a moral property is a naturalistic property through empirical evidence not dependent on the expressions in question having the same meaning. The most sophisticated version of this sort of proposal has been offered by Richard Boyd. Oversimplifying just a bit, Boyd's idea is that moral terms can refer to a property in virtue of a certain sort of causal connection between the use of the term and the property, just as the term 'water' can refer to H_2O in virtue of a causal connection between H_2O and our use of the term water. Since it need not be transparent to a speaker what object or property bears the right sort of relation to her use of a term, a competent speaker can remain ignorant of the identity in question. Hence linguistic competence will not be sufficient to close open questions about the identities of the properties involved (Boyd, 1988).[1]

Terry Horgan and Mark Timmons (henceforth abbreviated as H&T) have constructed a neat argument intended to refute Boyd's theory and all similar theories. If they are correct, their argument together with Moore's original open question argument leave us to choose between noncognitivism on the one hand and non-naturalism on the other, the same options available to our predecessors seventy years ago. Given these options, and given that their argument highlights the commendatory function of moral language, the authors suggest that a sophisticated noncognitivism is the preferred choice.

Horgan and Timmons make explicit what they take to be commitments of causal theories of reference of the sort Kripke, Putnam, and Boyd use to explain the functioning of scientific kind terms, and which Boyd also applies to moral terms. These theories are anti-Fregean in the sense that term reference was not determined by a descriptive sense grasped by a thinker or

[1] Some of my wording on this page duplicates wording in my (2004) encyclopedia entry.

speaker and uniquely satisfied by the referent of the term. Rather, reference is determined by the existence of a certain sort of causal connection between the speaker's use of the term and the referent. However, Horgan and Timmons argue, the thought experiments used by proponents to motivate such causal theories generally show that competent speakers are in fact aware that the terms refer to whatever stands in the appropriate causal relations to the use of a term when the causal theory is the appropriate theory of reference for that term. If this is right, then applying the same theory to moral terms would suggest that a competent speaker should at least tacitly know that the relevant moral term refers in virtue of the right sort of causal connection. Thus speakers' intuitions about whether or not the term refers to whatever has such a connection should be probative with respect to the truth of the theory of reference in question. Horgan and Timmons then construct a clever example involving ''Moral Twin Earth'' to generate intuitions in conflict with the assumption that causal regulation determines reference for moral terms, and conclude that the theory is false. Indeed, they claim their argument sounds the death knell for all descriptivist versions of naturalism.

The argument has spawned a number of replies, each designed to show that the example does not refute naturalism. Many of these replies argue that the target theory survives in the face of the Moral Twin Earth example (Geirsson, 2003). My argument takes a somewhat different line. I think the target semantic theory as understood by Horgan and Timmons is in fact refuted by the Moral Twin Earth example. And I think the internalist upshot of their argument—that a commendatory function is a constitutive feature of genuine moral discourse—is also correct. However, I will argue, we can construct a successor semantic theory to the one proposed by Boyd which takes advantage of his real insights while supplementing them in various ways. This theory is not refuted by the Twin Earth example and in fact incorporates the internalist[2] upshot of the example while classing moral property terms as genuinely referring expressions.

[2] Unfortunately, philosophers have used the words 'internalism' and 'externalism' to mark a number of different philosophical distinctions in different subfields of philosophy and at least two of these are relevant to this paper. Here I mean internalist in the sense which requires a necessary connection between accepting a moral judgement and being motivated to do what it recommends. At other points I will be defending views which are 'externalist' in a sense that does not contrast with this one, but which instead contrasts with claims that the meaning of a term and the contents of thoughts using that term are entirely determined by individualistic properties of the speaker or thinker using the term. Internalism in that contrasting sense is the view that meaning and content are determined by facts internal to the thinker's head skin. I hope that context will make clear which sort of internalism or externalism I have in mind.

The Dialectic, the Target Theory, and the New Objection

The old open question argument

The Moral Twin Earth argument is embedded in a dialectic that begins with G. E. Moore's open question argument. Moore's argument purports to show that goodness could not be identical with any naturalistic property. He challenged his readers to provide candidate natural properties to identify with goodness. He claimed that for any such candidate property it was open to a person capable of having thoughts involving the property to wonder whether it was in fact identical to the property goodness.[3] The fact this was possible could then be used as a premise in two sorts of arguments purporting to demonstrate the property goodness was not in fact the candidate property. One version rests on the assumption that an analysis ought to support substitution of one term for the other in any meaningful sentence. If goodness is to be analyzed as the property in question, it would then be appropriate to substitute the term for the property in question for any occurrence of the term 'goodness', including sentences expressing one's uncertainty about whether the property was itself goodness. But then one would be asking whether the property in question was itself—using the very same term to pick out the property. For example, "I wonder whether pleasantness is goodness," would become, "I wonder whether pleasantness is pleasantness." Since according to Moore the former makes sense and the latter does not, pleasantness is not a correct analysis of goodness. And thus, Moore believed, for any natural candidate. If we now add the assumption that identity claims are analytic—that is underwritten by correct analyses of one term employing another term for the thing—we arrive at the conclusion that goodness is not identical to any natural property.

A second version of the argument relies on Leibniz's law. It takes the fact that we might sensibly ask ourselves whether pleasantness is good but not sensibly ask ourselves whether goodness is good to show that we are in doubt about the goodness of pleasantness but not the goodness of goodness. Thus pleasantness has a property goodness lacks, and by Leibniz's law the two cannot be identical.[4]

[3] Moore actually used the term 'good' to refer to the property in question, but I think it is more natural to use the term 'goodness'. Moore's text is somewhat schematic and does not provide all of the premises needed to construct a valid argument, so I'm providing a reconstruction which I think is faithful to his intentions. More important for my purposes, I think my reconstruction captures what people took from his text and were persuaded by. See Moore (1903: 15–17).

[4] For textual support for this interpretation consider, "whoever will attentively consider with himself what is actually before his mind when he asks the question 'Is

The argument was widely influential in convincing people that no natural property was identical with either goodness or rightness. But in hindsight it is sometimes hard to see why the argument had such influence. Similar worries could be raised about almost any other informative identity claim inasmuch as it might be reasonable for someone to consider it a subject for investigation. Discussions of the paradox of analysis and Frege's puzzle should already have made this clear. And the first version of the argument contains a number of assumptions which we have reason to doubt, perhaps most importantly that identity claims must be analytic as opposed to synthetic. It is possible that those impressed by the argument thought the sorts of identity claims that philosophy was after—claims that could tell us about the real nature of an object or property—would have to be a priori, even if not all identity claims are. The Kantian idea that all claims with necessary modal force must be a priori plus the necessity of identity might underwrite the assumption. Such assumptions may make the open question argument hard to resist—even with seeming counter-examples ready to hand—at least until subsequent work on the semantics of natural kind terms made clear how identities might be matters for empirical investigation. Even now the second version of the argument employing Leibniz's Law remains somewhat persuasive until one is in a position to say which of the premises is false and why.

The target theory

This is where views which Horgan and Timmons dub "New Wave Moral Realism," come in. The new wave theorists, most notably Richard Boyd, provide a semantics for moral terms that explains how identity claims though necessary could be synthetic. The new wave theories also explain how we could sensibly have the sorts of doubts Moore's open questions express, even with respect to identical properties. The theories explain why Moorean doubts do not require us to give up Leibniz's law in order to defend the identification of moral properties with natural properties. These new wave theories are modeled on anti-descriptivist and externalist theories of meaning and content determination suggested by the work of (among others) Kripke (1972), Putnam (1973, 1975) and Burge (1979).

These theories were in part constructed to explain how any identity claim could be open to rational doubts on the part of even competent speakers of

pleasure (or whatever it may be) after all good?' can easily satisfy himself that he is not merely wondering whether pleasure is pleasant." (Moore, 1903: 16). For a nice discussion see Kalderon (2004).

the language which is used to express the identities. The basic idea was to deny that the relevant terms functioned as disguised descriptions known by competent speakers, sufficient for determining the referents of the terms. Without this assumption there is no reason to think linguistic competence makes available for each term some a priori equivalent description. It should then be no surprise that competent speakers could doubt any candidate identity, or that it should be a synthetic matter when identities are established.

Rejecting a descriptive picture of reference determination for a class of expressions carries with it the need for a replacement account of reference determination for those terms. Proponents of these externalist theories obliged by suggesting that an appropriate socially transmitted chain of causal and epistemic influence might be sufficient to secure reference for many classes of terms. The idea can be filled out in a number of particular ways. One of these is proposed by Richard Boyd and applied to moral terms.

Boyd takes moral terms to have their referents determined just as the referents for natural kind terms are determined. He thinks that referents of natural kind terms are determined by a causally composed feedback loop from the referent to the use of the term. He writes:

> *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 will typically include the existence of procedures which are approximately accurate for recognizing members or instances of *k* (at least for easy cases) and which relevantly govern the use of *t*, the social transmission of certain relevantly approximately true beliefs regarding *k*, formulated as claims about *t* (again excuse the slight to the use-mention distinction), a pattern of deference to experts on *k* with respect to the use of *t*, etc. . . . When relations of this sort obtain, we may think of the properties of *k* as regulating the use of *t* (via such causal relations), and we may think of what is said using *t* providing us with socially coordinated *epistemic access* to *k*: *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not to other kinds (properties, etc.) (Boyd, 1988: 195)

Thus, according to Boyd, a moral term such as ‘right’ will refer to whatever property causally regulates in the above-described manner our use of the term ‘right’.

The theory is ready-made to explain the possibility of open questions. Non-experts can have a thought with a certain content in virtue of being members of a speech community, and yet not know what the experts know. And it is not necessarily obvious to even experts, let alone ordinary speakers of a language, what naturalistic features the relevant property might have.

They may be largely in the dark about the nature of the property which lies at the other end of the causal-regulatory feedback loop, although over time they can expect to learn more about it. Given that the most competent speakers of the language may not know that the property which plays this role can truly be identified via some naturalistic description or other, such speakers may have doubts about the identity of the moral property in question and the property naturalistically described. Hence the possibility of open questions regarding the identity of such properties, even for experts and even when the right-hand side of the identity statement open to question picks out the property using features which are essential to it and thus pick it out rigidly.

Moral Twin Earth

It is this theory which Moral Twin Earth is designed to refute. H&T believe that if Boyd's proposal is true semantically competent speakers should at least tacitly recognize its truth. Thus, even if such speakers do not know which natural property regulates their moral terms, they should at least tacitly know they refer to the property which appropriately regulates their use. Further, H&T believe we should be able to elicit this tacit knowledge by presenting a speaker with the right sorts of thought experiments and asking whether the people in those examples are using the words in question to refer to the same thing we refer to. This is how it works with natural kind terms like 'water'. Putnam and Kripke argued against descriptive theories and for their own theories by eliciting audience responses to various scenarios. The scenarios were devised to generate verdicts concerning the referents of various terms which vindicate the externalist theories. For example, these theorists elicited verdicts about the meaning of 'water' on Twin Earth, a planet otherwise like ours except that XYZ takes the place of H_2O on Earth. Speakers' agreement that the term refers there to XYZ and not to H_2O is crucial in vindicating causal regulation as a component in the determination of meaning and reference. Thus Horgan and Timmons suggest we should expect such tacit knowledge whenever a term refers in virtue of similar causal regulatory roles (1992*b*: 162–3).

Moral Twin Earth is constructed to test the hypothesis that moral terms work in this way. We are to imagine a planet much like Earth on which people use moral terms such as 'good' and 'right' in much the same way we do. People on this twin of Earth apply these terms to persons' actions and institutions; they take the "goodness" or "rightness" of an option to be important, and they are normally disposed to do what they believe is "right" and to choose what they take to be "good". So on the surface

Earth and Twin Earth are indistinguishable. At the same time we are to imagine that one natural property causally regulates our use of the relevant moral term here on Earth, whereas a different property causally regulates the use of the same term on Twin Earth. The properties are similar enough to account for common ways the terms operate on the two planets, but they are still distinct. A bare minimum of subtle but real differences in the psychologies of the relevant populations is allowed, so that somewhat different properties can play the same roles for the groups on each planet (Horgan and Timmons, 1992*b*: 164–5).

H&T then ask us to decide whether we would translate the moral terms on Twin Earth with our counterpart terms, or not.[5] The verdict that we should will cause problems for Boyd's theory. If such translation is correct our counterparts must mean what we mean by the terms. But if our terms and theirs mean the same thing, it cannot be that the terms on Earth and on Twin Earth designate different natural properties. Thus intuitions that the two populations are in genuine disagreement would indicate that the terms cannot function to designate whatever natural property regulates the relevant population's use of the term, since by hypothesis these are different on Twin Earth than they are here.

Unfortunately for Boyd's theory, competent speakers do have the intuitions that Horgan and Timmons seek to elicit with their example. Most people who have read their article seem to agree that the two populations address one debate about moral goodness, and that they are not talking past one another in virtue of using words with different referents. On this basis, Horgan and Timmons claim that Boyd's theory stands refuted, along with other similar cognitivist theories.

Taking stock

I myself have the intuitions Horgan and Timmons expect when they present the Moral Twin Earth example, and I am inclined to believe they tell us something correct about our moral terms. The speakers in the two communities are using moral terms with the same meaning, so that their dispute over what to do is a real dispute. If Boyd's proposal is in conflict with this, something must be wrong with it. Granting that, it is worth

[5] Readers should be reminded here of Hare's (1952) famous missionaries and cannibals argument which shares many features with the argument here. Moral Twin Earth aims at targets not on the scene when Hare formulated his argument, and hence involves setting up the thought experiment in a somewhat different way than he did. The similarities are noted by Timmons (1999), though the Twin Earth argument was not consciously patterned on Hare's.

carefully considering how the example causes trouble for Boyd's approach and what we can learn from the example.

A First Bit of Instructive Complication

We might begin by looking closely at what sort of variation there can coherently be between Earth and Twin Earth. Only certain sorts of variation with respect to causal regulation can be built into the thought experiment consistent with the similarities between the planets. If the populations of each planet really constitute linguistic communities, the use of a word by one member of a community will play a role in explaining the use of that word by other members of the community. Obviously each member of a community does not miraculously and independently coin a term orthographically and phonologically identical to those used by her community to express the same contents she will express with her new term. Terms are passed on from one member of a community to another. People learn them from their parents, friends, and neighbors and repeat the terms they learn. All of this requires causal interaction and it is these causal mechanisms that are responsible for people speaking the same language. Thus when Horgan and Timmons ask us to imagine a place different from ours with respect to the causal regulation of terms they cannot ask us to build the difference into this part of the causal chains from properties to speakers.

Rather they must ask us to imagine some variation leading to the original use of a term with a certain meaning such that that same meaning can then be passed on to subsequent users. We need examples in which two different properties stand in the same relation to original meaningful usage which don't change things so much that the kind of relations between the property and the speakers also changes. To see if this is possible we need to pay attention to the kind of relation Boyd postulates and ask whether two different properties could stand in this very same sort of relation to the speakers in question. Boyd's theory requires a causally efficacious feedback loop, from the referent of a term back to our use of the term such that the referent itself plays a role in causally explaining how we come to modify our beliefs expressed using the term so that they become truer over time.[6] It certainly seems possible that two similar but not identical

[6] At least that is how Horgan and Timmons interpret him and I think that this interpretation is probably fair. If you look at the long quotation about regulation taken from Boyd he is less than fully explicit about this. Still in saying that the use of a term is "causally regulated by" a certain property he seems to suggest that the property or its instances causally impact the use in question.

properties would be suited to playing similar causal roles with respect to a community's use of a term, and hence there seem to be possible scenarios in which each one is related as Boyd suggests the referent of 'good' must be.

Actually the issue is a bit trickier than at first it seems. Boyd puts an epistemic constraint on the nature of the causal relation such that too radical switches in the properties at the end of the causal chain might all by themselves turn a relation which meets Boyd's specification into one that does not. The causal regulation must be such as to make the beliefs of the community truer of the referent over time. Thus each of the properties must be such that what people come to believe as a result of the relevant causal relation is more nearly true than what they believed previously. Not just any property which caused us to modify our beliefs will do. Some properties may play a causal role in belief formation and yet not be otherwise such that the beliefs formed about them will be true or more true than what was previously believed. Thus the causally regulating property in the twin scenario must be sufficiently similar to the property playing the relevant role in the actual scenario that as our beliefs about rightness or goodness evolve they become more true of the hypothetical twin property as well as of the property actually playing that role.

Suppose things are as Horgan and Timmons stipulate: property A (the one which fits the role determined by consequentialist theory A) regulates use of the term 'right' on Twin Earth, and property B (the one which fits the role determined by nonconsequentialist theory B) regulates the use of 'right' on Earth.[7] But suppose also that nonconsequentialism is correct and that B is the correct version of nonconsequentialism. On the theory we are testing, the proposition expressed by 'X is right' on Twin Earth is that X has property A, and the one expressed on Earth is that X has property B. Now on both planets the population believes that right actions are the ones that ought to be done and make the most sense to do, or at least most people believe this. This is to say, those on Twin Earth believe that actions with property A ought to be done and make the most sense to do. And those on Earth believe that actions with property B ought to be done and make the most sense to do. Furthermore they

[7] I switch the target term from 'good' to 'right' because it bypasses a problem with the example as formulated by H&T. Nonconsequentialists may not differ from consequentialists over what they believe to be good, but over how the relative goodness of an outcome determines the rightness of actions. Typically nonconsequentialists deny that we should always do what leads to the better outcome, though it is possible to model such views using an agent-relative measure of goodness. We can overcome this problem by switching the example from goodness to rightness, since consequentialists and nonconsequentialists do disagree about which actions are right.

have come to believe this as a result of regulation by A and B respectively. But, on the assumption that B is the correct moral theory, only those on Earth will have made their beliefs truer by taking this commitment on board.[8]

Whether or not this result is fatal to the coherence of the example would seem to depend partly on our metric for judging when beliefs are more or less true. For no doubt some of our beliefs expressed using the term 'right' will be true of property A and it may have played an appropriate regulating role in generating those beliefs. Depending on how we weight the true beliefs as against the false ones, perhaps we will still want to say that A too has regulated our beliefs expressed using the term 'right' so that they become truer of A as we move along. Perhaps not.

It may not matter to the overall point what we decide. For we can modify the H&T example slightly. Suppose that on one planet the moral term is not causally regulated in the appropriate way by any property at all. The term which is not appropriately regulated will have to possess a different semantic value from the same term appropriately regulated by some property, at least it must if we take the analogy with the direct reference theorist's treatment of natural kind terms seriously. A moral term not regulated in the appropriate way by any property will likely have to be treated as something like an empty name, or a purportedly referring expression for which there is no referent. Empty terms of this sort are not synonymous with non-empty terms of the same sort.[9] If the correct intuitions about this case are the same ones H&T elicit with their original case, we can bypass the worries and reach the same result. Causal regulation of the relevant sort does not determine the appropriate semantic values for moral terms.

Still there is something to learn from the near failure of the example. It may be possible to specify the kind of regulation that determines reference in such a way as to rule out twinning the relation. It may be that the addition of epistemic constraints on regulation—similar to the requirement that the regulation makes the beliefs truer—would be the sort of specification that might accomplish such a task. I will come back to this idea later.

[8] If you're worried that it is question begging to posit that one or the other of these theories is true, remember that the Moral Twin Earth thought experiment is first and foremost a challenge to the semantic theory embodied in Boyd's proposal. The conclusion that moral theories cannot be true had better not be an assumption of the argument on pain of circularity.

[9] Just as our term 'Santa Claus' does not mean the same thing as a phonologically and orthographically identical term referring to a fat guy in a red suit on another planet. For more on this see Kripke's (1973: 156–8) discussion of Sherlock Holmes.

A Second Bit of Complication

A second area where closer examination is instructive encompasses the features of the thought experiment which underwrite our attribution of a common meaning to the terms. H&T propose that the crucial feature has to do with the action-guiding nature of the judgments in question (1992*c*: 170). We can assess this suggestion by looking at the way Moral Twin Earth is introduced to us. We are told that (1) the terms 'good' and 'right' are used to reason about considerations bearing on well-being, that (2) people there are normally disposed to act in ways corresponding to what is 'good' and 'right', that (3) these people take the goodness or rightness of options to be important, even over-ridingly so, in deciding what to choose, and that (4) the terms apply to actions, persons, and institutions (1992*c*: 164). Since this list summarizes our grip on Moral Twin Earth's similarity to Earth, we should expect to find the grounds of our sense that the two populations mean the same thing in this rather short list of common features. Items (2) and (3) spell out the action-guiding character of moral judgments, and it does seem reasonable to conclude they are responsible for our regarding the terms as synonymous with ours.[10] If we have any doubts concerning this, we might ask ourselves if the features listed in (1) and (4) by themselves would be enough to sustain the verdict. It seems they would not and hence it seems that the internalist features of moral practice are essential.

But I should register my doubts that Horgan and Timmons have listed *all* of the features of moral terms which account for our ascribing the same meanings to any terms used in the same way. In addition to those they list, our moral terms are used in such a way that their application supervenes on the distribution of non-moral terms to the items up for appraisal (as H&T recognize elsewhere[11]). An action-guiding appraisal of actions, persons, and institutions that did not depend on the otherwise specifiable features of these items would not be moral judgment as we know it, and we might be loath to translate terms used in this unguided way with our moral vocabulary. I suspect most readers of the relevant papers just assumed in the spirit of charity that this feature of our practices was among

[10] The conclusion is reinforced by the similar verdict rendered when we are asked to consider Hare's missionaries and cannibals.

[11] Horgan and Timmons use almost identical language to describe the similarities between Earth and Moral Twin Earth in each of their papers employing the example (1991: 459; 1992*c*, 246–7; 1992*b*: 164.) Only the paper explicitly about supervenience (1992*c*) mentions it among the similarities. We might think of them as intending to include it in the others as well.

the similarities intended by H&T. In any case we should add supervenience to their list.

A couple of other features of our moral judgments should probably also be presumed in the same spirit. The terms 'good' and 'right' are our most general terms of moral appraisal; presumably the corresponding words on Moral Twin Earth are similar. And a full description of Moral Twin Earth had better add something to distinguish the roles of 'good' and 'right' so that we have some reason to translate each with the phonologically and orthographically identical terms. That said, it does seem that the action-guiding character of a set of judgments, together with these additional features, is sufficient for using a term with the same meaning as our terms. This is a second lesson to take from the example.

A Third Bit of Complication

This leads naturally to a third insight we might glean from a closer examination of the troubles Moral Twin Earth does and does not cause for new wave theorists: The kind of internalism supported by the example is social in nature and not to be interpreted in an individualistic fashion. Moral terms are described by H&T as being only for the most part action-guiding. They suggest that people are normally but not always disposed to do what they regard as right and to promote what they believe good. In this way of setting up the case they are certainly correct, as the literature debating the merits of internalism has made salient.[12] The connection with motivation is still necessary insofar as it is needed to sustain the verdict that the 'right' on each planet means right. But what is necessary is that this is the normal situation in the population, not that any given member of that population be so motivated.

This sort of action-guiding character actually vindicates one feature of moral practice that the new wave realists wish to emphasize. The content of sentences and thoughts expressed using moral language is a function of the community of which one is a member. People in a community can express certain thoughts and say what they say despite their not being typical members of the community in the ways they make those judgements or use those words. The original version of that idea was used in part to explain how members of a community could use a word with its standard meaning

[12] Stocker (1979) and Brink (1986) present examples that seem to show that not all competent speakers of moral language need be motivated by their moral judgements. Smith (1994) and Dreier (1991) suggest ways of accommodating these cases within a defeasible internalist view.

despite dissenting from "constitutive" truths about the referents of those terms. But the Moral Twin Earth example and our reaction to it suggests that there can similarly be constitutive features of moral practice which are not universally observed by even those competent enough to use moral terms with their ordinary meanings and to have thoughts which would be expressed using those terms. Even if there are some individually necessary conditions on possessing moral concepts and using moral terms with the meanings we do, not every constitutive feature of moral practice needs to be exemplified in that way. Some features, in particular the features H&T use Twin Earth to illuminate, need only be exemplified by subpopulations of the community of which one is a member.

A New New Wave Theory

I think we can accept all of the foregoing morals highlighted by the Moral Twin Earth example while retaining the central ideas of Boyd-style theories, namely the parts of the theory that enable it to make sense of open questions.

From a certain point of view, it can be surprising that *every* proposed identity between a moral term and any candidate property picked out in non-moral terms is subject to doubt.[13] It is this general thesis that grounds the open question argument and which the new wave theorists are in a good position to explain. They explain it by making the referent of a term a function of some fact or facts of which competent speakers may be unaware. But to be a general explanation of the purported fact that for any description a speaker can regard it as an open question whether it refers to goodness or rightness, the theory has to allow such ignorance in a pretty strong way. Thus the theory must deny there is any description sufficient for uniquely picking out the referent of these terms of which competent speakers must be aware. This rules out not only various first-level substantive

[13] In a reply to a paper of Sayre-McCord's which is itself a response to Horgan and Timmons, Ernest Sosa wonders how anyone could have found the open question argument at all surprising (Sosa, 1997: 304–7). For on any theory there will be synthetic identities open to doubt by competent speakers. That the President of the United States is the George Bush can be doubted by competent speakers, but that by itself does not show that he is not. (Evidence from Florida would be more telling than any open question argument in calling that identity into question, whether or not the identity in fact holds.) Puzzlement about the power of the open question argument should diminish when we remember that Moore claimed that *all* identity statements involving moral terms are open to similar doubts by those competent with the concepts. In the absence of an alternative proposal for what property terms mean it can seem hard to find a content for them without invoking some description or other. What the various "new wave" theories contribute is precisely that sort of alternative.

characterizations of the referent, but also descriptions which embody the semantic theory proposed to explain how words come to refer to their referents. In other words if competent speakers can doubt every descriptive analysis of moral terms, they must be able to doubt whether the term 'good' refers to whatever property regulates use of the term 'good' of people in the same community as the speaker.[14]

One upshot of this approach is that the meaning of a moral term should not be equated with any such reference-fixing description. For a speaker can count as competent enough to use the term meaningfully and to have thoughts of the sort we would use the term to ascribe without believing that any such description is satisfied by the referent of the term. Yet the term is supposed to have a determinate referent, in the present case a property. The causal regulation thesis serves to provide that determinate referent for Boyd's account. It provides a fact or set of facts about the speaker's use of the term which narrows down the available candidates for the referent from those which would be available by limiting ourselves only to properties satisfying descriptions accepted by the speaker. In fact the candidate may not be one of those satisfying what the speaker would assent to, since the speaker may well have false beliefs about the referent. Boyd's theory is thus a version of externalism about content. Facts about a speaker's

[14] Let me be clear, since it might seem that the claim is too strong to be credible. The theories deny that any description sufficient for uniquely picking out the referent is such that competent speakers must be unable to doubt it. Thus, any identity claims or analyses involving the referent will fall into the dubitable class. But the theories need not and do not deny that a speaker must have some knowledge about the referent to count as competent. Perhaps there is some fact about rightness that everyone possessing the concept must understand, or perhaps there are a number of different facts or sets of facts each of which is sufficient knowledge for competence. The point is only that such knowledge will not be sufficient to ground an identity claim of the sort involved in an analysis. H&T imply that the use of thought experiments to support the new wave theories is in tension with this claim, for they think that the experiments show that the question Q7 is open, and that its openness counts against Boyd's theory: "Q7 Given that the use of "good" by humans is causally regulated by natural property N, is entity e, which has N, good?" (H&T, 1992; 163). But if their argument really turned on the openness of this question, H&T would have gone through needless work to conceive of Moral Twin Earth. Q7 is one of the many open questions employed by the old-fashioned open question argument, and the complaint here is a version of that argument. If Q7's openness to competent speakers by itself refuted Boyd, H&T should merely have presented themselves as competent speakers who doubt the semantic theory postulated by Boyd and hence doubt that goodness is the property that causally regulates our use of the word 'good'. Lucky for all involved, the Moral Twin Earth example works independently of the original open question argument. It turns on the sufficiency of a population's practices to underwrite their using a word with the same meaning we do in circumstances where Boyd's theory predicts that the terms are not being used with the same meaning. It doesn't turn on the fact competent speakers can question Boyd's analysis.

usage and environment which can be unknown to that speaker contribute to determining the truth conditions for her thoughts and utterances employing the terms in question.

The particular external facts on which Boyd relies to determine reference have to do with which item in the world causally regulates a speaker's use of the term in such a way that people's beliefs about the referent of the term would over time become approximately true of the item in the world. Causal regulation by itself is not enough; the regulation has to also satisfy the epistemic constraint that it lead to truer beliefs over time. As we have already seen this makes it harder to construct the sort of Twin Earth scenario needed to refute the theory. I propose we see what we can do with just these sorts of epistemic facts, without requiring that the process in question be causal.

Suppose that (mimicking Boyd) we say something like this:

> *Roughly*, and for nondegenerate cases, a term *m* refers to a property *p* just in case there exist epistemically relevant procedures 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 procedures will typically require some members of the community to have an ability to recognize instances of *m* (at least for easy cases) which are employed in making judgments express using *m*, the social transmission of certain relevantly approximately true beliefs regarding *p*, formulated as claims about *t* (again excuse the slight to the use–mention distinction). When relations of this sort obtain, we may think of what is said using *t* as providing us with socially coordinated *epistemic access* to *p*. We have this sort of socially coordinated epistemic access when a good number of the beliefs about the referent of 'm' in the social context non-accidentally track what is going on with *p*. In other words, we have it when a good bit of what people say and believe using 'm' is approximately true, and it is said and believed because of how it is with *p*: *m* refers to *p* (in nondegenerate cases) just in case the socially coordinated use of *m* provides significant epistemic access to *p*, and not to other properties—that is when what is said using *m* expresses knowledge about *p*.

This modification of Boyd's proposal replaces causal regulation with epistemic regulation of roughly the sort that divides merely true beliefs from knowledge. We can think of this as a generalization of Boyd's idea regarding causal regulation. Knowledge requires that our beliefs be non-accidentally true. If our beliefs about some matter count as knowledge about some thing, they must not only represent matters correctly but we must have them because things are thus with the thing about which we have knowledge. For ordinary natural kinds, facts about those kinds can only explain our knowledge of those facts if we have causal contact with the kinds. When you and I believe that water is heavier than a similar volume of sawdust, our belief is not a lucky accident or a guess precisely because we have had contact with and thereby know something about water's weight. While XYZ on Twin Earth is also heavier than sawdust, there is no sense in which our believing that could constitute knowledge. For we would need empirical access to the stuff in order to know this.

Thus for many properties a causal connection to instances of its instantiation may well constitute the epistemic relation needed. But not every required epistemic relation will be such a causal relation. Some epistemic connections are a priori. It can be right to say that a person believes a necessary truth about an abstract entity *because* it is true, without their belief being caused by that entity or that truth. The 'because' here stands for some sort of relation that need not be causally implemented. I'm not sure I know what the relation is, but I think the notion is one we sometimes invoke and one which plausibly has an epistemic role to play (Nozick, 1981: 287). The above statement of the revised Boyd-style view employs it to cash out a more general version of his suggestion that the relation between a term and its referent might necessarily be epistemic.

I need to add one further refinement to this suggestion, one I will argue for in the next section. Some properties are more eligible candidates for the referents of a term than others. We can think of the most eligible candidates as the most natural properties available for the sort of endeavor speakers engage in when they are using a term to talk about a property. A term 'm' refers to the most eligible candidate property consistent with the constraint that speakers of the language who use 'm' use it to express knowledge about *p*. Since moral thought and talk has distinct purposes from other disciplines the most eligible properties will be natural moral kinds.[15]

We are now in a position to sketch how the revised account helps us handle the Twin Earth scenario. Since the Twin Earth argument is a semantic objection to moral realism, we can make various realist metaphysical and epistemic assumptions without begging questions against it. I'll suppose the following. Some actions really do make sense to do whereas others don't and some make more sense to do than others. Actions make more or less sense to do in virtue of other features that they have, and often these features are naturalistic. Furthermore, people are not entirely ignorant of this. In particular people sometimes do things because they make sense to do and when they do so it constitutes acting from knowledge.[16]

[15] Geoffrey Sayre-McCord (1997) in his paper on the Moral Twin Earth argument has suggested that we need to substitute moral kinds for natural kinds in the Boyd account. But without also modifying the story about the kind of regulation involved in the Boyd-type theory (from causal to epistemic) I don't think the suggestion by itself does the trick.

[16] I take the idea that we are talking about what it makes sense to do from Gibbard (1990: 49).

On these assumptions, people might well want/need a vocabulary to talk about what makes sense to do and what does not. Suppose such a vocabulary develops, and people use it in ways that somewhat track the facts about what makes sense to do, and that the judgments vary with some of the naturalistic features of the actions they are assessing. Suppose further that their judgments using a term ‘r’ come to more and more closely track what in fact makes sense to do, and that this is not accidental. Suppose still further that they tend to do what they label with ‘r’. We might be justified in thinking that they use this term ‘r’ as they do because of facts about which actions have the property of making sense to do. If that is correct, the proposed semantic theory for moral terms would license us to conclude that ‘r’ refers to the property actions have when they make sense to do. (Call this property **R**, to save space.)

Two communities could stand in such a relation to **R** without agreeing on all of their judgments. And they could stand in this relation to that property even at the limit of enquiry. One or both communities might be disposed to get things somewhat wrong about what makes sense to do, even though most of their judgments are correct and even though the correctness of these judgments is to be explained by their having knowledge of **R** which they express using the term ‘r’. In other words, they could be so disposed that they were causally regulated in such a way as to track a property somewhat divergent from **R** at least over a certain range of cases.

Yet, and this is the important point, if the causal factors that explain their judgments not accurately tracking **R** count from an epistemic view as mistakes, they do not for all that vitiate the claim that **R** is the referent of their term ‘r’. For our modified theory counts only epistemically relevant regulation as serving to determine the referent of a term. How do we determine when people are disposed to make judgments that track epistemically relevant factors as opposed to when they are disposed to make mistakes? That is a difficult matter, but one factor in this determination is whether they are disposed to judge in ways that track relevantly natural kinds. When two populations both use overlapping action-guiding terminology but are disposed to diverge over a range of cases, the population whose judgments most closely track the relevant moral kinds is getting things more right than the other population. But for all that, the proposed semantic theory will attribute the same content to their judgments and suggest that one population is making a mistake.

This is what I suggest is going on with the Moral Twin Earth example. The two populations are both tracking well enough what makes sense to do. At most one population might be disposed to get it right in the long run. This is how I take H&T’s stipulations regarding causal regulation in the example. But both populations may be using ‘right’ to refer to what makes sense to do, since both are using the term because they have (some) knowledge of what makes sense to do and they express that knowledge using the term 'right'.

Using Natural Kinds to Refine the Proposal

That anyway is the capsule summary of the position. In this section I'll give more detail on using natural kinds to winnow down the eligible candidates for term reference and motivate that component of the proposal. Then I'll provide some more commentary on how that helps explain the intuitions about the Moral Twin Earth cases.

We need some way to narrow down the range of referents beyond what even a causal regulation account could do on its own. For any finite series of events in which a particular property is instantiated, there are many properties that are instantiated by each of the events in the series. At least there are on the relatively liberal conception of properties where new properties can be created by conjoining and disjoining old properties. I take it that when we say that a property causally regulates people's use of a term, we mean that to be shorthand for the idea that events in which that property is instantiated play a role in causing the beliefs that people have and express using that term. The problem arises because each of these events involves the instantiation of many properties and no matter the number of instances involved there will be multiple candidates to be the designee of the term in question.

So we need to get from the plethora of different properties, the instantiations of which have causally interacted with a community of speakers, to the specific property which is the referent of the term. There are two suggestions one might immediately think of to try to narrow the range. One would be to limit the candidates to those properties which are relevant to causing people to have the beliefs that they do. We might hope that the instantiation of massiveness is causally relevant to people's beliefs about what is massive, but that the Cambridge property of being massive or in New York would not be. But there are problems with this approach that lead me to think it cannot be right. It might be that people are prone to a certain sort of error in certain circumstances, so that the absence of those circumstances is also causally relevant to our forming the beliefs that we do. Being massive and not in near zero gravity might be a causally relevant property when we are thinking about what causes our beliefs about massiveness. Yet we want the account to allow us to express beliefs about massiveness not just massiveness in higher gravitational environments.

Furthermore, being prone to error can bring in another sort of problem. We do in fact form beliefs in circumstances where we do make errors. We don't want it to turn out that our beliefs are true of some more complex property constructed out of the property we would ordinarily take ourselves to be designating along with the property the instantiation of which on a given occasion led us to make the error that we do. We want our beliefs about loudness to be beliefs about loudness, not beliefs about loudness or high distortion. But given the way the world actually works many of us mistake high distortion for volume. It could be that the beliefs that we normally express using the word 'loud' are causally best explained by citing our experiences of either loudness or high distortion. And it might even be that they are as true of that disjunctive property as they are of the non-disjunctive property of loudness.

These shortcomings may lead one to view a second suggestion as more promising. Why not allow the dispositions of the speaker's community to determine the property in question? But once again we face a problem. For sufficiently anomalous cases our community may be disposed to error about the extension of the property. Yet that would seem to be ruled out if the actual dispositions of the community to make judgments about the extension of a property made it the case that our judgments were about a property with that extension. This is the analogue for properties of the objection Kripke runs against the idea that dispositions can be used to determine which function we have in mind when it looks like our actual behavior is consistent with any number of candidate functions. At some point our dispositions run out (Kripke, 1982: 22–39). What I think this shows is that no causal or dispositional approach can all on its own uniquely determine the referent or semantic value of any such terms. It isn't that causation cannot play a role; it can narrow things down. But it needs supplementation.

So we need some way of narrowing down the range of candidate properties so that only some are eligible to be the contents of our judgments. David Lewis (1984) has suggested that we should employ the distinction between natural and unnatural properties, or better more and less natural properties.[17] In situations where our dispositions and practices under determine the referents for our terms, the more natural properties are eligible candidates for those referents whereas the less natural are not. Lewis's idea seemed to be that natural science would be the arbiter of naturalness, though I'm not entirely sure that is what he had in mind. In any case, I want to suggest a modification of this idea. Naturalness should be

[17] I thank Michael Smith for reminding me of the importance of Lewis's idea for the sort of semantics I am trying to work out here.

seen as discipline-relative. The kinds or properties which are more natural for the purposes of physics may not be the same as those which are more natural for purposes of biology. The more eligible semantic values for one's terms when engaged in the former may or may not be the same as the more eligible semantic values for one's terms when one is engaged in the latter. Probably the naturalness of a property or kind for a given discipline is relative to the subject matter and purposes of the discipline. I can't say exactly how this is supposed to work, but my sense that naturalness must be discipline-relative stems partly from thinking that when I'm being appropriately responsive to the questions posed within a discipline and the evidence we have for different hypotheses, certain methods of classification and of picking out properties relevant to the theory seem more natural to use than others.[18] And it seems to me that what seems natural to me in one sort of domain is not what seems natural to me in another. Tables seem perfectly natural for anthropological purposes, but not for the purposes of physics.

With this modification, I propose to follow Lewis in employing the naturalness of kinds to determine the eligibility of kinds for referents of terms. Even though the actual dispositions of speakers and thinkers are insufficient to rule out gruesome interpretations of their talk, the greater eligibility of the more natural properties to be the referents of terms can be used to single out the more natural property as the referent. Once we have such a notion to be employed this way, it can also be employed to overcome the related problem that people can be disposed to make mistakes. For such cases we can allow the greater eligibility of the more natural kinds to override the actual dispositions of speakers. Two speakers who are differently disposed to use some term over a range of cases can still be said to be using it with the same meaning because we expect them both to be referring to more or less natural properties, and thus we chalk up their divergence as due to error.[19]

[18] The idea that naturalness might be discipline-relative is not particularly original with me, but it does seem to deviate from Lewis's proposal. His discussion allows that different theories will employ different classificatory kinds, but has all of them ordered along one scale for naturalness. Thus, ordinary artifact terms such as 'table' select a less natural kind than electrons, though the fact that tables form a more natural kind than more disjunctive or gruesome kinds still figures in 'table' designating tables. My own intuitions are that tables are just as natural as electrons unless we happen to be asking the sorts of questions that physics aims to answer.

[19] In a nice paper, David Copp (2000) suggests that the referential intentions of speakers, plus their interests can be used to similar effect in fixing the referent of terms within a Putnam inspired naturalistic semantics for moral terms. And he uses this to suggest that terms on Earth and Twin Earth might mean the same thing. This idea may be very similar to the suggestion I am making here, though he thinks that there is an important contrast between Putnam and Boyd with respect to the applicability of this

It should be relatively easy to see how the idea extends to speakers causally regulated by different properties in their use of a term. For I take it that the idea of causal regulation involves judgments about which properties are in fact causing the speaker in question to use the word as they do. Normally different dispositions with respect to their use of the term will reflect difference in which properties are causally relevant to their uses of the term, even where those dispositions are not manifested. On the indicated approach to kind reference, differences with respect to those dispositions can sometimes be ignored in the interpretation of a speaker's utterances. And on the modification of the Boyd approach that I advocate here, we are allowed to make a similar move. What matters is not which kinds actually regulate a speaker's use of a term, but which kinds it makes sense to think that community members are talking about, given the overall uses to which they put the terms and the ways in which they make judgments about when the terms apply.

This is also the picture suggested by some of the well-known examples employed in the externalist literature about the meaning of natural kind terms. As Putnam (1973) has pointed out, we regard people's use of the term 'water' prior to 1700 as designating and meaning just what *we* designate and mean by that term. But our use of the term as we now use it depends on things that we have learned since 1700. It can't be just the fact that we do use this term in this way that makes it the case that those in the past use it with the same meaning as we do. It has to be because it is non-arbitrary that we so use it. We use 'water' to refer to H_2O and not liquids containing either H_2O or some nearly indistinguishable liquid of a different chemical makeup which they were not in a position to rule out as the referent of their terms. Our ruling that alternative out could only be a reason to interpret their meaning as our meaning if we think that our choice is governed by the evidence and rationally responsive to that evidence. If we retrospectively decide that we have made a mistake, either because we lack some evidence or because we used bad judgment, we should change our judgments about what meaning the term 'water' had in 1700. And we should change our view of what the meaning and referent was all along, not think that our discovering a mistake changed the meaning.

On this way of looking at things, once a certain basis for use of the term has been established, there are two further factors determining the referent. There are facts about the kinds to which the speakers have epistemic access and there are facts about which kinds are most natural. Neither by itself is

idea and he thus does not try to work it into a regulation-based story of the sort Boyd suggests. I'd like to downplay the role of speaker's referential intentions more than Copp would, partly because they can often be very vague and I would like a somewhat less individual conception of speaker's interests than he seems to employ.

sufficient to determine a referent, since speakers will have some access to gruesome kinds related to those we intuitively regard as the referents and since there may be some natural kinds to which they don't have sufficient epistemic access.

Once we have this much, we should allow that even interpretations that past users of the term have ruled out might in the end come to be seen as the correct interpretation of their language. This anyway is the upshot of another well-known example, Dalton's use of the term 'atom' (Burge, 1993). Dalton apparently defined atoms as the smallest particles into which matter could be divided. He suggested that all matter was made up of such atoms and also apparently believed that something like the periodic table captured differences in features of different kinds of atoms corresponding to certain kinds (Burge, 1993: 316). People are generally inclined to take Dalton to refer to atoms when he used the term 'atoms'. Yet his choice of definition seems to have explicitly ruled out any interpretation on which the referent was not the smallest indivisible particle of matter. But, if we regard him as offering a mistaken definition of his term, a term which has its reference determined in part by the true things he thought about atoms but which he did not take to be definitive, this need not be *our* verdict. It is fair to suppose that the experiments he did to determine that something like our periodic table correctly represented real features of what he called 'atoms' gave him knowledge of atoms. And these could form enough of a foundation to make it the case that his term and ours have the same meaning and the same referent. So to use a term correctly a speaker or community of speakers must have some knowledge about the referent, but that knowledge need not be what they themselves would offer as the defining features of the referent of the term.

Filling in the Details for Moral Twin Earth

Boyd's proposal as understood by H&T[20] uses the tendencies of the causal mechanisms over time to narrow down the range of candidates beyond what can be determined from the actual events that have occurred so far. And it is precisely this that leaves him vulnerable to the Moral Twin Earth objection. Given that the relevant mechanism is causal and hence only nomically necessary, we can coherently imagine alternative mechanisms which would focus the range on another candidate property. Thus it is fair for H&T to stipulate examples in which two populations have different dispositions

[20] Or at least one fair interpretation of that proposal, which Horgan and Timmons adopt in constructing their counter-argument.

to judge, so that the sets of properties instances of which regulate their judgments diverge or will diverge over time.

The modification I defend here uses two related ideas to avoid the problem. First, by emphasizing that the tracking criterion that makes our terms designate is epistemic and not always causal, we see how that idea can still yield common designees for our terms in the absence of the sorts of common causal mechanisms the original Boyd story required. Regulation of a community's use of a term by a property is only relevant to the designation of that property when that regulation yields knowledge of the property. And if some not-wholly-causal epistemic route generates the relevant sort of knowledge it can be relevant to determining the referent. This allows us to characterize the causal or dispositional divergences in Twin Earth cases as constituting errors or dispositions to err and discounting them in determining the referents of the relevant terms. Secondly, the modified approach suggests that the naturalness of a classification is epistemically relevant and that naturalness is discipline-relative. It is partly because of this that we can classify certain responses as mistakes. Sometimes a particular judgment will be best interpreted as involving mistakes about a natural kind rather than knowledge of an unnatural kind. By using discipline-relative facts about naturalness to narrow the range of candidates for the referents of moral terms we explain why relevant twinning scenarios cannot be constructed, since they require assigning an unnatural kind as the referent of one of the terms in question.

Relatedly, the proposal cites both necessary and contingent features of our world to determine the designee of the term 'right'. Since the relevant contingent features are stipulated to be the same across Earth and Twin Earth they can be used to generate the desired semantics. And insofar as the necessary features are necessary they will be available for that purpose on both planets. The contingent designation-determining facts are that (1) the term 'right' is used to reason about considerations that bear on well-being, that (2) people there are normally disposed to act in ways corresponding to what is 'right', that (3) these people take the rightness of options to be important, even over-ridingly so, in deciding what to choose, that (4) the term applies to actions, that (5) the rightness of an action is treated by competent speakers as being determined by features of the actions describable without using that moral term, and (6) 'right' is the most general term of moral appraisal which applies specifically to actions. The first four are the features of speaker's use of the term 'rightness' that Horgan and Timmons hold fixed between Earth and its moral twin, and the last two are features I claim need to be added for their argument to go through.

The relevant necessary designation-determining facts are that (1) some actions really do make more sense to go in for than others, (2) that if an action makes more sense than an alternative that fact is a reason to do it rather than the alternative, (3) that options make more or less sense because of the non-moral features that they have, and (4) that among the relevant features are things like how they effect the well-being and happiness of oneself, other people, and other creatures, and so on.

When a someone here on Earth tells me that they are doing an action because it is kind to so and so, and that its kindness makes it right, I take them to be expressing and acting on knowledge. For I take it that kindness does (often) make actions right, and that this is a reason to do actions of that general kind. The knowledge here expressed combines necessary and contingent facts and facts which are knowable only empirically with facts that seem to me to be discoverable through reflection and which hence are a priori. It is an empirical discovery (though often an obvious one) of contingent truths that certain ways of treating people make them happier and better off, and we need to know which kinds they are to treat people kindly. It is a necessary truth that making people happier and better off is something that makes sense to do and which we have reason to do (other things equal). This, I think, is knowable a priori though it may be that some empirical input is crucial in figuring it out at least for some people.

Does anything about the shift between Earth and Twin Earth shake our confidence that Twin Earth speakers have knowledge of these same facts? It is built into the example, partly in virtue of the internalist features of the Twin Earth scenario that there will be people like our speaker here on Earth, who take actions to be right because they are kind, and who act on that knowledge by doing what they take to be kind and right. The twinning operation did not in fact shift the major facts of human psychology and biology. Thus kindness on Twin Earth should be the same as it is here, and the same sorts of epistemic procedures should be relevant to finding out what it takes to be kind to another. In cases where Earthlings have knowledge about kind actions, their counterparts on Twin Earth will as well. What about the fact that in these circumstances kindness constitutes rightness? This, I think, is one of the a priori knowable and necessary truths. One does not know an a priori knowable truth just in virtue of believing it. One might believe it for the wrong reasons or might not have gone through the a priori reasoning, imagination, or reflection necessary for putting oneself in a position to know. But supposing that our Earthling friend has put herself in a position to know by reasoning appropriately, her twin will have done the same.

It is this basis which I think forms the foundation of knowledge sufficient to determine a common referent for the terms here and on Twin Earth. But it can do so only if we can view any divergence over the extension of the term (beyond the sort attributable to ordinary vagueness of terms) as involving a mistake by one or both parties.

Suppose Earthlings and their counterparts are disposed to diverge in their judgements about what is right and what is not over a significant range of cases. One way this could come about is if the empirical component of their epistemic processes diverges. They might come to disagree about which acts are kind due to disagreement about what sort of nervous system one needs to feel pain. Here we have no trouble ruling out one view or another as just a mistake, relevant neither to the referent of 'pain', nor to the actual extension or referent of 'kindness'. Suppose instead that the disagreement comes out in disagreements over which naturalistically described actions count as right such as when kindness can be too demanding. If this is not just a disagreement about a borderline case and realism about rightness is correct, there is an answer to this question. The answer here must be necessary, whatever it is.[21] The dispute here is over which way of conceptualizing rightness gives us a more natural moral kind, that is over which conception cuts moral reality at its joints. Except for borderline vagueness, at most one view will be correct, so at most one speaker will be reasoning correctly and getting things right. And since one can be justified a priori in accepting views about what makes an action right, if the person who gets it right has gone through the relevant epistemic reasoning, her beliefs may be justified and count as knowledge.

This helps explain the difference in our intuitions between the standard Twin Earth scenarios and the Moral Twin Earth scenario. Where the referent is a naturalistic kind, and it is a contingent matter whether the truth of our beliefs about a kind is accidental or not we can stipulate relevant twin scenarios. In the original Twin Earth story, much of what we believe about water is equally true of XYZ as it is about H_2O•. But only with H_2O is what we believe non-accidentally true. That's because most of what we believe about water is composed of contingent empirical claims. To say that they are contingent is to say that they could have been otherwise and hence we need information from the actual world to figure out whether they are true, and that is to say they are empirical. And to say that they are empirical is to say that these contingent facts must have a causal impact on our powers of observation. It is because H_2O has had that sort of impact on us that it is no accident that we believe what we do about water. • Q1

On the other hand, with the kind of moral case we are now envisioning, no such epistemic procedures are relevant. If we are in dispute about how much kindness can morally demand, and we have the issue clearly

[21] The correct way of conceiving of the supervenience of the moral on the non-moral will assure that much.

defined, no actual experimental evidence seems relevant to settling that issue. What could further empirical information tell us? The issue in dispute is hypothetical and does not turn on contingent and empirical matters. If kindness requires a certain sort of sacrifice is it still morally required? This does not mean that our beliefs about this could not be accidental and hence not knowledge. But it does mean that what it takes for them not to be so is not a matter of causal contact with the property in question. Rather we have to have thought about the issue in the right way so that our beliefs are correct because we have thought about the relevant issues correctly or nearly so. If I'm right this will require correct reasoning and vivid and sympathetic imagination. If after imaginatively and sympathetically thinking things through we arrive at the correct view about what to do, this will be no accident, and the epistemic constraint that I think forms the heart of the Boyd-type stories will be satisfied. Some of our beliefs about rightness will count as knowledge and thus our term 'rightness' will refer to rightness.[22] So will the analogous beliefs of those on Twin Earth.

It may be possible to stipulate a scenario as a limited test case for the modified theory. Imagine that the things we express using the terms 'right' and 'rightness' are not known because it is only an accident that people believe them even though they are true. It is hard to describe the sort of case required but it is worth trying. The scenario requires that our twins use their term 'right' in a supervenient way, not because they think it makes more sense to treat like cases alike in evaluating actions, but for some other reason. And similar things must hold for their treating rightness and well-being as linked and their thinking that the rightness of an action counts in favor of doing it. They must have reached these conclusions not by thinking imaginatively and reasonably about the alternatives but through some epistemically irrelevant process. Maybe throwing darts at a special dartboard would be an example of such a process. If that is the story I no longer think their word means what ours does. If so, this is some confirmation of the proposal.

To conclude this section I'll summarize. A suitably modified version of the sort of externalist theory that Boyd uses can explain how reference to properties can be secured for moral terms even while competent speakers

[22] Sayre-McCord (1997) has suggested that changing the Boyd account to specify that the property in question be a moral as opposed to natural kind would dissolve the troubles caused by H&T's scenario. I agree with using moral kinds as the relevant sort of natural kind for the domain. But by itself this move doesn't overcome the problem. The H&T example is a problem for the reference-determining mechanism proposed by Boyd and not for the particular referent proposed. So changing the referent without changing the mechanism will not by itself provide an answer to the objection, though providing an answer will involve changing the referent from what Boyd thinks it is.

can raise questions about their reference. And it can do this while remaining immune from Moral Twin Earth counter-examples, chiefly because it places a greater emphasis on the epistemically relevant features of the proposal, and because moral epistemology makes it harder to shift the necessary epistemic facts in such a way as to provide a counter-example.

Why Believe Any of This?

Is there any reason to accept the semantic story, apart from allowing us to avoid Horgan & Timmons's clever counter-example? Obviously I think so. Metaethical arguments are always a species of argument to the best explanation. A certain range of phenomena are taken to be data about moral practice and we construct theories to explain those data. And the proffered semantics for moral terms is part of a package which, as I've employed it, includes a commitment to a certain sort of rationalism. Whether we should accept this package depends on how it fares in explaining the various data about ethics that we think needs to be explained. In addition to being a clever counter-example, the Moral Twin Earth story highlights a number of the data that a good metaethical theory needs to explain. So a theory constructed to handle that example will, if all goes well, score rather well in the contest that arguments to the best explanation set up.

One datum is that competent speakers can doubt the identity of rightness and any property picked out in other terms. The theory I offer shares with Boyd's the ability to explain open questions about true identities wherever those identities are empirically established or only empirically establishable. But it surpasses Boyd's theory insofar as it explains how open questions are possible even when the reasons to accept a true identity statement are a priori. One thing the Twin Earth example does is to highlight the need for this. If competent thinkers can doubt the identity of rightness with any property picked out using some other form of words, either it must be possible to question even a priori establishable identities, or there must be no a priori arguments for thinking rightness identical with any property. Given the game that many of us are involved in—offering relatively a priori arguments for the identity of moral properties, we need a story to tell about how such doubts are possible. And the theory on offer suggests that we can find our explanation by noting that even a priori truths may be knowable only through a process of reasoning and that we can always sensibly wonder whether our reasoning about some matter has been correct.[23]

[23] The point is obvious but a surprising number of theories flout it. Take e.g. models of belief revision on which all of the a priori truths automatically get a credence of 1.

Another datum, put very blandly, is that moral judgments have some sort of tight connection with action. Yet that tight connection seems compatible with some people not being moved by moral judgments they accept. On the one hand we would not translate a term using our term 'right' if people were not normally inclined to do what they took the term to apply to. On the other hand it seems perfectly conceivable that some people not be so moved, and even that some people expressed doubts about the rationality or sensibleness of doing those things. Not only does it seem conceivable, but actual people seem sometimes to be like this. Insofar as Earth and the stipulated Moral Twin Earth share these features, the fact that we take them to be speaking univocally suggests that this sort of tight connection may be grounded in just these similarities.

Our package endorses this hypothesis and explains why it might be correct. If enough people in a community get enough things right about a property, so that in principle there is available to all speakers in a community an epistemic pathway to finding out more about that property, then all members of that community can be credited with thoughts and talk about that property, even if some of what they say is false and even necessarily false. One way (among others) that the belief that something makes sense to do can manifest itself is by doing it when one is in a position to do so (van Roojen, 2002). My suggestion is that those who do mostly do what they think right are acting on a belief whose content is that these things make sense to do. And, I suggest, this belief counts as knowledge if it stems from noticing that some features of actions make them make more or less sense to do, when it is the features of the action that in fact make it make sense that grounds their judgment about the case, and when they reliably (though not infallibly) discriminate the sensible from the inadvisable. Thus most of the populations of both Earth and its moral twin meet the conditions for getting things right much of the time. Those who are not motivated to do what they regard as right are making a mistake, perhaps culpable perhaps not. But given that their fellow inhabitants get things right and that their use of the term 'right' depends on the practices of their fellows, they too use the term with the same meaning and it refers to the same property in their mouths as it does in the mouths of their friends.[24]

The theory then supports the sort of internalism suggested by the Twin Earth argument. When and only when most people in a community are

[24] Conceptual role semantics of the sort advocated by Ralph Wedgwood in the conference draft of his paper for this volume may also be able to explain this form of internalism requirement allowing some people not to be appropriately motivated, but only if it takes the form of broad rather than narrow conceptual role semantics. Wedgwood prefers the narrow version.

guided by the application of a certain term in choosing what to do, and when certain other requirements are met we will think of it as an evaluative term. But this does not rule out the sort of amoralist often invoked by externalists to refute internalism. We can think of these amoralists as something like the patient who thinks that he has arthritis in his thigh—someone whose membership in a community gives him competence enough to make judgments using a concept, even when the most competent members of that community would regard some of what he believes as incoherent.[25]

There is one further desideratum that many metaethicists find important. This is that the theory fit into a broadly naturalistic world view, whatever that comes to.[26] So it might be appropriate to say something about whether the resulting theory counts as naturalistic. My answer is that it depends, and that a lot of what it depends on is not special to ethics.

Naturalism is usually stated as a contingent thesis. The world might have been such as to contain non-natural things but it does not. And while this idea seems coherent enough when applied to concrete entities, how to apply it to properties is not all that straightforward. It seems like there could be three sorts of properties—those that could be had only by natural objects, those that could be had only by non-natural objects, and those which might be had by either. Leaving aside the difficult question of what it takes for an object to be natural, only the third sort of property cannot be instantiated in a world with only natural objects. So probably naturalism should be conceived in a way consistent with the existence and even the instantiation of properties which could be had by non-natural entities.

Now it looks like rightness and goodness ought to be thought of as this sort of property. On any favored partition of entities into natural and non-natural, it seems like the non-natural entities would—if there were any—be morally evaluable, and actions with effects on them might vary in rightness in virtue of those effects. If we think naturalism rules out gods and ghosts, it does not seem to follow that killing a god or a ghost would not be wrong, if only we could do it. Nor does it follow that some gods or ghosts could act wrongly or be morally evil.[27] These examples invoke only one way of partitioning the natural from the non-natural, but it seems like

[25] See Burge (1979). I discuss this analogy in more detail in my manuscript, 'Moral Rationalism and Rational Amoralism'.

[26] One might be forgiven for suspecting that a non-cognitivist analysis of 'natural' and 'naturalism' has something going for it.

[27] One of the best reasons to doubt the existence of the sort of god postulated by most Americans is precisely that no such god could have all of the other features commonly supposed and still have acted rightly. Such a god would be very bad (contrary to what most believe), but the truth of that claim relies on their being properties that could be had by nonexistent non-natural entities which would not be natural if there were any.

the point is pretty general. For any other coherent sort of thing that you might think there could have been but as a matter of fact is not, it seems like features of these things might be morally relevant.

The point here is really not specific to evaluative properties, since I think there are other properties that could be instantiated in a world of only natural items, but might also be instantiable in non-natural worlds. Existing in close proximity to five other distinct things seems like a candidate. So if you think that naturalism is incompatible with properties of this sort I think you should conclude both that my story is not compatible with naturalism, but also that naturalism is false. That option creates no problem for my account.

Sometimes in metaethics when people suggest that they are naturalists they don't seem to mean to rule out very much. They admit that the possible extension of moral properties would go beyond any very substantive delineation of the natural, admitting ghosts, forces, or entities not substantiated by natural science, or anything else we can come up with. What these folks seem to care about is that moral properties not be distinct from their supervenience bases. Given the argument above, these supervenience bases will include an awful lot of pretty weird stuff—stuff that seems pretty unnatural to me. So the concern here is different from the concern that motivates those who emphasize that naturalism is a contingent hypothesis.

My response to this specification of naturalism is that my favored theory is neutral on it, but that the issue turns on nothing all that special to ethics and much more on general metaphysical principles. If there is reason to identify properties with the same extension as infinitely disjunctive properties, that reason would apply here. Whether that identification is best thought of as a reduction of the higher level property to the lower is also an interesting question.[28] But whatever the answer to these questions, the sort of view I suggest here seems ready to accommodate the verdict. Moral terms refer to properties about which some of us have knowledge which we express using those terms. If those properties are identical to properties picked out with infinite disjunctions of non-moral terms then our moral terms refer to those. If they are not identical they do not. There is no problem for the account if this sort of naturalism turns out to be true. And in fact insofar as the theory remains able to explain the possibility of open questions it should be welcomed by naturalists who identify moral properties with such constructions from natural properties.

In summary then, the proposal allows us to use a modified version of Boyd's original idea to handle the Moral Twin Earth example. And this

[28] See John Gibbons's as yet unpublished paper, 'Identity without Reduction', for an interesting discussion of the issue.

version also has the virtue of explaining two features of the example that we should wish to explain anyway. First it explains how any identity claims about moral properties can be open for competent speakers. And second, it explains how a very weak sort of internalism—one which postulates a necessary connection between moral judgments and action-guidingness for many members of a community—could be required. And it does both of these things without violating any naturalist strictures that we should be inclined to accept.

REFERENCES

Boyd, Richard, 'How to be a Moral Realist', in G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 181–228.

Brink, David, 'Externalist Moral Realism', *Southern Journal of Philosophy*, supplement (1986), 23–40.

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

Burge, Tyler, 'Individualism and the Mental', *Midwest Studies in Philosophy*, iv, ed. P. French (Minneapolis: University of Minnesota Press, 1979), 73–121.

—— 'Concepts, Definitions, and Meaning', *Metaphilosophy*, 24/4 (1993), 309–25.

Copp, David, 'Milk, Honey and the Good Life on Moral Twin Earth', *Synthese*, 124 (2000), 113–37.

Dreier, James, 'Internalism and Speaker Relativism', *Ethics*, 101/1 (1991), 1–26.

Geirsson, Heimir, 'Moral Twin-Earth: The Intuitive Argument', *Southwest Philosophy Review*, 19 (2003), 115–24.

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

Hare, R. M., *The Language of Morals* (Oxford: Oxford University Press, 1952).

Horgan, T., and Timmons, M., 'New Wave Moral Realism Meets Moral Twin Earth', *Journal of Philosophical Research*, 16 (1991), 447–65.

—— and Timmons, Mark, 'Troubles on Moral Twin-Earth: Moral Queerness Revived', *Synthese*, 92 (1992*c*), 212–60.

—— and —— 'Troubles for New Wave Moral Semantics: The "Open Question Argument" Revived', *Philosophical Papers*, 21 (1992*b*), 153–75.

—— and —— 'Nondescriptivist Cognitivism: Framework for a New Metaethic', *Philosophical Papers*, 29 (1992*a*), 121–53.

Kalderon, Mark, 'Open Questions and the Manifest Image', *Philosophy and Phenomenological Research*, 68 (2004), 251–89.

Kripke, Saul, *Naming and Necessity* (Cambridge, Mass.: Harvard University Press, 1972).

—— *Wittgenstein on Rules and Private Language* (Cambridge, Mass.: Harvard University Press, 1982).

Lewis, David. K., 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), 343–77.

Moore, G. E., *Principia Ethica* (Cambridge: Cambridge University Press, 1903).

Nozick, Robert, *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981).

Ogden, C. K., and Richards, I. A., *The Meaning of Meaning* (New York: Harcourt Brace & Jovanovich, 1923).

Putnam, Hilary, 'Explanation and Reference', in G. Pearce and P. Maynard (eds.), *Conceptual Change* (Dordrecht: Reidel, 1973), 199–221.

____ 'The Meaning of Meaning', in K. Gunderson (ed.), *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science*, vii (Minneapolis; University of Minnesota Press, 1975).

Salmon, Nathan, *Frege's Puzzle* (Atascadero, Calif.: Ridgeview, 1991).

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

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

____ 'Internal Reason', *Philosophy and Phenomenological Research*, 55 (1995), 109–31.

Soames, Scott, *Beyond Rigidity* (Oxford; Oxford University Press, 2002).

Sosa, Ernest, 'Water, Drink, and "Moral Kinds"', *Philosophical Issues*, 8 (1997), 304–12.

Stocker, Michael, 'Desiring the Bad', *Journal of Philosophy* (1979), 738–53.

Timmons, Mark, *Morality without Foundations* (Oxford; Oxford University Press, 1999).

van Roojen, Mark, 'Humean and Anti-Humean Internalism about Moral Judgements', *Philosophy and Phenomenological Research*, 65/1 (2002), 26–49.

____ 'Cognitivism vs. Non-Cognitivism', *Stanford Encyclopedia of Philosophy* (Spring 2004 edn.), ed. Edward N. Zalta: <http://plato.stanford.edu/archives/spr2004/entries/moral-cognitivism/>.

Queries in Chapter 6

Q1. We have changed ‘0(zero)’ to alphabet ‘O’ for the compound ‘H_2O’ in this paragraph. Please check.