Reference Magnetism as a Solution to the Moral Twin Earth Problem*

Billy Dunaway  Tristram McPherson
University of Oxford  Virginia Tech

28 February 2014

Abstract

‘Moral Twin Earth’ thought experiments constitute a central semantic challenge to naturalistic normative realism. This paper argues that naturalistic realists can meet this challenge by appealing to the idea of reference magnetism. According to this influential idea, some properties are reference magnets, which (roughly) means that they are easy to refer to in virtue of carving nature at its joints. We argue that reference magnetism can smoothly vindicate plausible judgments about Moral Twin Earth cases. Further, it is a unified account of reference determination that can nonetheless explain the striking asymmetry between the apparent semantic behavior of normative terms and that of standard natural kind terms like ‘water’. In the second half of this paper, we sketch two contrasting ways of implementing our response, which are, respectively, conservative and revisionary of our intuitive semantic judgments. We conclude that both implementations of the idea are plausible fully general responses to the Moral Twin Earth argument. The availability of the two implementations of our response makes it an attractive resource for a wide range of naturalistic normative realists.

One might think that, whatever other challenges they face, naturalistic normative realists should be on strong ground with respect to semantics. After all, they take our normative terms and concepts to refer to the same sorts of properties as our mundane naturalistic terms and concepts, so they might hope to simply adopt the most promising general naturalistic theory of reference. However, in a series of papers, Terence Horgan and Mark Timmons have shown this optimism to be hasty (1991; 1992a; 1992b; 1996; 2000; 2009). In these papers, Horgan and Timmons marshal a series of ‘Moral Twin Earth’ thought experiments to argue that leading naturalistic theories of reference determination will fail to account for the range of

*We are indebted to Dave Chalmers, Anthony Eagle, Matti Eklund, Allan Gibbard, John Hawthorne, David Manley, David Plunkett, Peter Railton, Geoff Sayre-McCord, Laura and Francois Schroeter, Brett Sherman, Mark Schroeder, Laura Schroeter, Michael Smith, Pekka Väyrynen, and Sean Walsh, audiences at Bowling Green, the University of Leeds, Oxford University, the Monash Mind and Morality Workshop, SPAWN, and the Australian National University, and to students in a Spring 2012 Virginia Tech graduate seminar Semantics of Normative Realism for helpful discussion of drafts of this paper.
circumstances in which speakers can substantively disagree with each other about moral matters. Horgan and Timmons claim further that these failures generalize: any naturalistic theory of reference determination for moral terms which succeeds in securing a satisfactory degree of referential determinacy will fail to adequately account for the full range of semantic disagreement (2000, §1).

In this paper, we grant the force of Horgan and Timmons’s arguments against the particular theories of reference that they have addressed. We will show that their attempt to generalize their argument fails, in light of making overly restrictive assumptions about the possible structure of plausible naturalistic theories of reference-determination. We build on this diagnosis by introducing an independently motivated theory of reference-determination, and showing how it allows the normative realist to meet the Moral Twin Earth challenge (§2). The theory we defend holds that certain properties are *reference magnets*: they are easy to refer to in virtue of carving nature at (or close to) its joints. Reference magnetism is an influential proposal in the foundations of reference, and should be independently attractive to many naturalistic realists. We acknowledge, however, that it involves significant metaphysical and epistemological commitments, which we sketch and motivate (§3). We then use these commitments to defend our account against three salient objections (§4). In closing, we outline an alternative way of appealing to reference magnetism in response to Moral Twin Earth for those who find the commitments of our first approach too much to swallow (§5). In short, this paper both undermines Horgan and Timmons’s ambitious attempt to refute naturalistic normative realism on semantic grounds, and develops a substantive metasemantic proposal that should be of interest to many normative realists.

1 The Twin Earth challenge to naturalistic normative realism

We begin by introducing the background necessary to understand the central argument of our paper. We first sketch the naturalistic normative realism that we will defend against Horgan and Timmons’s challenge. We then explain the basic structure of Horgan and Timmons’s challenge, and why they take it to generalize to all plausible variants of such naturalistic realism.

1.1 Naturalistic normative realism

Recent metaethics has been marked by a shift in focus from morality narrowly understood to practical normativity, construed broadly.\(^1\) We take this shift in focus to represent real progress: there are strong reasons to take practical normativity to be the fundamental object of metaethical analysis. As we characterize it, naturalistic normative realism (henceforth *naturalistic realism*, for brevity) is a family of views about the nature of practical normativity.

We can provisionally characterize the notion of practical normativity at issue

---

\(^1\)For example, see Enoch (2011), Gibbard (2003), Schroeder (2007, 2008), Street (2008), and Wedgwood (2007).
as follows. Sometimes different sorts of considerations for or against an action can conflict. This can be true whether narrowly moral considerations are central or irrelevant by the agent’s lights (e.g. ‘Shall I keep my promise or not?’ vs. ‘Shall I study philosophy or poetry?’). In many such cases, it is natural for an agent to ask a central deliberative question: what ought I to do? In virtue of this central role in deliberation, the practical ‘ought’ appears to be clearly and distinctively normative.

We will take naturalistic realism about practical normativity to consist in four claims. The first is descriptivism: this is the claim that ordinary declarative sentences in which ‘ought’ features—sentences like ‘Alice ought to intervene’—are to be understood as representing the world as having a certain character. This contrasts with views on which normative words mark distinctive speech-acts, or conventionally express desire-like psychological states.

The second claim is non-indexicality. Some words are indexicals or mark implicit relativity. For example, consider assertion of ‘It is raining here’ or ‘Driving on the left is illegal’. Interpreting sentences containing such words requires filling in certain information from the context of utterance, such as the place of utterance or the intended legal system. Relativistic and contextualist metaethical views (e.g., Björnsson and Finlay (2010)) take fundamental normative terms to work in roughly this way, while naturalistic realism denies this.

The third claim is that normative properties are metaphysically explanatory. (Throughout, we use ‘properties’ broadly, to refer both to monadic properties, and to relations; for a relevant discussion of metaphysical explanation, see Fine (2001).) This commitment distinguishes realism from the error theory, which claims that, as with Early Modern witch discourse, there is nothing in the world for normative thought and talk to refer to. It also distinguishes realism from the quasi-realist program in metaethics (c.f. Blackburn (1993); Gibbard (2003)), which claims that we can ‘earn the right’ to claim that normative sentences are descriptive, that these sentences are true or false, and that there are normative facts. The quasi-realist aims to achieve all of this without invoking normative facts or properties as explanantia. The explanatory dimension of realism likewise rules out quietist realism, which rejects expressivism but echoes the quasi-realist’s denial that normative talk involves substantive metaphysical commitments about

---

2We do not wish to legislate how the term ‘realism’ is to be used, and acknowledge that there are uses of the term which do not match the description we give below. We do think, however, that our characterization captures one intuitive use of the term.

3One might try to arrive at a unified conception of realism by subsuming cognitivism and non-indexicality under the dimension of explanatory metaphysics. We leave this a question for another time; for more, see Dunaway (MSb).

4See Mackie (1977, Ch. 1) for (what is most naturally read as) error theory only about categorical normativity. However, Bedke (2010) argues that Mackie’s case extends to all normativity, and Streumer (2013) also suggests an error theory about practical normativity.

5Compare Dreier (2004); see also Dunaway (MSa) for discussion.
the normative.\footnote{Scanlon is the leading example of a quietist in this sense who simultaneously claims to be a non-naturalist realist (Scanlon (1998, 55-64), Scanlon (2003), Scanlon (2014, Ch. 2)). For critical discussion see ?. Compare also Dworkin (1996, 2011), Kramer (2009), and Skorupski (2010).}

Finally, a realist view is \textit{naturalistic} if it takes normative properties to be reducible to, or metaphysically continuous with, the sorts of properties discovered by the sciences. This is inconsistent with \textit{Moorean non-naturalism}, according to which the normative is an additional, \textit{sui generis} component of reality.\footnote{Defenders of Moorean non-naturalism include of course Moore (1903), as well as Enoch (2011), Fitzpatrick (2008), Huemer (2005), and Shafer-Landau (2003). For a more careful discussion of the naturalism/non-naturalism distinction see McPherson (2013).}

\subsection*{1.2 The Normative Twin Earth challenge}

Horgan and Timmons’s argument is explicitly posed as an objection to naturalistic moral realism. However, as we have noted, our focus is on practical normativity, rather than morality. Because of this, we will introduce a version of Horgan and Timmons’s argument that is adapted to target naturalistic realist theories of reference-determination for ‘ought’. This is the \textit{Normative Twin Earth} argument. We take this to be a charitable extension of Horgan and Timmons’s original Moral Twin Earth argument: the premises of this argument are at least as plausible when applied to ‘ought’, and the adapted argument closes off some important avenues of response to the original argument (as we explain in n. 11 below). In light of these considerations, we will freely adapt Horgan and Timmons’s argument to apply to normative realism. We will also adapt the views of some of their interlocutors, who also tend to frame their discussions in terms of morality.

Horgan and Timmons’s challenge can helpfully be introduced by rehearsing a now-familiar dialectic concerning G. E. Moore’s ‘open question argument’ (1903, §13). Adapted to ‘ought’, the Moorean claim is that for any proposed naturalistic analysis\footnote{By ‘analysis’, we just mean any (possible very complex) expression that (i) doesn’t contain the term under analysis (in this case, ‘ought’), and (ii) refers to a property that is necessarily co-extensive with the property referred to by the term under analysis (in this case, obligation). While there are uses of ‘analysis’ that suggest stronger relations between the \textit{analysans} and \textit{analysandum}—such as conceptual connections, reduction, or isomorphic ‘structure’ (cf. King (1998))—we wish to explicitly avoid these stronger connotations in our use of the term. Our use allows, for instance, that Moorean non-naturalists can accept that there are correct naturalistic analyses of even fundamental normative terms.} \(N\) for ‘ought’, the question:

\begin{quote}
I know that doing such-and-such is \(N\), but ought I to do it?
\end{quote}

could be asked sensibly, without displaying conceptual confusion or lack of semantic competence. Put differently: all naturalistic analyses of ‘ought’ will have an \textit{open feel} for speakers competent with the term. We will grant this controversial Moorean claim for the sake of argument.
Once granted, this claim creates pressure on theories of reference for normative terms to explain the existence of the open feel among competent users of ‘ought’. One attractive strategy begins by appealing to other terms that display the same feel, such as natural kind terms. As Saul Kripke (1980) and Hilary Putnam (1975) pointed out, someone might competently deploy a term like ‘water’, without knowing that water is H₂O. For such a speaker, the question:

I know that the stuff in this glass is H₂O, but is it water?

would have a similar open feel. There are many potential ways to explain the open feel of this question, but these examples suggest that doing so requires appeal to reference-determining mechanisms that operate independently of the knowledge of competent speakers.

Richard Boyd (1989) offered an account of one such reference-determining mechanism for natural kind terms, which he then applied to moral terms as well. Simplifying greatly, Boyd’s proposal was that the reference-determining relation is causal regulation. Because a kind can causally regulate a speaker’s use of a term without the speaker knowing that it is this kind which is doing the regulation, this account entails that competent speakers can use natural kind and normative terms without knowing which kind they refer to, under a naturalistic description. Adapted to ‘ought’, this account thus explains why ‘ought’-analyses (like ‘water’-analyses) have an open feel.

Horgan and Timmons’s initial Moral Twin Earth arguments (1991; 1992a; 1992b) were directed against Boyd’s account. Their argumentative strategy was straightforward. If natural kind terms are (à la Boyd) a helpful model for the metasemantics of normative terms, then normative terms should mimic the distinctive semantic features of natural kind terms as well. Horgan and Timmons argued that moral terms fail to mimic one such feature famously illustrated by Putnam’s Twin Earth thought experiment (1975, 222-4).

Briefly, that thought experiment is as follows. Imagine a Twin Earth that is identical to Earth except that the stuff in their lakes and streams and sinks and bodies is not H₂O. It is rather a complex chemical (which Putnam dubbed ‘XYZ’) with very similar macro-qualities to H₂O, but a radically different microstructure. On Earth, we typically apply the English term ‘water’ to samples of H₂O. The Twin-English word ‘water’ is typically applied to samples of XYZ. Counterparts on Earth and Twin Earth are otherwise very nearly qualitative duplicates: in particular, they have near-identical dispositions to token the term ‘water’ (in their respective languages) of the clear potable liquids in their environments. Putnam’s observation was that despite this fact, it is plausible that we refer to H₂O with our term ‘water,’ and our counterparts refer to XYZ with theirs.

Boyd’s causal regulation theory of reference-determination seemingly explains this datum: we and our twins refer to different things by our use of our words ‘water’, on this view, because different natural kinds in our respective environments (H₂O vs. XYZ) causally regulate our usage.
It is natural to suspect that, if the reference-determining mechanisms for ‘water’ and ‘ought’ are the same, and if these predict the existence of a Twin Earth scenario for ‘water’, then there should be a similar Twin Earth-style scenario for ‘ought’. Horgan and Timmons show that there is indeed such a scenario on the Boydian theory: imagine that the difference between Earth and Normative Twin Earth is that the property causally regulating the use of the word ‘ought’ on Earth is a consequentialist property (e.g., the maximization of happiness). By contrast, the property regulating the use of the word ‘ought’ on Normative Twin Earth is a deontological property (e.g., the maximization of happiness constrained by some prohibitions, such as on promise-breaking). Crucially, besides these differences, the role of the words ‘ought’ on the two planets—in deliberation, self-monitoring, interpersonal criticism, etc.—is stipulated to be nearly identical.

Because the case has been tailored to have two different properties causally regulating use of ‘ought’ on the two planets, Boyd’s theory entails that these words do not corefer. Just as ‘water’ is regulated by different properties on Earth and Twin Earth, and so refers to different properties in the mouths of Earthlings and Twin Earthlings, ‘ought’ is likewise regulated by different properties on Earth and Normative Twin Earth, and so refers to different properties in the mouths of Earthlings and Normative Twin Earthlings.

In order to turn this result into an objection to Boyd’s theory, one needs a further assumption about the relationship between reference and disagreement. Suppose that Alice arrives on Twin Earth, meeting her twin. They appear initially to mean the same thing by their uses of ‘water’. They then learn that the stuff in Alice’s glass is XYZ and not H₂O. Suppose that Twin-Alice then says (in Twin English) ‘that glass is full of water’, and Alice says (in English) ‘that glass is not full of water’. This is plausibly a merely verbal disagreement, not a real disagreement: it is like an apparent disagreement over whether Athena is Australian, when we are talking about two different Athenas.

These cases would be neatly explained by a thesis linking coreference and real disagreement, which Horgan and Timmons take normative realists to be committed to: if two communities are capable of having substantive disagreements with their use of a term, then the term in their mouths has the same referent (e.g. 1992b, 165). Since—according to the Boydian theory—the speakers on Earth and Normative Twin Earth don’t refer to the same property, they do not substantially disagree in their utterances of superficially inconsistent ‘ought’-sentences. Their disagreement is merely verbal, just like Alice’s disagreement with her twin. But this consequence of Boyd’s theory is implausible: when you apply ‘ought’ to a given act (perhaps an act of happiness-maximizing promise-breaking), and thereby claim that you ought to perform that act, while your twin refrains from applying their functionally similar term ‘ought’ applies to the same

---

9Horgan and Timmons move from Boyd’s talk of ‘kinds’ doing the regulating work to talk of ‘properties’ doing so. One might worry about the legitimacy of this alteration. But see Soames (2002, 259 ff.) for an account of how to extend Kripkean semantics for natural kinds to at least some properties.
act, this is plausibly an instance of real—and not merely verbal—disagreement (1992b, 164-5).

With these points in hand, Horgan and Timmons’s case against Boyd can be summarized simply. In the Normative Twin Earth case, it is plausible that we can have real normative disagreements with our Twins, using ‘ought’. And this requires coreference of our words ‘ought’. But the case has been constructed to ensure that Boyd’s theory entails that our word ‘ought’ does not corefer with the Twins’ word ‘ought’. So Boyd’s theory is implausible.

This argument should worry the naturalistic realist for at least two reasons. First, it shows that an initially promising approach to explaining the openness phenomenon is in fact highly problematic. Second, nothing in this argument casts doubt on Boyd’s account of reference-determination for natural kind terms. If generally attractive theories of reference-determination fail for normative terms, the normative realist is seemingly in danger of needing to posit an unappealingly disjunctive overall account of reference-determination, in order to save her view. In light of these points, Horgan and Timmons’s argument— even when applied only to Boyd—shows at the very least that the realist surrenders a substantial hostage to metasemantic fortune.

1.3 The challenge generalized

Horgan and Timmons do not rest content with this already substantial result. They have gone on to apply variants of the argument to other salient naturalistic realist accounts of reference-determination (1996; 2000; 2009). More ambitiously, however, they have suggested that they need not attack such accounts one-by-one. Rather, they suggest that their argument generalizes to any naturalistic theory of reference (2000, 149). Their basic line of reasoning can be summarized as follows. Any theory of reference determination for ‘ought’ that the naturalistic realist offers must generate plausibly determinate referents: it must be, for instance, that we determinately refer to deontological, rather than consequentialist properties. Moreover, such determinacy cannot be grounded solely in conditions required for speaker semantic competence: the open question phenomenon entails that competent speakers needn’t know which property bears the reference-determining relation to their use of the term. According to Horgan and Timmons, any theory with these features must predict the existence of a Twin Earth scenario:

it appears that whatever relevant, putatively reference-fixing, relations we bear to certain natural properties (e.g., to consequentialist functional properties), there will be a Twin Earth scenario in which the moral twin earplings bear the same kinds of relations to distinct natural properties (e.g., deontological functional properties). (2000, 146; see also their 1992b, 167.)

We conjecture the following line of reasoning leading to this ambitious conclusion. Begin with the assumption that any adequate theory of reference-
determination for normative terms would need to explain the open feel of ‘ought’ analyses. In the original Twin Earth case, Putnam famously and plausibly proposed that ‘water’ rigidly (and hence necessarily) refers to H\textsubscript{2}O. However, the fact that the stuff in our environment that ‘plays the water-role’ is H\textsubscript{2}O is contingent. Call this latter sort of fact contingency in reference-determination. We don’t need the details of Boyd’s account of reference-determination to see that this contingency in reference-determination will (one way or another) explain why competent speakers can find ‘water’-analyses to have an open feel. Any adequate theory of reference-determination for natural kind terms will imply that one needn’t know the contingent fact that H\textsubscript{2}O plays the water-role in order to be competent with ‘water’. This might in turn suggest that any adequate explanation of the open feel of a term must similarly appeal in this way to contingency in reference-determination.\textsuperscript{10}

The role of this sort of contingent fact in fixing reference is exactly what permits the construction of a Putnam-style Twin Earth case, however. To construct such a case, one simply holds fixed all else that is relevant, and varies the crucial contingent fact. In the case of ‘ought’, the relevant features to hold fixed include the role that ‘ought’ plays in deliberation, self-monitoring, interpersonal criticism, and the like. Since these do seem evident to competent speakers, the relevant contingent facts will need to be something only contingently connected to these roles. And this seemingly puts Horgan and Timmons in a position to construct a reductio of any possible naturalistic theory of reference-determination for ‘ought’, because the following (alleged) commitments of naturalistic normative realism are jointly inconsistent:

**Contingency** The open feel of normative and natural kind analyses is explained by the contingency in reference-determination for normative and natural kind terms.

**Twin Earthability** If a term exhibits contingency in reference-determination, then it will be possible to construct Twin Earth scenarios for that term, where speakers use the term to refer to a different property than the property it refers to in the actual world.

**Disagreement** In some of the Twin Earth scenarios for normative terms, speakers in the scenario can use ‘ought’ to have real (and not merely verbal) normative disagreements with speakers in the actual world.

**Coreference** Real normative disagreement with speakers in such scenarios requires coreference of ‘ought’.

\textsuperscript{10}We are indebted to Mark Schroeder for suggesting something like this hypothesis to us.
We saw the motivations for Disagreement and Coreference in §1.2.\textsuperscript{11} And we have just sketched initial motivation for Contingency and Twin Earthability. But together, this tetrad suggests that naturalistic realism is hopeless: it suggests that no naturalistic realist theory can explain the open feel of ‘ought’, while respecting Disagreement and Coreference. Notice that the argument against Boyd is just an instance of this general schema. His theory implies Contingency, and Horgan and Timmons aim to show by construction of their cases that it thus satisfies Twin Earthability.

There are a number of potential ways to reply here. One might try to refute Disagreement by example, by showing that given a particular plausible reference-relation, any communities that refer to different properties will also intuitively disagree only verbally (Merli (2002, 215 ff.)). Alternatively, one might argue that some naturalistic theories of reference determination are in a position to dismiss the probative force of Disagreement on principled grounds (Dowell (MS)). In this paper we will not pursue either of these strategies. Instead, we will grant (for the sake of argument) that if Contingency were correct, then the Normative Twin Earth argument would generalize to cast doubt on naturalistic normative realism in any form.

We will argue that the naturalistic normative realist can and should reject Contingency for normative terms, and thereby reject the possibility of constructing the relevant Twin Earth counterexamples. There is a general reason to be suspicious of the assumption that any fact that suffices to explain the open feel of an analysis of ‘ought’ will also be metaphysically contingent. The open feel is psychological-cum-epistemic phenomenon, and contingency is a metaphysical phenomenon. This gives us a recipe for constructing a counterexample to Contingency: we must find a theory of reference on which necessary truths can play a significant role in determining the referent of a term, without knowledge of those truths being a condition on being a competent speaker. Happily, there is a highly attractive theory of reference determination which perfectly fits this recipe, as we now explain.

\section{Reference magnetism}

In this section we show that Contingency is false, in virtue of an influential and attractive approach to reference-determination: reference magnetism. We show\textsuperscript{11} Horgan and Timmons are arguably in a stronger position vs. normative realism (the position we wish to defend) than against moral realism (their explicit target). This is because one of the most important sorts of response to Horgan and Timmons on behalf of the moral realist challenges Coreference, by suggesting that our judgments of real disagreement might be explained not by the synonymy of the relevant moral terms, but by underlying normative (but non-moral) disagreement. For example, Merli suggests that the apparently real disagreement in the Moral Twin Earth cases may be non-moral but normative disagreement about what to do ((2002, 232-3), cf. also Copp (2000, 123)), and Plunkett and Sundell (2013) argue that it can be understood as normative disagreement about which broadly moral concepts to deploy. This sort of strategy is much harder to deploy in response to a Normative Twin Earth challenge.
that reference magnetism both provides a counterexample to Contingency, and can be used to explain the plausible judgments that Horgan and Timmons appeal to in Normative Twin Earth cases. Further, we show that adopting reference magnetism as a (partial) theory of reference-determination for both normative and natural kind terms permits us to explain the independently plausible differences between Putnam’s Twin Earth case and Normative Twin Earth, without requiring any ad hoc assumptions about reference-determination for normative terms. We begin by briefly explaining and motivating the idea of reference magnetism.

2.1 A brief introduction to reference magnetism

Many philosophers have in recent years been attracted to a pair of theses that can be found in David Lewis’s (1983). The first thesis is that some properties are metaphysically more significant than others: they constitute the ‘joints of nature’. Lewis calls these properties, variously, ‘perfectly natural’, ‘elite’, and ‘joint-carving’. To avoid confusion with the distinct notion of the natural at play in metaethics, we will call these elite properties. Lewis thinks that fundamental physical properties like being negatively charged are good candidates for being perfectly elite. Eliteness is a gradeable phenomenon: being negatively charged is more elite than being acidic which in turn is more elite than being colored grue.

The second Lewisian thesis is that elite properties are ‘reference magnets’: more elite properties are easier to refer to than less elite properties. A more rigorous characterization of this phenomenon is difficult in the abstract because reference magnetism is by nature a partial theory of reference-determination that needs to be combined with other ingredients. As such, the theory is silent on what other ingredients play a role in determining reference. However, the flavor of the thesis can be given by examining its contribution to a toy complete theory of reference.

Consider two illustrative examples. First: let a gappycat be the mereological sum of most of the time-slices of a cat, where those that are not included are replaced with the contemporaneous time-slice of the nearest non-identical cat. Given plausible assumptions about the use of ‘cat’, gappycats will fit as well as cats do with this use. But on the combined theory, the higher degree of eliteness of cat would break the tie in fit with use, and single out the kind cat above the kind gappycat as the referent of ‘cat’.

The maximizing structure of the combined theory permits reference magnetism to do more than break ties in this way. Consider a second example that brings

---

12van Roojen (2006, 179-83) also proposes using reference magnetism as part of his response to the Moral Twin Earth challenge. While we are inclined to accept much of what van Roojen says, we explore the virtues and vices of the approach in much greater detail. van Roojen’s approach also differs from ours in two important ways. First, he weds reference magnetism to a knowledge-maximization constraint on reference; we are proceeding here without that additional machinery. Second, van Roojen’s notion of naturalness (i.e. eliteness) is what he calls ‘discipline-relative’; we instead prefer to work with a single unified notion of eliteness.

13See also Lewis (1984, 1986). Contemporary discussions can be found in Hawthorne (2007) and Sider (2012).
Reference Magnetism as a Solution to the Moral Twin Earth Problem

this out. Suppose that a community of early humans discovered both gold (i.e. the element Au), and ‘fool’s gold’ (i.e. iron pyrite), and were unable to tell the difference between the two substances. Suppose that they introduced the term ‘goldite’ to refer to the substance they were discovering. Consider two candidate extensions for the referent of ‘goldite’: the element Au, and the disjunction of Au with iron pyrite. The disjunctive referent provides a better fit with use. But because Au is more elite than the disjunction, it isn’t implausible to say that the former best maximizes the combination of eliteness and fit with use in this community. The toy theory, then, can say that this community referred to gold with their term ‘goldite’.

The toy theory, of course, is just that. However, reference magnetism can provide the attractive structural features advertised by the two examples (indeterminacy reduction and referential bias towards the joints of nature) when combined with sophisticated alternatives to the bare idea of fit with use. It may help to have an illustration to fix ideas. Figure 1 illustrates the relationship between other reference-fixing factors and the relevant reference magnet, given such a combined theory of reference:

![Figure 1: natural joints as reference magnets](image)

Here, we think of an **intension** as a function from possible worlds to term-referent pairs, or less strictly, as the distribution of the referent of a term across possible worlds. Please note that this figure simplifies for expository purposes. It assumes that that the rest of our vocabulary has already had its intensions fixed, thereby eliminating the threat of radical referential indeterminacy through inferential connections to other (assumed determinate) terms. What we are left with is local referential indeterminacy suggested by use of a specific term \( t \) (or some other candidate reference-determining mechanism). In the figure, the smoothness of the oval represents the eliteness of the reference magnet. The thickness of the irregular line represents local indeterminacy of reference if we held fixed only use (e.g.). The gaps between the two lines represent the way that reference magnetism can override fit with use in determining reference (as in the case of ‘goldite’).

Why find the idea of reference magnetism appealing? Its defenders have emphasized the impressive philosophical work that it can do. Most notably, it
Reference Magnetism as a Solution to the Moral Twin Earth Problem

provides principled and unified solutions to the deepest puzzles about reference, including Quine’s case for the indeterminacy of translation, Putnam’s permutation argument, and the rule-following paradox from Kripke’s Wittgenstein. And it promises to do these things in conjunction with an attractive approach to metaphysics that has eliteness at its center.

2.2 Reference magnetism and Normative Twin Earth

With this partial sketch in hand, we can explain why reference magnetism is well-suited to help the naturalistic realist reply to the Normative Twin Earth challenge. To begin, we explain why the existence and plausibility of reference magnetism refutes Contingency, which is central to our explanation of Horgan and Timmons’s confidence that no naturalistic realist can accommodate Normative Twin Earth. Contingency assumes plausibly that the openness of candidate analyses of ‘ought’ is explained by the fact that competent speakers can be ignorant of some of the reference determining facts for ‘ought’. However, it also assumes that in any plausible theory of reference-determination, every reference-determining fact that competent speakers can be ignorant of is contingent.

A theory of reference-determination that incorporates reference magnetism has two features that together undermine this latter assumption. Competent speakers can be ignorant of which properties are elite. But if a property is elite, it is necessarily so: for example, it makes little sense to imagine that negative charge is elite in the actual world but that there are possible worlds in which being negatively charged is instantiated but not elite (Lewis (1986, 61 n. 44); Dorr and Hawthorne (2013, 31)). In light of these two features, reference magnetism is a counterexample to Contingency. Reference magnetism can explain why ‘ought’-analyses can seem open: one needn’t know which candidate analysis refers to the most elite property. But this is ignorance of a necessary fact, not a contingent one. Contra Contingency, explaining the open feel does not require contingency in the reference-determining relation.

Reference magnetism is thus a counterexample to Horgan and Timmons’s attempt to generalize Twin Earth-style arguments against naturalistic realism. However, this isn’t a mere technical complaint about the formulation of Contingency. As we now show, the features that have this implication also provide the normative realist with an attractive explanation of the very Normative Twin Earth judgments that Horgan and Timmons rightly find plausible. This suggests that there can be no general argument against naturalistic realism on Twin Earth grounds, since the naturalistic realist can always appeal to reference magnetism.

See Quine (1960), Putnam (1981), and Kripke (1982). See Sider (2012) for an extended defense and application of this approach to metaphysics. And see Dorr and Hawthorne (2013) for a useful critical overview. Of course, one might argue against either reference magnetism per se or our metaethical use of it. Our point is that these would necessarily be piecemeal arguments, and not applications of Horgan and Timmons’ master argument.
The application of reference mangetism to Normative Twin Earth can be illustrated using the toy theory of reference introduced above. To begin, suppose (just for the moment) that the deontological property is much more elite than other candidate referents for ‘ought’ (it is like the smooth oval in Figure 1). Because eliteness facts are necessary, the deontological property will be highly elite in all possible worlds, and hence on Normative Twin Earth. The deontological and consequentialist properties have quite similar intensions. But we know that reference magnetism can outweigh some variance in use. So even if we stipulate that the consequentialist property is the best fit for use of the Twins’ word ‘ought’, the toy theory of reference-determination would entail that our twins refer to the deontological property with their word ‘ought’.

We have noted that competence with ‘ought’ does not require knowing which properties are elite. So, relax the illustrative assumption that the deontological property is highly elite, and replace it with the assumption that there is some single highly elite property in this vicinity. Here, reference magnetism suggests that we and our Twins refer to the same property with our uses of our respective terms ‘ought’. In other words, reference magnetism provides a principled explanation of the core semantic judgment that is the heart of the Normative Twin Earth challenge.

2.3 ‘Water’, again

It is not enough for the normative realist to give a plausible explanation for Normative Twin Earth cases. These results should fall out of a general theory of reference-determination that applies to normative and non-normative cases alike. And this might seem to leave the normative realist in a bind, because Horgan and Timmons appear to show that contrasting judgments are plausible in the Twin Earth and Normative Twin Earth cases (judgments of merely verbal and real disagreement, respectively). It is thus natural to worry that reference magnetism will, if it is sufficiently general, have implausible implications when we consider Putnam’s original Twin Earth scenario. As we now show, this is not so: the account of reference magnetism that we offer straightforwardly entails the intuitively plausible contrast between the two sorts of Twin Earth cases.

This is because of an important contrast between Putnam’s Twin Earth case and Horgan and Timmons’s Normative Twin Earth. It is easy to imagine a world with XYZ but no H₂O: perhaps such a world is one where there are no protons, electrons, etc., but rather alien subatomic particles that enter into alien bonding relations; the resulting substance is one which differs substantially from H₂O at the micro-level but behaves (very nearly) like water at the macro-level. By contrast, it doesn’t make sense to think that a world could have options which maximize happiness out of a certain constrained set of options (thereby instantiating the deontological property), while lacking options that maximize happiness relative to the total set of the agent’s options (thereby instantiating the consequentialist property).
Horgan and Timmons notice this contrast, but go on to dismiss it:

One might think that this difference is significant and that it can be exploited by the moral naturalist to her advantage. Clearly, there is just such a difference, but far from helping the moral naturalist overcome the Moral Twin Earth argument, the fact that both planets are ones in which both consequentialist and deontological properties are eligible referents for moral terms makes things worse for the naturalist [by exacerbating the threat of indeterminacy]. (2000, 145)

However, reference magnetism inverts the significance Horgan and Timmons ascribe to the difference. On the one hand, reference magnetism is tailor-made to address the indeterminacy worry that they raise. On the other, the contrast Horgan and Timmons note permits the reference magnetist to explain the intuitive appearance that in the relevant Twin Earth cases, uses of ‘ought’ corefer, but uses of ‘water’ do not.

In Putnam’s case, the Earth English word ‘water’ picks out H\textsubscript{2}O, which is intuitively a highly elite kind, but on Twin Earth, the lakes and streams and showers are not filled with H\textsubscript{2}O, but rather a different kind: XYZ. Because the relevant Twin Earth environments lack H\textsubscript{2}O, reference magnetism cannot plausibly entail that H\textsubscript{2}O is the referent of the Twins’ word ‘water’. By contrast—as Horgan and Timmons correctly grant—the consequentialist and deontological properties will be instantiated in all of the same worlds, and with very similar intensions. If one of them is much more elite than the other, this very feature allows the normative realist to explain why there is coreference across all of the alleged Normative Twin Earth scenarios.

In this section, we have introduced the idea of reference magnetism, and applied it to the Normative Twin Earth challenge. We have emphasized three attractions of this application: first, reference magnetism constitutes a clear counterexample to Contingency, the thesis which seems to underly Horgan and Timmons’s confidence in the generalizability of their challenge. Second, it allows the naturalist to smoothly vindicate the plausible judgment that there could be real normative disagreements between Earth English speakers and Normative Twin Earthlings. Third, reference magnetism provides a unified (partial) account of the reference-determining mechanisms for normative and natural kind terms. This account can explain the ‘openness’ of analyses involving both kinds of term while also explaining the contrast between plausible judgments about Twin Earth cases involving natural kind terms versus those involving normative terms. We take these to be highly impressive results that warrant close attention from normative realists.

3 Metaphysical and methodological commitments

We have just argued that reference magnetism is an extremely promising way for the naturalistic realist to seek to meet the Normative Twin Earth challenge.
However, reference magnetism is not a free lunch. Both the core thesis and its deployment in the case of Normative Twin Earth inevitably entail some significant metaphysical and methodological commitments. The realist has some flexibility with respect to the details of these commitments; in this section, we aim to clarify—and to some extent motivate—a representative package of such commitments that we find attractive. Besides its intrinsic interest, this discussion will put us in a position to address three important objections to our view, in the following section. However, some readers may find the metaphysical and methodological diet that we prescribe here too rich. In §5, we will sketch an alternative way to implement our approach that may be more attractive to those readers.

The view we sketched in §2 not only requires the existence of highly elite properties; it is also committed to the existence of a highly elite property in the vicinity of our use of ‘ought’, and that there be a *unique* such property. These are *metaphysical* commitments. In addition, such a view also requires an account of how we can *know* that these metaphysical claims are true. This is an additional epistemological commitment of our view. These commitments may appear worryingly optimistic. A sharper worry is that our commitments are almost certainly inconsistent with David Lewis’s own substantive account of relative eliteness. We begin by addressing this worry, and use our response to motivate a broader defense of our metaphysical and epistemological commitments.

### 3.1 Lewisian canonical definitions

Lewis’s account has two crucial elements. First, all and only the fundamental microphysical properties are perfectly elite (1984, 228). Second, all properties can be predicated by terms in a *canonical language*: a language that contains only simple predicates standing for perfectly elite properties, and some privileged logical connectives (a complication that we set aside). All not-perfectly-elite properties can be predicated by logically complex ‘definitions’ constructed from the simple predicates, using the privileged connectives. On Lewis’s view, the degree of eliteness of a property corresponds to the length of its definition in this canonical language: the longer this *canonical definition*, the less elite the property (1986, 61; see also 1983, 347). This predicts that *being negatively charged* is more elite than *being furniture*. The former is perhaps perfectly elite (or at least very close to it), while the latter would presumably require an extremely long and complex canonical definition.

If true, this account would render our reply to Horgan and Timmons hopeless. As we have emphasized, our strategy requires that some plausible candidate referent for ‘ought’ (perhaps the deontological or the consequentialist property, for example) is markedly more elite than the other candidates. This is required to secure coreference between our two imagined communities, despite their slightly

---

17See Williams (MS, §2.3) for a related view that does not involve these commitments.
Reference Magnetism as a Solution to the Moral Twin Earth Problem

divergent use. But the deontological property surely does not have a markedly shorter canonical definition in terms of ‘negative charge’, ‘spin’, etc. than the consequentialist property. Both definitions will be highly complex (perhaps even infinitely long) when stated in the canonical language. The same point holds for any reasonable hypothesis about what ‘ought’ picks out. (A contrasting unreasonable hypothesis take obligatoriness to be identical to some perfectly elite microphysical relation.)

The good news for the naturalistic normative realist is that we know for independent reasons that the Lewisian approach to degree of eliteness cannot underpin a plausible theory of reference magnetism. Reference magnetism, we have been emphasizing, is a partial but completely general theory of reference determination. But, as John Hawthorne has forcefully argued, macroscopic candidate referents systematically fail to be distinguished from nearby macro-level objects and properties by virtue of their canonical definitions: the problem we noted for ‘ought’ will also face ‘chair’ or ‘rabbit’, thereby predicting intolerable referential indeterminacy or instability, two of the central problems that reference magnetism was advertized to solve ((2006, 206), (2007, 434); compare also Williams (MS)). Lewis’s problem is arguably even worse than this: as J. R. G. Williams has shown (2007), if certain mathematical structures can be finitely constructed out of perfectly elite properties, then it is possible to prove that intuitively absurd mathematical interpretation of all of our terms better satisfy Lewis’s theory of reference than intuitively plausible referents. Together, these points strongly support Hawthorne’s suggestion that we should revise Lewis, to permit relative eliteness to float free from microphysical definability (2007, 434).

3.2 An epistemology for primitive degrees of eliteness

One might reasonably worry that, by abandoning Lewis’s theory in favor of primitive degrees of eliteness, the theory of reference magnetism becomes objectionably unconstrained, able to be tailored however one might desire. We address this worry by sketching a principled account of how relative eliteness can be determined. This puts us in a position to motivate our account as a naturalistic amendment to Lewis’s own views about the metaphysics and epistemology of perfect eliteness.

To begin, consider Lewis on perfect eliteness. Metaphysically, it is a primitive: what it is to be perfectly elite can’t be defined in further terms. But that does not make Lewis’s account objectionably unconstrained. This is because Lewis accepts a second, epistemological thesis: that the discipline of fundamental physics is our crucial epistemic guide to which properties are perfectly metaphysically elite.

Next, consider some methodological reasons to be suspicious of Lewis’s definitional approach to less-than-perfect eliteness. Practicing scientists offer explanations that appeal to higher-level scientific kinds and properties, for which even the

---

18We thank Laura Schroeter for powerfully pressing a variant of this worry.
in-principle availability of physicalistic reductions are highly controversial. And even if canonical definitions for each of these higher-level kinds could be found, they certainly wouldn’t typically be short and simple: imagine, for example, what the Lewisian canonical definition of *gene* might look like. In light of these points, Lewis’s assumption that fundamental physics *alone* among the sciences provides distinctive insight into the elite structure of reality appears dubious from a naturalistic methodological perspective.

These points motivate a principled departure from the Lewisian picture. First, we can take *relative* eliteness to be metaphysically primitive, instead of just perfect eliteness. Thus, on our view, one property can be much more elite than another property which has an equally long canonical definition in microphysical terms.\(^{19}\) We couple this metaphysical thesis with the obvious naturalistic amendment to Lewis’s dubious epistemology of eliteness:

**Liberal** One can know which properties are highly elite by knowing which properties are countenanced by naturalistically credible theoretical disciplines including (but not limited to) physics.

By adopting **Liberal**, our account—like Lewis’s—ensures that theorizing about relative eliteness is not objectionably unconstrained.

Our account includes more primitivism than Lewis’s, and any commitment to additional primitivism entails some cost to a theory. However, this in our view is more than made up for because our account avoids the two problems we identified for Lewis. First, because our account allows macro-eliteness to float free of microphysical eligibility, it allows that ‘rabbit’ can refer to a relatively elite kind, and hence have a determinate and stable referent. Second, **Liberal** allows that higher-level sciences (and not just physics) can successfully reveal the elite structure of reality.

**Liberal** leaves open the question of which disciplines count as naturalistically credible, and hence as windows into relative eliteness. This, we take it, is a matter for substantive debate in the philosophy of science. For our purposes, a crucial question concerns whether normative theorizing counts as a naturalistically credible theoretical discipline. Our framework is compatible with any of a variety of influential ways for naturalistic realists to answer this question. For example, a naturalistic realist might be inspired by Richard Boyd’s sketch of moral theorizing as methodologically continuous with the sciences (1989, §4.4), or Geoff Sayre-McCord’s defense of the idea that ethics is a methodologically autonomous but still naturalistically acceptable discipline (1997). Alternatively, a naturalistic realist might defend a reduction of the normative facts in terms of facts that can in turn be investigated by a naturalistically credible discipline (for example, see Mark Schroeder’s recent case for a psychologistic reduction of normative facts in his (2007, Ch. 4)). The essential point is that any adequately naturalistic realism

\(^{19}\)Hawthorne (2006, 206) calls this view ‘emergentism’. See also Hawthorne (2007).
Reference Magnetism as a Solution to the Moral Twin Earth Problem

already needs a naturalistically acceptable epistemology for the normative. And our account is compatible with the central plausible proposals in this arena. A working account of reference magnetism for normative terms needs some substantial metaphysical and epistemological assumptions, if it is to deliver what was promised in the previous section. In this section, we have sketched an account of why these assumptions should seem independently plausible to naturalistically-minded philosophers.

4 Three objections to the reference magnetic solution to Normative Twin Earth

We have just motivated some of the broader philosophical assumptions that are required for our solution to the Normative Twin Earth challenge to work. However, one might think that our solution fails even if one grants these assumptions. In this section, we consider three objections that purport to reveal quite general problems with the reference magnet approach to Normative Twin Earth.

4.1 The assumption of a unique reference magnet

To begin, consider a worrying analogy: suppose that Twin Earthlings play a game that appears identical to chess, except that the Twins proceed as if their “Knights” can move diagonally. It would be absurd to think that, due to reference magnetism, we and our Twins are playing the same game—one group of us systematically less competently than the other! But our reference-magnetic approach is committed to an exactly parallel claim about Normative Twin Earth. Our account provides an explanation of the asymmetry: chess and ‘Twin chess’ are plausibly approximately equally elite games. Because of this, a theory of reference determination that includes reference magnetism will entail that we refer to chess with our use of ‘chess’ and the Twins refer to Twin chess with theirs. By contrast (as we have noted previously) our account insists that in the normative case, there is a single highly elite property in play. This assumption is what explains coreference on Normative Twin Earth, given reference magnetism.

With the chess analogy before us, however, this assumption might seem like a bit of ad hoc wishful thinking. We make two points in reply. First, any use of reference magnetism with respect to higher-level theoretical terms will need to make similar claims. For example, a reference-magnetic explanation of how biologists manage to refer to genes instead of some nearby gerrymander will need to claim that the difference between biology and other biology-like disciplines is not like the difference between chess and Twin chess. Second, recall from §1 that one of the core metaphysical commitments of naturalistic realism is that the property of obligatoriness is metaphysically explanatory. A central feature of the elite property approach to metaphysics is that only highly elite properties can be metaphysically explanatory. Thus naturalistic realism, as we conceive of it, already includes a commitment to normative properties having a distinctive mark of highly elite properties.
This objection contends that even if reference magnetism can accommodate the most familiar instances of the Normative Twin Earth challenge, it will inevitably fail to a structurally similar thought experiment. The objector notes that reference magnetism is especially plausible in the canonical case because of the substantial intensional overlap between the deontological and consequentialist properties. But suppose we instead consider linguistic communities whose use of a term seems to best fit a quite different intension, while continuing to hold fixed the role of ‘ought’ in deliberation, self-monitoring, interpersonal criticism, etc.

To make this idea vivid, consider Normatively Inverted Twin Earth. The case is similar to the standard Normative Twin Earth scenario we have been discussing: the only difference is that, rather than appearing to track the deontological property, on Normatively Inverted Earth, values appear to be inverted: (almost) all and only what we take to be obligatory is treated there as prohibited, and (almost) all and only what we take to be prohibited is treated there as obligatory. (The ‘almost’s are inserted because what is obligatory includes, e.g., criticism of people who do certain things that are prohibited, and the Normatively Inverted Earth case will not want to invert those relations. More on this below.) To the extent that it is possible, the social and deliberative functions of the Inverts’ word ‘ought’ are stipulated to otherwise be like those of our word.20

It is worth emphasizing that this case is dialectically weaker than Horgan and Timmons’s case in two respects. First, it is quite difficult to concretely imagine such a community, so we should be cautious about placing much weight on cases like this. Second, it is not at all obvious that the Inverts do semantically disagree with us in their use of their word ‘ought’. Despite these points, let us grant for the sake of argument that the Inverts do semantically disagree with us in such uses.

The objector thinks that reference magnetism could not explain this (alleged) fact. But we think that two considerations mitigate in favor of (cautious) optimism here. First, recall the realist’s commitment that the especially elite candidate normative properties are very sparsely distributed. In light of this, the best case for the objector is to have the Inverts’ use of ‘ought’ pick out some genuine normative property other than being obligatory. But this will not be easy. One suggestion is that the property of being prohibited, which is likely highly elite on our view, is the referent of the Inverts’ use of ‘ought’. But this is unpromising, because being prohibited has structural features that do not pattern with the relevant uses of ‘ought’. For example, it is common for many of an agent’s mutually exclusive options to be prohibited, but rare (impossible, on some views) for multiple options to be obligatory.

A second pass at the objection might claim that there must be some highly magnetic property distinct from being obligatory that magnetizes the Inverts’ use

---

20 This example is a more extreme version of Hare’s ‘missionaries and cannibals’ example (1952, §9.4), which is arguably the ur-text of Moral Twin Earth.
of ‘ought’. But we see very little reason to accept this claim, once we consider the potential semantic significance of the connections between different normative terms. Intuitively, the property of being obligatory enters into very complex relations with agency, experience, and with self- and inter-personal criticism and interpretation. Different metasemantic theories will accord such apparent relations different sorts of significance, but at first blush, a plausible candidate for the referent of ‘ought’ will need to accommodate many such relations. Thus, when seeking to identify the referent of the Inverts’ word ‘ought’, we prima facie need to find a relatively elite property that also supports plausible connections to plausible relatively elite referents for ‘agency’, ‘self-criticism’, etc. In light of these considerations, we take it to be very difficult to mount a compelling Normatively Inverted Twin Earth case.

4.3 An objection from the autonomy of ethics

A final objection claims that appeal to reference magnetism flies in the face of attractive views about the autonomy of ethics. If reference magnetism is a correct partial theory of how ‘ought’ gets its referent, then a community’s use of ‘ought’ refers to a property partly in virtue of its relative eliteness—a metaphysical property. If this is correct, then it might seem that normative theorizing should defer to metaphysical theorizing concerning which properties are elite. But, the objector insists, this does not seem right: we should determine what we ought to do by engaging in ethical theorizing, not by doing metaphysics.

Stated this way, this challenge can be undercut by appealing to the Liberal naturalistic epistemology of eliteness, introduced above. Consider a parallel case: biologists certainly don’t defer to non-biological metaphysical theorizing about what is elite. Liberal endorses this: biologists can and often do identify relatively elite biological properties by doing biology. A naturalistic normative realist can hope to offer an analogous reply. For example (as we saw above), one way for the normative realist to implement Liberal claims that we can and sometimes do identify relatively elite normative properties by engaging in normative theorizing. On such a view, the objector’s complaint would be inert: normative theorizing just is doing (the relevant) metaphysics.

A more contentious version of the objection might press an alleged disanalogy between normative and biological investigation. While it is plausible that even procedurally ideal biological theorizing could potentially get the fundamental biological facts wrong, one might press this objection by claiming that procedurally ideal normative theorizing could not possibly lead us astray. On this view, elite ‘joints of nature’ might seem irrelevant to determining reference. Here, we are inclined to simply deny the assumed infallibility of procedurally ideal normative enquiry. Such denial is a standard mark of robust forms of realism about a subject

21 Thanks to Pekka Väyrynen for raising a version of this objection.
22 Sturgeon (2002, §III) defends a relevant combination of methodological naturalism and methodological autonomy.
matter, and we see no reason that normative realists should not be robust in this way.

Suppose, however, that we grant the infallibility of procedurally ideal normative enquiry as an analytic truth (for example). This thesis amounts to holding that the intension of ‘outcome of procedurally ideal theorizing’ covaries with the intension of ‘ought’. Once we understand the thesis in this way, it should be clear that it fails to threaten the reference magnet theory at all. Reference magnetism will simply predict that the very same elite property will play the same role in magnetizing reference for ‘procedurally ideal theorizing’ in addition to ‘ought’.

5 Is the language organ a reference magnetist?

We take the case developed in the preceding three sections to constitute a formidable case for the reference magnetism solution to Normative Twin Earth. However, we recognize that not everyone will be comfortable with the metaphysical and methodological assumptions required by the solution as we have sketched it so far. An initial reply to such readers is that Moral Twin Earth is usually advertized as a purely (meta-)semantic challenge to the naturalistic realist. So even if we have succeeded only in forcing the debate into metaphysical and methodological waters, we take that as a (partial) victory. In this section however, we aim to win over the metaphysically wary, by sketching an alternative picture on which the normative realist can appeal to reference magnetism without taking on the sorts of commitments sketched above.

Suppose, then, that one granted part of the story we have described above: that our intuitive reactions to the Normative Twin Earth cases reflect implicit confidence that normative properties are easy to refer to. One might simultaneously deny that normative properties actually are highly elite and hence easy to refer to. From this perspective, reference magnetism provides a correct diagnosis of our judgments about the Normative Twin Earth cases, without actually supporting the claim that those judgments are correct. In terms of the tetrad that we introduced in §2, this amounts to a rejection of Disagreement—intuitions of real disagreement between speakers in Twin Earth scenarios are, on this view, driven by mistaken judgments of the presence of a reference magnet. In this section we explain how the naturalistic realist could develop this account, by appealing to a familiar naturalistic strategy, which is arguably well-motivated in this case.

Many naturalistic realists have emphasized that their theories of the normative will be revisionary or reforming to some extent (cf. Brandt (1979, 10)). For example, Peter Railton proposed a theory of morality that (he admits) fails to accommodate all of the apparent ‘objective prescriptivity’ that is intuitively inherent in moral norms (1986, 201). In offering these reforming theories, naturalists reasonably assume that the best overall theory of a subject will sometimes be inconsistent with some of our intuitive assumptions about that subject. In light of taking common judgments about Normative Twin Earth cases to be incorrect, the proposal that we are considering in this section is a reforming theory in this sense.
As Railton notes, such reforming theories are most satisfying when they are coupled with an explanation of why the folk would have the relevant erroneous beliefs. For example, Railton himself takes the idea of objective prescriptivity to be tempting because people assume that morality cannot have ‘authority’ without it. And he goes on to sketch what he takes to be the outline of an adequate alternative account of this authority. The revisionary approach to Normative Twin Earth that we propose can similarly appeal to an apparently satisfying diagnosis of error.

This diagnosis begins by considering a putatively analogous case: our judgments about ‘red’. It is clear that the property of reflecting light wavelengths between 620 and 680 nm is no more metaphysically elite than the property of reflecting wavelengths between 619 and 679 nm, or the property of reflecting light wavelengths between 621 and 681 nm. Thus—ignoring vagueness, and assuming our toy theory of reference—if ‘red’ refers to one of these properties, the question of which it refers to is entirely a matter of which property fits with use better than the others. In such a case, ‘red’ will exhibit what Hawthorne calls semantic plasticity: the intension of ‘red’ could (with slight variations in use) easily have been different (2006, 196). By the same token, communities which differ only slightly in their use of terms like ‘red’ may end up referring to different properties. For instance, suppose we tend not to apply ‘red’ to most objects that reflect light at 619 nm but do tend to apply ‘red’ to most objects that reflect light at 680 nm. By contrast, our otherwise similar Twins tend to apply ‘red’ to the 619 nm objects but not to the 680 nm objects. Given our simple theory of reference introduced above (or other natural assumptions about reference-determination for ‘red’), these communities will refer to different properties.

But some intuitive judgments about these communities will betray an implicit assumption that ‘red’ is not plastic, but rather semantically stable. For example, we are intuitively inclined to treat the following counterfactual as true:

If we were part of the Twins’ linguistic community, we would still have believed that stop signs are red.

It is however a plausible constraint on counterfactual reports like this that one can truthfully report that the speaker in the counterfactual situation has a belief about redness only if she uses ‘red’ to refer to redness. And given a natural understanding of the case, plus the assumption that ‘red’ is semantically plastic, our ‘Twins’ term refers to a property distinct from redness. For example, when they token the sentence ‘stop signs are red’, their belief is not about redness, but rather is about stop signs reflecting light at wavelengths between 619 and 679 nm.

One diagnosis of this kind of case is that our inclinations to treat counterfactuals of this kind as true reflects a deep but mistaken assumption that ‘red’ is semantically stable. Normative Twin Earth might be another instance of a similar phenomenon: our best accounts of metaphysics and reference-determination for

---

23See Dorr and Hawthorne (MS) for a much more thorough discussion of the issue and available responses.
normative terms might imply that ‘ought’ is semantically plastic. But our intuitive judgments about disagreement might reflect an implicit but mistaken assumption that normative terms are semantically stable. If we seem systematically vulnerable to such errors across domains, and if many such errors appear to have a common source, this would strongly support a revisionary view about our judgments concerning the Normative Twin Earth cases.

One conjecture about such a common source of error is this: perhaps the folk have what we might call an implicit ‘epistemology of eliteness’. Applied to normative talk, this hypothesis would suggest that even if no candidate normative properties are distinctively elite, the cognitive systems responsible for producing judgments of agreement and disagreement, and assessing counterfactual reports of speakers’ beliefs, function as if there were such reference magnets. That is, on the hypothesis we are exploring here, the ‘language organ’ produces judgments that reflect intuitions of semantic stability because the language organ treats normative properties as reference magnets, which can secure semantic stability. (It must be emphasized that—as with most ascriptions of theoretically interesting features to folk psychology—this hypothesis is the possible conclusion of an interpretive inference to the best explanation, not something that we would expect the folk to directly report about their own psychology.)

It is important to distinguish this sort of revisionism from full-blown error theory about normative discourse. The proposal we are exploring rejects some pre-theoretic claims about disagreement and counterfactual report that are apt to be endorsed by competent users of ‘ought’, but doesn’t require an endorsement of error theory. Consider an analogy: the liar paradox is arguably generated by judgments one must find compelling in order to be competent with ‘true’ (see Eklund (2002) for discussion). But that does not entail the error-theoretic thesis that no sentences are true. In both cases, a more plausible response is to offer a slightly revisionary account of our terms: one that rejects as false some claims that competent speakers find compelling.

To reiterate, we find the metaphysical and methodological commitments defended in §3 to be plausible. We have offered the revisionary hypothesis sketched in this section as an alternative for our metaphysics-wary colleagues.

Conclusions

We take Moral Twin Earth-style thought experiments to constitute the most serious challenge to the (meta-)semantic dimension of the naturalistic normative realist’s research program. Such thought experiments appear decisive against some important theories of reference determination for normative properties, suggesting at the very least that the naturalistic realist must surrender some hostages to semantic fortune. However, proponents of the challenge sometimes make a stronger claim: that these thought experiments suffice to refute any otherwise plausible theory of reference determination for normative terms. In this paper, we have diagnosed the apparent deep basis for this stronger claim. This
is the assumption that the only plausible explanations of the alleged openness of all ‘ought’ analyses must appeal to the role of contingent facts in the reference-determining relation. The existence of this contingency is what ensures the alleged availability of a compelling Moral Twin Earth objection to any such theory.

We then showed that the existence of reference magnetism as a serious candidate theory of reference-determination refutes this assumption. On this theory, facts about the eliteness of properties contribute to fixing reference. Such facts are also necessary, but speakers need not know which properties are elite in order to count as competent. This permits an explanation of openness that does not appeal to contingency. We next argued that reference magnetism provides a satisfying explanation of semantic phenomena elicited by Moral Twin Earth cases, explaining why we corefer with our twins. We also showed how it simultaneously explains the apparent contrast between Moral Twin Earth and Putnam’s original Twin Earth case. We take this to be an impressive set of virtues for a candidate theory of reference-determination for normative terms.

Even in light of these virtues, one might worry that committing the naturalistic realist to reference magnetism surrenders too great a hostage to fortune. We offer three points in reply. First, the tremendous work that reference magnetism can do in the foundations of reference makes it a far from idiosyncratic commitment. This has important dialectical consequences: many potential objections to reference magnetism for normative terms risk generalising into objections to reference magnetism for any macrophysical properties whatsoever. Given the work reference magnetism can do, such objections would need to be powerful indeed. Second, as we have emphasized, reference-magnetism is a partial theory of reference determination. Thus, it could be used to amend semantic pictures as diverse as Boyd’s, on the one hand, and the ‘moral functionalism’ of Frank Jackson and Philip Pettit (1995), on the other. It thus permits the naturalistic realist to take various positions on many of the central controversies about the foundations of reference. Third, as we have argued in §5, there is a principled way for the normative realist to offer the idea of reference magnetism in a purely diagnostic spirit, while declining to adopt the central metaphysical or semantic commitments of the view.

In short, the thesis that reference magnetism plays a role determining the reference of normative terms is powerfully motivated on independent grounds, capable of beautifully explaining what otherwise threaten to be intractable semantic challenges to normative realism, and compatible with a wide range of other commitments that a naturalistic normative realist might hold. Together, we take these attractions to constitute a nearly irresistible resume for a candidate theory of reference for the naturalistic normative realist.

References


Cian Dorr and John Hawthorne. Semantic Plasticity and Speech Reports. MS.

Janice Dowell. Millikan, Metasemantics, and Moral Twin Earth. MS.


Billy Dunaway. The Metaphysical Conception of Realism. MSb.


Reference: Magnetism as a Solution to the Moral Twin Earth Problem


Reference Magnetism as a Solution to the Moral Twin Earth Problem


J. Robert G. Williams. Lewis on Reference and Eligibility. MS.