Moral Language and Moral Discovery:

Making Do Without Twin Earth

Daniel Muñoz

I. Planning for the Twin Apocalypse

For a while, it looked like Moral Twin Earth had done away with *Causal Semantic Naturalism* (CSN)—the view that moral terms like 'good' pick out their causal regulators, just as 'water' picks out H_2O (Horgan and Timmons 1992; Boyd 1988; Brink 1989). For suppose that pleasure regulates our word 'good' and beauty regulates its Twin English homophone, and imagine that Twin Daniel tries to deny my claim that only pleasure is good. If CSN is right, it seems to follow that my twin and I aren't having a substantive disagreement: we're just using different words. But intuitively we *are* disagreeing, so CSN must be false.

Janice Dowell (forthcoming) has shown that things aren't so simple. She claims the argument from Moral Twin Earth rests on an intuition with "no probative value": the intuition that the moral vocabulary of Twin English is semantically near enough to ours to allow for substantive disagreement across languages—that our word 'good' means roughly the same as its Twin English homophone (3). But why trust *our* intuitions about *Twin English*, a made-up language we don't competently speak? As Dowell notes, our intuitions about meaning are only credible when they stay within the bounds of our linguistic competence, and our competence is with actual languages like English, not their imagined rivals. So, CSN isn't refuted by the familiar arguments from imaginary

disagreements—disagreements involving Moral Twin Earth, Smith's "rather different community" (1994: 32-3), and Hare's missionary (1952).

Even if the familiar arguments survive Dowell's objection, they probably won't ever seem as decisive as they did before. That's why I think we opponents of CSN should plan for the Twin Apocalypse: we need an alternative to Moral Twin Earth that does the same work without the same dubious intuitions.

I think we can find a suitable replacement in cases of *moral discovery*. While disagreement cases turn on an implausible difference in meaning within two rival communities, discovery cases involve a difference across time within just *one* speech community—our own, pre- and post- discovery. Moral discoveries may cause a radical change in what regulates our term 'good', but they shouldn't cause a change in *meaning*. Even if I discover that equality is a distinct kind of good, and equality becomes a regulator of my use of 'good', I haven't *expanded* the term's reference. On the contrary, what makes my finding a discovery rather than a whimsical stipulation is that equality has been a referent of 'good' *all along*, in the same sense of 'good' that I've always used.

The plan is to argue that CSN can't be squared with these sorts of plausible beliefs about moral discovery, and that such beliefs don't rest on the intuitions that Dowell tries to debunk. In doing so, I hope also to establish a general point about Twin Earth-style arguments. Arguments that rest on intuitions about a hypothetical *rival* language can sometimes be recast, so that it's about a hypothetical *revision* or *extension* of our own language. And in some of these cases, the recast version will rely only on intuitions within the bounds of our competence. If I noticed that English changed slightly, as it did when NASA stipulated a new definition for "planet," and when "salad" started applying to dishes made from fruit, I would remain a fully competent user of the changed terms. And

so long as the change isn't too dramatic, I should also be able to *imagine* what English would be like if it were to undergo such changes—I just pretend that NASA made some new press release, or that the Society of Salad Lovers managed to persuade everyone that salads aren't fruity, and I note how my linguistic intuitions would change as a result. This seems to me like a fine method for turning a bad argument into a good one, and I think it would work for plenty of other argument that Dowell's objection seems to sink (e.g., Kripke 1977). When this method is applied to disagreement arguments like Moral Twin Earth, the result is the argument from moral discovery.

After laying out the argument, I'll consider some objections that appeal to naturalness and dispositional versions of CSN. I argue that neither of these work: no matter how much we soup it up, CSN will rule out a kind of moral discovery that, plausibly, could happen at any time.

II. Moore, CSN, and Moral Twin Earth

This section gives a brief recap of the dialectic up to now—CSN, Moral Twin Earth, and Dowell's objection.

First up is Moore's Open Question Argument, which tries to show that that 'good' can't be analyzed (1903). The argument comes in two versions: positive and negative. The positive one goes like this.[1] Consider some arbitrary analysis of "x is good" into "x is F," where F is a natural property. If we come to think that the analysis is correct, we'll think that goodness just is F-ness, and the sentence "The good is F" ought to have the same cognitive significance for us as "What's good is good," and "What's F is F." But

[1] The negative form says that "the good is F" can never be a contradiction in terms, no matter what natural property we plug in for F.
[2] Following Horgan and Timmons (1992: 161), let's say that the question of whether Fs

these last two are trivial, while the first strikes us as significant. Of course, it might turn out that goodness and F-ness go together in all cases. But even if this is so, establishing that something is F doesn't *thereby* close the question of whether it's good.[2] Until one does some further distinctly moral thinking, the question stays open, even for competent users of 'good' like us.

The burden is now on the naturalist: if goodness really just amounts to F-ness, whence the cognitive gap for competent speakers? Enter CSN, which holds that 'good' picks out a natural property, though we don't have to know which one in order to competently talk about goodness. The idea inspiring CSN is that there are two ways for a description to be relevant to reference. First is for the description to *secure* reference, as definite descriptions do. 'The nearest star' picks out the sun in virtue of the fact that the sun satisfies the description; the description is therefore part of the meaning of the term. So, if someone knows that 'the nearest star' picks out the sun, but doesn't realize that this is because the sun is closer to us than other stars, she doesn't fully understand the meaning of 'the nearest star', and she's not fully competent with it.

But there's another way for descriptions to get into the picture: instead of playing a semantic role by securing reference, they might play a *metasemantic* role by *fixing* reference (Putnam 1975; Kripke 1980). A reference-fixing description is one that helps a term latch onto something, but then evaporates rather than sticking around as part of the meaning. We might fix the reference of a name like 'Dartmouth', for example, with a description like 'the institution founded at the mouth of the Dart River'. The fact that

[2] Following Horgan and Timmons (1992: 161), let's say that the question of whether Fs are good is *closed* just in case the answer is obvious to someone who (i) is competent with "F" and "good," and (ii) "brings her competence to bear" on the question of whether Fs are good. A question is *open* just in case it's *not* obvious to such a speaker.

Dartmouth is at the mouth of the Dart explains how its name *came to refer* to it rather than Yale or MIT. But once reference is fixed, the description drops out of the meaning of the name. The name then functions as a rigid designator—in every possible world where Dartmouth College exists, 'Dartmouth' picks it out. Moreover, it does so directly, not by way of any description.

If a term picks out its referent directly, it doesn't take much to be competent with it. In particular, competent users don't have to know how the term's reference was fixed. Plenty of Anglophones have no idea where Dartmouth was founded, even though they know perfectly well how to use its name. So even though these speakers are competent with 'Dartmouth', for them it's an open question whether Dartmouth College fits its reference-fixing description.

The basic idea behind CSN is that the same is true of 'good': its reference is *fixed* by a naturalistic description (something like 'the causal regulators'), but not *secured* by it. So: (i) 'good' refers to a natural property F, even though (ii) it might be an open question for competent speakers whether the Fs are good.

Here's how Horgan and Timmons (1992: 455) define CSN's thesis: "Each moral term t rigidly designates the natural property N that uniquely causally regulates the use of t by humans."

The meaning of 'good' doesn't consist in a reference-securing description—'the natural property N…'—rather, it consists in the natural property that satisfies the description. To say what this property is would be to give a synthetic "natural definition," in Boyd's terms; a verbal definition, by contrast, would typically involve some descriptive content. A competent user has to know a term's verbal definition, but she needn't know the first thing about its natural definition. That's how CSN defends against Moore's

argument: it's no wonder why competent users of 'good' feel that it's an open question whether F is good—it feels open because the definition of 'good' is natural, not verbal, so speakers don't need to know about it.

But what about the speakers who *do* know the natural definition of 'good' because they know how its reference was fixed? If CSN were true, then knowledge of how 'good' has been causally regulated ought to close the question of whether F is good. But this question seems just as open as the one Moore started with.

We can grant CSN for the moment and do a bit of reasoning:

1. Fs are good if F causally regulates 'good'.
2. Fs causally regulate 'good'.
3. So, Fs are good.[3]

Certainly the conclusion follows from the premises, but according to Horgan and Timmons, it ought to still be of cognitive significance. They argue that the following two questions are open:

- Given that the use of 'good' by humans is causally regulated by natural property N, is entity e, which has N, good?
- Given that the use of 'good' by humans is causally regulated by natural property N, does entity e, which is good, have N?

If Horgan and Timmons are right, then CSN falls prey to a new version of the Open Question Argument. But it obvious that these two questions are open?

Enter Moral Twin Earth, whose purpose is to get you to answer 'yes'. Suppose you know that our word 'good' is regulated by natural property N_1. Now imagine that

[3] An analogous argument holds for "Good things are F."

you take a trip to a world that's very much like ours, except that on this world our doppelgängers use a homophone of 'good' ('good*') that's regulated not by N_1 but N_2. You and your Twin both know about the history of how your words were regulated.

One day, you two get into an argument over the evaluative status of sadistic pleasure; you say it's good, Twin You says it's not good*. You know that 'good' is regulated by sadistic pleasure, but 'good*' isn't. If CSN is right, you and your Twin are having a verbal disagreement. Since 'good' and 'good*' are regulated differently, you two are simply using different words, so you're talking past each other. You might resolve your disagreement by accepting that sadistic pleasure is good but not good*, without giving up any substantive beliefs.

But this isn't the intuitive take on the situation. Intuitively, you and your Twin *are* disagreeing about something substantive: whether sadistic pleasure is good, of value, a fitting object of pro-attitudes, etc. And if your debate is substantive despite your knowledge of causal history and competence with your respective terms, it would seem that the above questions really are open, and CSN falls to Horgan and Timmons just as old-school naturalists fell to Moore.

Dowell's objection to all this, as stated above, is that it relies on our linguistic intuitions about Twin English, which are defeated by the fact that this isn't a language we competently speak. But we don't need these intuitions to revive the OQA, if we can run an argument from discovery instead of disagreement.

III. Moral Discovery

Here's the argument from discovery:

1. If CSN is true, we can't make a certain kind of moral discovery.

2. We might make such a moral discovery.

3. So, CSN isn't true.

The next section will defend these premises. But first, I ought to say what I have in mind when I say "moral discovery," since it's the argument's key term. And rather than give elaborate arguments for the possibility of moral discoveries, I'll just try to home in on the sort of discovery that goes on when we discover a new value.[4]

Crucially, I *don't* mean to talk about the discovery of morally relevant natural facts—for example, the discovery that solitary confinement causes suffering. This sort of discovery helps us flesh out the situation we're considering. The role of moral discovery, however, is to tell us the moral status of a situation that's already been fleshed out. Given that solitary confinement causes suffering, is it permissible to impose it on prisoners (in such-and-such conditions)? When you come to know the right answer to this sort of question, it's because you've made a distinctly moral discovery.

We should also distinguish three further kinds of discovery. The first is *nature discovery*, the kind that happened when we discovered that water is H_2O. When we discover something's nature, we learn some propositions that tell us about what it is to be that thing, as well as what the thing is like. To make this notion of discovery more precise, we might add that the propositions we learn are about the thing's grounds, intrinsic properties, and essence. But the details don't matter; whatever you think is going on in the cases of water and H_2O, molecules and atoms, heat and mean kinetic energy—that's

[4] There's at least this argument: (1) What I mean by 'good' wouldn't change if I changed my mind about whether beauty is good; (2) beauty is either good or it's not; (3) so, it's either true that I could make a moral discovery by converting to nihilism, or true that I could make one by giving up nihilism and going back to thinking that beauty is good; (4) either way, moral discovery is possible.

nature discovery.

Second is *new kind discovery*. Roughly (and for nondegenerate cases!), we discover a new kind of thing when our term for it starts being regulated by new properties. Consider your own discovery (if you've had one) that bitter foods and drinks can be tasty. Pre-discovery, you were disposed to use 'tasty' to describe only things that were sweet, salty, sour, or savory in certain proportions. But post-discovery, your disposition became more liberal. You learned that bitterness had been a unique way of being tasty all along, a new kind of tastiness, and you adjusted your use of 'tasty' accordingly.

The difference here is simple. Nature discovery is learning more about the things you already knew about. New kind discovery is learning that a thing you know about comes in a kind that you didn't know about—this is the sort of discovery that's relevant to my argument. I'm arguing that CSN closes the question of whether we'll discover new kinds of value, when intuitively the question is wide open.

How new are the new kinds? In some cases, not very. A new kind of metal or wood will have a somewhat different nature from the familiar kinds, but it will share a cluster of core properties. Metals are generally hard, shiny, opaque, and good for conducting electricity. Woods are hard and organic, and they come from trees. New kinds of metals and woods are no exception; they'll have these core properties to a significant degree.

But not all kinds of a thing will share core properties. Bitterness and sweetness are quite different on the level of flavor, though they're both kinds of tastiness. Equality, virtue, and pleasure may not have much in common naturalistically, but (I'll assume) each

is a kind of goodness.[5] I'll return to this point later—that as a matter of first-order normative fact, 'good' may be more like 'tasty' than 'metal', in that the good things may not share a common cluster of naturalistic properties.

Finally, there's *non-canonical discovery*.[6] My favorite examples of this are foods and music genres. Did anyone discover that pies could be made with cheese and pepperoni? That rock 'n' roll could be made with mostly electronic instruments? That salad could be made from fruit? (For that matter, did anyone discover that science isn't a part of philosophy, as natural philosophy seemed to be?) There's a lot to be said about these cases. As foods and genres evolve, our terms for them stretch to cover new uses while also trying to keep their shape in order to preserve the truth of uses in the past. Past uses are subtly reinterpreted; marginal cases are hotly debated.

The gist of what's going on here is that we come to understand something better, and that new understanding gets absorbed by our language, changing the meanings of our moral terms. We don't *learn a proposition* of the form "Salads can be made from fruit"; rather, we come to *assent to different sentences* because we've learned something that made a more permissive view of salads seem reasonable. (Maybe we realize that what matters to us about salad is that it doesn't overpower an entrée, or something like that.)[7]

I think it would be fascinating if some moral discoveries turned out to be non-canonical. Perhaps this is what's happening when we give up racial slurs or other problematic thick terms (e.g., 'chaste', 'masculine') in favor of more enlightened verbiage. Still, I don't think this story is plausible for thin terms like 'good' and 'right'. If it were,

---

[5] I'm happy to talk in terms of different *kinds* of value, or different values, or different ways of being good. This shouldn't affect my argument.
[6] Cf. Plunkett & Sundell (2013) on non-canonical disagreements.
[7] Some non-canonical discoveries are more serious than foods and genres; one example might be the discovery that marriage can be between two people of the same sex.

that would be a serious objection to my view, since CSN allows for non-canonical discovery; it's only new kind discovery that it has problems with.

Fortunately, I've got three reasons for thinking that moral discovery is canonical. First is that there doesn't seem to be much room for semantic evolution; the terms are just too thin. A less glib reason is that moral discoveries *feel* canonical—they feel like new kind discoveries. This is a point how discoveries strike us, rather than a point about how the terms themselves do.

When a person discovers a new value (for herself, if not for her whole speech community), she gets the sense that *this has been good all along*, in the exact sense of 'good' that she's been using all along. She seems to learn the truth of a moral proposition, not an ancillary lesson to be absorbed into her moral vocabulary. Things are a bit different when one discovers that salads can be made of fruit, or that marriages can take place between gay couples. One gets the sense that one is finally seeing what salads should be all about, or what marriage should be getting at, but unlike with moral discoveries, one doesn't get the sense that fruits have been salad-eligible all along.

The most important reason to doubt that moral discovery is non-canonical, however, is that some moral discoveries involving thin terms can only be captured in those terms. When I learn that beauty is good, the proposition I learn is *that beauty is good*; the only other way to capture this would be to give some analyzed version. But for a non-canonical discovery that beauty is good, what's learned can be expressed in terms that don't include beauty, goodness, or even any analyses thereof.

Consider some other examples. The non-canonical discovery that gay couples can get married might consist in one's learning more about what gay relationships are like, or learning about the importance of companionate love. The non-canonical discovery about

salad might consist in the realization that one point of the meal (right now, for us) is to eat something healthy for once, which makes fruit a good candidate ingredient. These two discoveries lead me to assent to different sentences about marriage and salads, but what I learn can be expressed in terms that don't involve either. This just doesn't seem to be so for moral discoveries.

Finally, the moral discoveries featured in my argument are discoveries about *foreign* values—ones that we've never come into causal contact with. These are the ones that cause trouble for CSN, which entails that the question of whether such values exist should be closed for any competent user of 'good', though intuitively the question should be open.

IV. The Argument from Discovery

Now that we know what kind of discovery we're dealing with, we can put the argument more precisely:

1. If CSN is true, the question of whether we might discover a foreign kind of value is closed.
2. The question of whether we might discover a foreign kind of value is not closed.
3. So, CSN isn't true.

Why should CSN get in the way of such moral discoveries? Because CSN tells us that 'good' designates its causal regulators, and the only things that can be regulators are things that we've encountered. And of course, a foreign value is precisely one we haven't encountered. Nothing surprising here: CSN says that the reference of 'good' is fixed by causal contact with something; so, no contact, no reference, and this is all scrutable a

priori. The same goes for other terms that get the causal semantic treatment, like 'water' and 'tiger'.

The second premise is also pretty intuitive; it's just a new spin on the familiar open question intuition. Except this time, we don't even need to consider whether some particular natural facts close the question; if CSN is right, the question is closed *automatically*, since we can know a priori that we've had causal contact with all referents of our moral terms. At the start, it certainly seems like an open question whether we've encountered all the values, and the claim that we have strikes us as substantive. Moreover, if it were true, it wouldn't be trivial, but marvelous! What a lucky speech community we'd belong to!

So, since the two premises seem straightforwardly true and the argument is valid, we should accept the conclusion that that CSN is false.

V. Objections

Now that the argument's been laid out, I'll consider some objections. Both suggest revisions of CSN that try to make it so that our moral terms refer to foreign values, as well as the ones we know and love, so they both target premise 2 of the argument.

a. Dispositions

What about a version of CSN on which our moral terms refers to the natural property that we're *disposed* to have our usage regulated by? Such a view wouldn't rule out the possibility of foreign values, since we equally disposed to be regulated by these as we are by the values we happen to have encountered.

I only need to mention one problem with this objection: it vitiates the motivation for being a causal semantic naturalist, since it undermines the original Twin Earth case. Oscar from the 17th century doesn't know that his term 'water' is regulated by H_2O rather than XYZ. So, he's disposed to be regulated by XYZ just as much as H_2O.[8] But of course that shouldn't establish that 'water' refers to XYZ; that goes against the whole point of Twin Earth.

To defend against this, the friend of CSN might suggest an *ideal disposition* account: moral terms refer to what would regulate them *if* we were to discover lots about the nature of their actual regulators. But this won't do anything to stop the argument from discovery, since nothing in the argument turned on *regulator discovery*; it works no matter how much or how little we know about what our terms' regulators are like. The ideal disposition version of CSN will still make it a closed question whether there are foreign values, so it won't save CSN.

b. Naturalness

The objection from naturalness grants that we may not have encountered all of the values, but insists that if we've encountered any, that's enough to secure reference to the whole class. This objections that among the closed questions is whether goodness is natural; the answer is yes! On a simple version of the view this suggests, 'good' doesn't just refer to any old property that regulates it: it refers to the most *natural* property that does so, in the Lewisian sense of 'natural' (1983). This property gets to be the privileged regulator, we might say.

[8] If you worry that this isn't true because Oscar isn't near enough to any XYZ, just suppose that there's an XYZ aquifer beneath his house that's poised to burst next week.

The problem with this view is that there might be several equally natural properties vying to be the privileged regulator, or there might be a unique winner that doesn't refer to the whole class of values. Here's how that might happen: suppose pleasure is the only good we've encountered, though there's a further foreign value. Now, pleasure is extremely natural—perhaps even perfectly natural. So why should we expect goodness to be more natural? If anything, CSN suggests that goodness is extremely *unnatural*; this would certainly be the case if it were a "homeostatic property cluster," of the sort described by Boyd (1988).

Furthermore, we can turn this into a pincer objection, since it might be that the most natural property doing the regulating is much more *general* than goodness, rather than being too specific. (The property of being naturalistic might count.)

There's a more basic objection, however, which is that it's an open question whether the good things share any (Lewisianly) natural non-moral properties. Certainly some distinguished philosophers think so (Thomson 1997). And philosophers in the intuitionist tradition think the same about the deontic side of ethics (Ross 1930; McNaughton 1996; Prichard 2002; Dancy 2004).

But if the naturalness of 'good' is an open question, then we can't appeal to this naturalness to close the question of whether there are further, foreign values.

VI. Conclusion

The argument from discovery tries to do what Moral Twin Earth did, but without the cross-linguistic intuitions to which Dowell has objected. I hope that in formulating this argument I've given a demonstration of a general style of argument-CPR. If an argument

seems vulnerable to Dowell's objection, we can try to change its appeal to a hypothetical rival language into an appeal to a hypothetical revision of a familiar language. In doing so, we may sometimes find that our basic ideas didn't rest on anything having to do with distant planets and foreign languages. We may find that we wanted to say something quite simple, and if we do, we should try to express our thoughts with arguments that are rich enough to be compelling, but restrained enough that they don't outstrip our intuitions.

Works Cited

Boyd, Richard (1988). "How to be a Moral Realist", *Moral Realism*, ed. G. SayreMcCord. Ithaca: Cornell University Press.

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

Dowell, J. L. (forthcoming). "The Metaethical Insignificance of Moral Twin Earth." In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*. Oxford.

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

Horgan, Terence & Timmons, Mark (1992). "Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived." *Philosophical Papers* 21 (3):153-175.

Kripke, Saul A. (1980). *Naming and Necessity*. Harvard University Press.

Lewis, David (1983). "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61: 343–377.

McNaughton, David (1996). "An Unconnected Heap of Duties?" *Philosophical Quarterly* 46 (185):433-447.

Plunkett, David & Sundell, Timothy (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13 (23).

Prichard, H. A. (2002). *Moral Writings*. Ed. MacAdam, Jim. Oxford University Press.

Putnam, Hilary (1975). "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7:131-193.

Ross, W.D., (1930). *The Right and the Good*, Oxford: Oxford University Press.

Thomson, J.J., (1997). "The Right and the Good", *Journal of Philosophy*, 94: 273-298.

24.502 Topics in Metaphysics and Ethics

Fall 2014