Reference Magnetism as a Solution to the Moral Twin Earth Problem

Billy Dunaway
University of Missouri – St Louis

Tristram McPherson
Ohio State University

Abstract:

'Moral Twin Earth' thought experiments constitute a central semantic challenge to naturalistic normative realism. This paper first outlines a general framework for understanding the challenge, according to which (i) central normative terms are semantically stable in ways that contrast with many other paradigmatic descriptive terms, and (ii) realists should expect to have a *unified* meta-semantic theory that explains the difference in stability between the normative and descriptive terms in question. The most attractive way of meeting this challenge, we argue, appeals to the idea of reference magnetism. According to this influential idea, some properties are *reference magnets*, which (roughly) means that they are comparatively easy to refer to. We argue that (together with other plausible assumptions) reference magnetism can provide an attractive explanation of both the general phenomenon of varying semantic stability, and the distinctive semantic stability of normative terms. We illustrate this by showing that reference magnetism can smoothly vindicate plausible judgments about Moral Twin Earth cases. We conclude by offering an alternative gloss on our account, for those wary of the metaphysical commitments we propose. The alternative account adapts our proposal to provide a debunking explanation of the apparent semantic stability of normative terms.

Introduction

It is a striking fact that some important normative terms appear to refer to the same property *stably*, even given significant changes in speaker usage, and the environment in which speakers use them. For example, the prevailing pattern of application of the term 'ought' has shifted substantially over the last century, but this does not seem to signal a change in the

term's referent. This is in stark contrast with some paradigmatically descriptive terms. For example, suppose that our usage of 'red' were to consistently and without confusion shift to include a wider range of orangey shades. It is plausible that this difference in *usage* would entail that we would refer to a property distinct from redness. Alternatively, a community's *environment* might also be substantially different from ours, in ways that generate shifts in reference. For example, we can imagine a community of speakers who are disposed to use a word to refer to the watery stuff around them, much as we do with 'water.' If their environment did not include H_2O, their term might nonetheless refer to a different kind than our word 'water' does.

We can (at a first pass) describe this distinguishing feature of 'ought' by saying that it is more *semantically stable* than terms like 'water' and 'red'. The apparent contrast between the semantic stability of central normative terms and the comparative semantic instability of many paradigmatically descriptive terms presents a puzzle for the proponent of descriptivism about normative language.[1] We take this puzzle to be the unifying thread among the most influential arguments against descriptivist semantics for normative terms, including anti-descriptivist uses of G. E. Moore's open question argument (e.g. Gibbard (2003)), R. M. Hare's translation argument (1952), and Terence Horgan and Mark Timmons' Moral Twin Earth argument (1991; 1992a; 1992b; 1996; 2000; 2009). The most compelling general version of the challenge to the descriptivist as having two parts. The first is to provide a descriptivist-friendly meta-semantic account that explains the striking semantic stability of certain normative terms. The second is to do this by appealing to a unified meta-semantic theory that also explains the variance in semantic stability across different parts of our language. This not a trivial challenge: as we will show, some familiar descriptivist meta-semantic accounts fail one or both parts of this challenge.

This paper systematically addresses this two-part challenge. We begin in Section 1 by exploring what semantic stability amounts to, and why descriptivism about normative terms is not guaranteed to succeed in explaining the phenomenon. In Section 2 we introduce the metaethical view we aim to defend from the challenge: a version of normative realism. In Section 3 we present a package of independently-motivated metaphysical and meta-semantic views – including a version of reference magnetism – that we use to explain semantic stability. Because Horgan and Timmons' Moral Twin Earth argument is the most influential form of the stability challenge in the contemporary literature, we discuss this argument in Section 4. Horgan and Timmons have sometimes claimed that *no* naturalistic realist theory of reference could explain the Moral Twin Earth phenomenon. We argue that a meta-semantics that includes reference magnetism fits neatly with a natural diagnosis of the weakest element of Horgan and Timmons' case for this general conclusion. We illustrate this diagnosis by showing how the package of views we develop here accounts for the core thought experiments that motivate Moral Twin Earth arguments. In Section 5, we consider some objections to our metaethical use of reference magnetism. We conclude in Section 6 by offering an alternative gloss on our account, for those wary of the metaphysical commitments we propose. The alternative account

[1] See Section 1 for a more thorough discussion of semantic stability, and Section 2 for more detail on how we understanding descriptivism.

adapts our proposal into an explanation that debunks the apparent semantic stability of normative terms.

There are other important recent proposals for explaining the distinctive semantic stability of normative language in the literature. Matti Eklund (*forthcoming*) favors the idea that appeal to inferential role may be able to explain semantic stability. And Robbie Williams (Ms.) explores a view on which semantic stability follows from the interpretive requirement that speakers be as reasonable – in an appropriate sense – as possible. We do not directly address these important proposals in this paper. Our aim is instead to explore what we take to be the most elegant and general explanation of the phenomenon: reference magnetism. Portions of non-normative descriptive language exhibit a variety of degrees of semantic stability. As we shall show, reference magnetism plays a central role in a natural and plausible explanation of this variation. This gives the proponent of reference magnetism a prima facie methodological advantage, especially from a realist's perspective: it doesn't exploit distinctive aspects of normative language, but rather makes the semantic stability of 'ought' a straightforward consequence of features the normative shares with other important naturalistic domains. We will explore how far we can take this elegant explanation of the semantic stability of the normative.

1. Semantic stability

What is the phenomenon of semantic stability? We will understand it as a gradeable modal property of terms in a language: a term is *semantically stable* to the extent that its referent tends to remain unchanged over counterfactual variance in semantically significant properties of the term.[2] What these semantically significant features might be is a question for the meta-semantic theorist to investigate; there are potentially many such features. We will illustrate the phenomenon of semantic stability by considering two classes of properties that have a strong claim to either be – or be closely correlated with – semantically significant properties: usage and environment. These will allow us to highlight how terms might be semantically stable to different degrees.

Usage of a term encompasses the ways that speakers of a language use – or are disposed to use – the term in question. Plausibly, no term is *perfectly* stable over usage. To see this, choose any term you like: if throughout history, we had used (and been disposed to use) that term in exactly the ways that we in fact have used 'cat', then presumably that term would have referred to cats. However, stability over usage nonetheless seems to vary between terms. To return to our initial examples, 'water' appears to be more semantically stable over usage than 'red'. It is plausible that a small systematic and community-wide shift in usage, on which 'red' was applied to a wider range of slightly orangey shades, would entail a commensurate change in what 'red' referred to. By contrast, it is plausible that no such small change in the usage of 'water' would change the referent of that word.

[2] For relevant recent discussions of semantic stability, see Dorr and Hawthorne (2013), Schoenfield (2016) and Manley (Ms.).

But even 'water' is not stable in every respect, since there is another strong candidate for a semantically relevant property: the *environment* that a term is used in. The idea that environmental properties can affect the referent of that term, independently of facts about usage, is suggested by Hilary Putnam's famous Twin Earth thought experiment (Putnam 1975). This thought experiment suggests that were relevant parts of our environment to have contained XYZ instead of H_2O, the term 'water' would have referred to XYZ . This suggests that 'water' is relatively semantically *unstable* over relevant changes in the environment. By contrast, the term 'watery stuff' – which we are disposed to use to refer to any actual or possible clear potable (etc.) liquid – is arguably more environmentally stable: holding fixed facts about usage, it would refer to the same collection of kinds across many possible relevant changes to environment.

As we have mentioned, some central normative terms are paradigms of semantic stability with respect to both usage and environment. We will focus on just one example in this paper, an important precisification of 'ought.' It is plausible that the word 'ought' is highly context-sensitive, easily accommodating a variety of norms and other contextually supplied factors. For instance there are some contexts where the term clearly expresses an epistemic or end-relative notion. We will not focus on these precisifications.

One important precisification of 'ought' is deliberative. 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 the deliberative question: "What *ought* I to do?" The 'ought' being deployed in this deliberative context is not obviously linked to morality, but it does appears to be understood by the deliberator as invoking a distinctively deliberatively significant sort of normativity (see McPherson *forthcoming* for further discussion).[3] In what follows, we will treat this precisification as read into the meaning of 'ought', throughout.

The most straightforward way to motivate the idea that 'ought' is distinctively semantically stable is to consider counterfactual scenarios where we vary relevant usage and environmental properties for 'ought'. Consider our current usage of 'ought.' We often apply it to actions that produce the most good, but sometimes (and in some not obviously systematic ways) we refrain from applying 'ought' to optimific actions. For example, many people refrain from applying it to optimific actions that require dramatic coercion or manipulation.

Now imagine two possible counterfactual changes to this world. First, imagine that – holding fixed its deliberative role – our usage of 'ought' was slightly different. For example we might be much more consistent in applying 'ought' to only optimific actions, making fewer exceptions for coerced and manipulated acts. It is intuitively plausible that given these differences in use, we would still be *talking about the same thing* with our use of 'ought', rather than having shifted to using a semantically different homonym. Second, suppose that we held fixed our usage of

[3] Recent metaethics has been marked by a shift in focus from morality narrowly understood to this sort of deliberative or practical normativity. For example, see Enoch (2011), Gibbard (2003), Schroeder (2007, 2008), Smith (2013), Street (2008), and Wedgwood (2007).

'ought' and counterfactually varied the environment in which this usage occurs. For instance, suppose our actual usage occurs in an environment where uses of 'ought' are causally connected to a feedback mechanism that tends to make our usage approximately track the property of *being optimific-with-exceptions-for-autonomy-and-coercion*. Then the counterfactual world would be one where the causal connections are such that a different property is causally connected to uses of 'ought.' For example the property of *being optimific* simpliciter might play the relevant causal role in this world. It is again plausible that this change would not alter the referent of 'ought'.

Explaining what grounds the semantic properties of linguistic entities is one task of *meta-semantic* theories.[4] Because our explanandum is referential stability, we will focus on theories that explain what grounds the fact that a given term refers to one part of the world rather than another. This discussion of the apparent semantic stability of 'ought' allows us to refine the general challenge that we set out in the introduction. The refined challenge consists of three substantial constraints which a descriptivist meta-semantics for normative terms needs to meet.

First, use is not the only semantically significant feature: since communities can use 'ought' differently and still be talking about the same thing, there must be some additional feature of the relevant worlds which explains why this is so. Second, the additional feature is not simply causal regulation via a feedback mechanism: in a counterfactual environment where a different property causally regulates use of 'ought', we will not thereby be in a world where 'ought' refers to a different property. Instead we need a meta-semantic theory that explains why 'ought' is highly stable with respect to both usage *and* environmental change. Third, a meta-semantic theory for normative terms will be most attractive if it can explain the stability of 'ought' by appealing to plausible resources that are also capable of explaining the range of degrees of semantic stability displayed by non-normative terms as well. As a cursory survey of non-normative terms like 'red' and 'water' shows, not all descriptive terms will exhibit similar semantic behavior in counterfactual worlds which contain changes in use and environment. Together, these three constraints constitute the heart of the meta-semantic challenge to descriptivism about normative terms.

2. Naturalistic normative realism

In developing our response to this challenge, we will assume that some version of *naturalistic normative realism* about deliberative normativity (henceforth *naturalistic realism*, for brevity) is correct. This is a metaethical thesis. What does it mean?

As we characterize it, naturalistic normative realism is a family of views about the nature of deliberative normativity characterized by four commitments.[5] The first is *descriptivism*: this is the

[4] For illuminating discussion of the variety of metasemantic projects, see Burgess and Sherman (2014). For relevant discussion in a similar context to ours, see Williams (2015) and (Ms.).

[5] We 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.

claim that ordinary declarative sentences in which 'ought' features—sentences like 'Alice ought to intervene'— are to be understood as representing the world as being a certain way. This contrasts with views on which normative words mark distinctive speech-acts, or semantically express desire-like psychological states. The commitment to descriptivism as we understand it makes an explanation of semantic stability particularly challenging for the realist: since normative and non-normative terms alike have the same semantic role, there is a strong methodological incentive give a unified meta-semantic theory which convers both cases.

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. Naturalistic realism grants that the word 'ought' is context sensitive (as we explained in the previous section), but claims that once attention is appropriately focused on the deliberative context, there is no further context-sensitivity to be explained.

The third claim is that central normative properties – including the property picked out by 'ought' – are *metaphysically explanatory*.[6] (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.[7] 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.[8] The explanatory dimension of realism likewise rules out *quietist realism*, which rejects the quasi-realist's characteristic expressivism but echoes her denial that normative talk involves substantive metaphysical commitments about the normative.[9]

Finally, a realist view is *naturalistic* if it takes normative properties to be reducible to, or metaphysically continuous with, the sorts of properties discovered by the sciences. This is

---

[6] One 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 (Ms.).

[7] See 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.

[8] Compare Dreier (2004); see also Dunaway (2016).

[9] Scanlon is the leading example of a quietist in this sense who simultaneously claims to be a non-naturalist realist (see especially Scanlon (2014, Ch. 2)). For critical discussion see Enoch and McPherson (*forthcoming*). Compare also Dworkin (1996, 2011), Kramer (2009), and Skorupski (2010).

inconsistent with *Moorean non-naturalism*, according to which the normative is an additional, *sui generis* component of reality.[10]

3. Meta-semantics and reference magnetism

Because naturalistic normative realism is committed a descriptive semantics for normative terms, it faces the central meta-semantic challenge for the descriptivist introduced in Section 1: to explain the distinctive semantic stability of 'ought.' Our thesis in this paper is that the most elegant way to achieve this goal is to deploy a meta-semantic theory that includes reference magnetism. On a common gloss (to be elucidated below), *reference magnetism* is the thesis that properties can be ordered as more or less *easy to refer to*, independent of facts about language-users. In this section we explain the version of reference magnetism we endorse, why it is appealing to think that reference magnetism will be a part of any general theory of reference-determination, and how reference magnetism will apply to 'ought'. We begin by introducing our metaphysical assumptions about the type of properties that are referentially privileged.

3.1 Eliteness

Our account develops and modifies an influential account of metaphysics and meta-semantics that was introduced in David Lewis's (1983) "New Work for a Theory of Universals".[11] The first Lewisian thesis that we accept 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 gradable phenomenon: *being negatively charged* is more elite than *being acidic* which in turn is more elite than *being colored grue*. Crucially, if a property *P* is elite to a certain degree, then it is necessarily elite to that degree (cf. Lewis 1986, 61, n. 44). We will call this the *Necessity of Eliteness*. Because Necessity of Eliteness is crucial to our argument, and because some philosophers have entertained doubts about it (e.g. Cameron (2010, 284)), it is worth briefly explaining the case for it.

Eliteness is a plausible and attractive metaphysical posit in virtue of being a single property that plays a number of diverse explanatory roles (Lewis (1986, 61 n. 44); see Sider (2011) for an ambitious and systematic discussion and defense). Necessity of Eliteness is crucial to many of these roles; here we briefly offer two examples. First, the eliteness of instantiated properties is claimed to explain objective similarity. Roughly, this means that any two possible objects that share a highly elite property thereby objectively resemble each other to some degree. For

[10] 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 Dunaway (2015) and McPherson (2015).
[11] See also Lewis (1984, 1986). Contemporary discussions can be found in Hawthorne (2007) and Sider (2011).

example, if *being negatively charged* is perfectly elite, then any two possible negatively charged things resemble each other to some degree, in virtue of negative charge being perfectly elite. If Necessity of Eliteness were false, eliteness would not play this role: it would be possible that, while negative charge is (actually) perfectly elite, there are possible negatively charged things which do not resemble other negatively charged things. Second, the perfectly elite properties are claimed to form a complete supervenience base: this means that no two worlds that are identical in their distribution of perfectly elite properties can differ in any respect, period. A denial of Necessity of Eliteness is inconsistent with this. Such examples can easily be multiplied: Necessity of Eliteness is similarly crucial to the ability of eliteness to play other advertised roles, such as explaining lawhood or duplication. So our case for Necessity of Eliteness is that it is crucial to the ability of eliteness to play many of the roles that make it a promising theoretical notion (*cf.* Dorr and Hawthorne (2013, 31)).

As we emphasized in the previous section, a familiar commitment of naturalistic normative realism is that normative properties are metaphysically explanatory. The most elegant way to spell out this idea within the eliteness-centric approach to metaphysics is to conclude that normative properties are highly elite. This assumption will be crucial to the argument to come.

3.2 Reference magnetism

The second influential Lewisian thesis that we adopt is that elite properties are easy to refer to or 'reference magnets'. The point of the latter metaphor is to convey the idea that elite properties attract reference in a way analogous to the way that magnets attract iron. Just as a magnetic field can exert force on a piece or iron in proportion to its strength, so a property attracts reference in proportion to its eliteness. And just as magnetism is not the only force that acts on iron, so eliteness is not the only determinant of reference. 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.[12]

Let a *semantic theory* be an assignment of referents to terms in a language, as used by a linguistic community. The *Toy Theory* is a meta-semantic theory according to which candidate semantic theories are to be evaluated along two dimensions. The first is a *fit with use* dimension: roughly, we will say that a candidate semantic theory for a language achieves fit to the extent that it makes the sentences of the language accepted by members of the community, come out *true*. The second dimension is *reference magnetism*. A candidate semantic theory could assign a more or less elite referent to a given term. Candidate semantic theories do well on this dimension to the extent that they assign highly elite referents to terms in the language. Theories might do well on one of these dimensions, and poorly on the other. For example, some

[12] Like Lewis, we would ultimately prefer to implement reference magnetism within a broadly 'headfirst' approach, which assigned referents to mental contents first, and explained linguistic contents parasitically on mental content. We ignore this complication only to simplify discussion.

theories might do especially well on the fit dimension by assigning hopelessly non-elite referents to terms in a community's language. These theories thereby fare poorly on the eliteness dimension. According to the Toy Theory, the *correct* semantic theory is the candidate theory that maximizes the sum of fit with use and reference magnetism.

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. It is very easy to imagine a linguistic community much like ours, whose usage of 'cat' fits gappycats at least as well as ordinary cats. Here the reference magnetism dimension of the Toy Theory comes into play: this community still refers to cats, on that theory, because gappycathood is much less elite than cathood.[13]

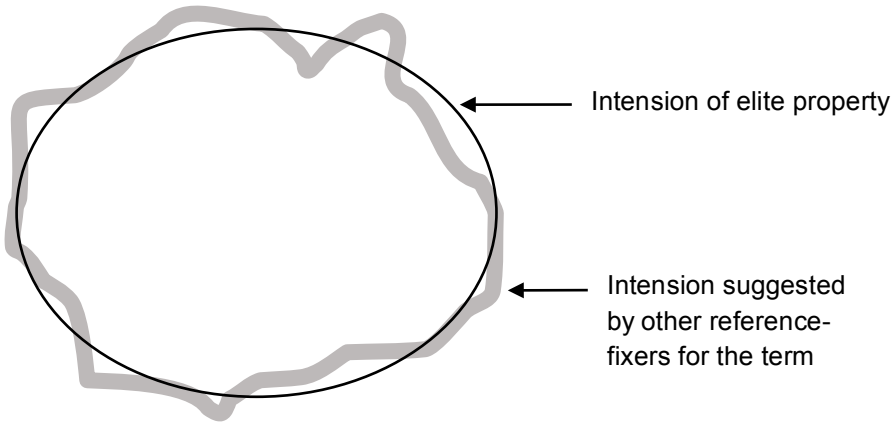
Second, suppose that a community of early humans discovered both gold (i.e. the element Au), and 'fool's gold' (i.e. iron pyrite), and were unable to tell the difference between the two substances. Suppose that they introduced the term 'goldite' to refer to the substance they were discovering. Consider 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'.

This example illustrates an important aspect of our theory, which is characteristic of orthodox discussions of reference magnetism: according to this theory, reference magnetism is grounded in an objective metaphysical feature – eliteness – and not by what (if anything) speakers believe about eliteness. Reference magnetism can thus operate even when competent speakers of a language are ignorant or misled concerning which properties or kinds are elite.

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 the relevant reference magnet for a term, and other reference-fixing factors (exemplified by fit with use, on the Toy Theory), given such a combined theory of reference:

[13] This conclusion assumes that all else is equal between the gappycat-assigning semantic theory and the cat-assigning theory. The Toy Theory is *holistic*, so if the cat-assigning theory were to make assignments elsewhere in the language that fit poorly with use, or if it were to assign referents elsewhere that were comparatively less elite, then the gappycat-assigning theory might turn out to be the correct theory, after all.

Figure 1: natural joints as reference-magnets



Here, we think of an *intension* as a function from possible worlds to sets of objects. A semantic theory assigns intensions to terms in a language (or entities which determine an intension, such as properties), settling the distribution of the referent of a term across possible worlds.

The natural kind in this figure is shown as a smooth oval, representing the eliteness of the joint, while the indeterminate and gerrymandered status of the intensions which are suggested by appealing solely to other reference-fixing factors, are represented by the thickness and irregularity of the represented boundary. In the figure, the smooth oval is not entirely enclosed within the imprecise boundary suggested by the other reference-fixing factors. This reflects the fact that reference magnetism does not merely resolve indeterminacy left by these other factors, but can override the contribution of use, as in the 'goldite' example. Reference magnetism suggests that the term represented in the image would refer to the natural joint, just as we suggested that 'goldite' would refer to Au.

This figure simplifies the role of eliteness in determining reference. Most importantly, it only illustrates reference magnetism for terms that refer to highly elite properties. On the Toy Theory, many terms in our language will refer to less elite properties, for two reasons. First, usage sometimes tends to track less elite properties: for example, there may not be any highly elite property in the vicinity of our usage of 'dinner.' Second, usage will encode *inferential connections* between terms across a language. For example, usage is not limited to how 'cat' and 'mammal' are applied in concrete situations: it also includes the fact that we routinely accept the inference 'if *x* is a cat then *x* is a mammal'. Because of such inferential connections, the Toy Theory will sometimes maximize the sum of fit and eliteness by assigning less elite referents to some terms.

We have focused here on providing a simplified introduction to how reference magnetism works. Proponents of reference magnetism have advertised its ability to provide principled and unified solutions to some of 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.[14] As we have suggested above, reference magnetism promises to do these things as part of a broader philosophical program—with eliteness at its center— which promises to explain objective similarity, lawhood, duplication, and many other phenomena. These credentials make reference magnetism a central part of one of the most powerful systematic approaches to contemporary philosophy. Reference magnetism is thus an appealing candidate ingredient for theorizing reference-determination.

We have emphasized that the normative realist should seek to explain the semantic stability of 'ought' by appeal to a completely general metasemantic theory. We are now in a position to offer an initial version of this explanation. This explanation has two parts: the first is a metaphysical thesis: that the actual referent of 'ought' – *obligatoriness* – is highly elite. As we have emphasized, this is a claim that the normative realist should be sympathetic to. The second part of the explanation is a completely general theory of reference determination that includes reference magnetism (such as the Toy Theory). Given these assumptions, obligatoriness is a reference magnet. Our Toy Theory of reference determination then entails some reason to think that 'ought' will be semantically stable. For example, just as with the 'goldite' example, 'ought' will refer to the elite property of obligatoriness given a range of plausible variation in use.

3.3 Unique eliteness

This provisional explanation of the semantic stability of 'ought' depends upon a further crucial assumption that we will call *Uniqueness*: that among the properties that fit moderately well with the use of 'ought' as a normative term, there is a single property that is much more elite than any of the others.[15]

To see why this assumption is crucial to our explanation of semantic stability, imagine that Uniqueness failed to be true. For example, imagine that there was a highly elite property that perfectly fit with a certain consequentialist theory, and a distinct, equally elite property that fit perfectly with a distinct deontological property. Given these assumptions, the Toy Theory would predict that (other things being equal) the referent of 'ought' would depend on which of these properties our usage of 'ought' best fit with, thereby exemplifying instability.

It is important to emphasize that not all broadly normative words are semantically stable. For example, consider the counterfactual scenario in which our dispositions and conventions around chess had evolved differently, such that it was universally accepted that "Knights" can move diagonally. It would be absurd to think that, due to reference magnetism, 'legal chess move' has a stable referent across this counterfactual change. Our explanation of this asymmetry between

[14] See Quine (1960), Putnam (1981), and Kripke (1982).

[15] Uniqueness, as stated, may need to be qualified in order to accommodate certain plausible views about ethical vagueness. There may be cases where it is vague what one ought to do: for example, it might be vague whether, given her circumstances, Janet ought to give a penny more to charity above what she has already given. This needs to be addressed, but we are confident that this can be done, and we won't dive into the possible routes for the reference magnetist here. See Schoenfield (2016) for more on vagueness and ethical realism, though we suspect there are more options for the realist to take than Schoenfield acknowledges.

'ought' and 'legal chess move' is that the analogue of Uniqueness is false for the rules of chess-like games: no such set of rules constitutes a uniquely comparatively elite real pattern in the world.

This explanation of the asymmetry between 'ought' and 'legal chess move' is not some ad hoc patch on behalf of the normative realist. Recall from Section 2 that one of the core metaphysical commitments of naturalistic realism is that the property of obligatoriness is metaphysically explanatory. And a central tenet of the elite properties approach to metaphysics is that only highly elite properties are metaphysically explanatory. We take this commitment to generalize to comparative explanatoriness: crucially, we think that few naturalistic normative realists would be comfortable with the thought that deontology provides the correct account of obligatoriness, but that the consequentialist property is also highly metaphysically explanatory.

These points about Uniqueness can be developed into a crucial element of our *general* explanation of variation in semantic stability. The general idea is that the analogue of Uniqueness is plausible for the central theoretical terms of successful sciences. For example, it is plausible that the referent of 'gene' is significantly more elite than the referent of any nearby gerrymandered term. Given this assumption, the Toy Theory will suggest that the referent of 'gene' is stable over changes in usage. Conversely, consider the continuum of shades from reddish orange to orangey red. Plausibly, no shade in that series is significantly more elite than the others. On the Toy Theory, this explains why we expect color terms like 'red' to exhibit much less stability over usage than 'ought' or 'gene'.

3.4 Lewisian canonical definitions

Our appeal to Uniqueness entails that we must reject David Lewis's account of relative eliteness. This account has two crucial elements. First, all and only the fundamental microphysical properties are perfectly elite. 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.

Lewis's account would render our explanation of semantic stability hopeless, exactly because it would undercut Uniqueness. For example—returning to our example—a typical deontological property surely does not have a markedly shorter canonical definition in terms of 'negative charge', 'spin', etc. than a typical consequentialist property: both definitions will be highly complex (perhaps even infinitely long) when stated in the canonical language. The same point holds for most reasonable hypotheses about what 'ought' picks out. Applying Lewis's view of

relative eliteness to the normative thus deprives the naturalistic realist of all the benefits of reference magnetism from the previous section.

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 pointed out, 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 arise 'chair' or 'rabbit', thereby predicting intolerable referential indeterminacy or instability. Lewis's theory of relative eliteness is thus incompatible with his own claims about magnetism's potential to solve central problems of reference-determination (Hawthorne (2006, 206), (2007, 434); compare also Williams (2007)).

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 a positive suggestion made by Hawthorne: in order to elaborate a workable theory of reference magnetism, we need to reject Lewis's definitional approach to degrees of eliteness, and to permit relative eliteness to float free from microphysical definability.[16]

3.5 An epistemology for primitive degrees of eliteness

We take this lesson to heart: degrees of eliteness are not definable by—or otherwise reducible to—the perfectly elite. However, abandoning Lewis's theory in favor of primitive degrees of eliteness may seem to make the appeal to reference magnetism objectionably unconstrained. Absent a principled account of how we come to know which properties are highly elite, reference magnetism may seem to amount to a way of redescribing the facts about reference which need to be explained.[17] We address this worry by sketching an account of how we can know facts about relative eliteness. We argue that this account is a principled 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:

[16] There are other important alternatives to this approach, which we do discuss at length (for reasons of space). One could, for instance, dispense with the need for degrees of eliteness altogether, and instead propose that many properties not countenanced by fundamental physics are perfectly elite. See Schaffer (2004) for a version of this idea. Alternatively, one could propose that there are two distinct notions of eliteness: one characteristic of the fundamental physical properties of mass, charge, and the like, and the other applicable to macroscopic, non-fundamental properties. See Weatherson (2013) for a version of this approach.

[17] We thank Laura Schroeter for powerfully pressing a variant of this worry.

that the discipline of *fundamental physics* is our crucial epistemic guide to which properties are perfectly metaphysically elite.[18]

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 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] The relative eliteness of a property is not explained by its relationship to the perfectly fundamental. We couple this metaphysical thesis with the obvious naturalistic amendment to Lewis's dubious physics-centric 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.[20] The broad thought behind Liberal extends Lewis's epistemology: metaphysical naturalists should adopt a realist philosophy of science, on which the ontological commitments of natural science provide (at least the heart of) our best guide to elite reality. We should be clear that this gloss unashamedly offloads two substantive questions onto the realist philosopher of science.

First, there is the question of how to identify the naturalistically credible disciplines. This is a version of the familiar demarcation problem in the philosophy of science (for a useful introduction see Hansson (2015)), though in fact we are dealing with a more general version of

---

[18] Lewis (1984, 224)

[19] Hawthorne (2006, 206) calls this view 'emergentism'. See also Hawthorne (2007).

[20] There is a superficially related use of reference magnetism in this context in van Roojen (2006) that should be compared to Liberal. Van Roojen makes use of the notion of 'discipline-relative' naturalness (or eliteness), which, like our use of Liberal, gives special place to a wide range of legitimate naturalistic areas of inquiry (and not just physics). But while Liberal posits a single scale of relative eliteness, van Roojen posits many different eliteness-like properties, For van Roojen, there is elite-relative-to-physics, elite-relative-to-biology, elite-relative-to-psychology, and so on. A property such as *gene* or *obligation* might have one of these eliteness-like properties while lacking others. Moreover the various notions of discipline-relative eliteness are not primitive: presumably there is something about the practice of biology that makes *gene* elite relative to biology but not physics. We have some worries about the meta-semantic role of discipline-relative eliteness (for example: how to handle terms deployed across multiple disciplines?), but our main point here is simply that Liberal is quite different from van Roojen's idea.
.

the question. We wish to treat mathematics, logic, and (perhaps) ethics as relevantly categorized alongside genuine natural sciences. These disciplines should count, according to Liberal, as theoretical enterprises that countenance very elite properties, just as physics and chemistry do (and, as astrology and alchemy do not). We don't have anything illuminating to say about this question here. But we think the distinction is an intuitive one, and are optimistic that an informative characterization is in the offing.

Second, there is the question of how to identify the ontological commitments warranted by the work in a scientific discipline. Here is a very schematic approach to this question that is congenial to our project. To begin, take the "law-like" generalizations of a scientific theory, and form a list of the properties which are referred to by the generalizations, or which must exist if the generalizations are true. These are the properties that are *countenanced* by the theory. According to Liberal, the fact that a property is countenanced by a scientific theory will give us reason to think that it is elite to some degree. A natural way of extending Liberal can give us guidance about relative eliteness as well. First, some of the properties countenanced by a certain theory have a wider explanatory role, or more basic explanatory status, than other properties countenanced by that theory. Other things being equal, this provides reason to take the former properties to be more elite than the latter. Second, there will be distinctions *between* theories: some deal with a more widespread and foundational aspects of reality (e.g., chemistry) while others are concerned with higher-level, less general phenomena (e.g., ecology). A natural extension of Liberal will have it that the former properties are more elite than the latter, other things being equal. (For related ideas, see development of the idea of a *wide cosmological role* in Wright (1992, 196-7).)

One illustrative case for the relationship between theoretical structure and eliteness comes in the form of *reductionist* theories. Claims of reduction are related to the eliteness-facts in several ways (for general discussion, see McPherson 2015). One central aspect of the relationship for our purposes is the following: featuring in non-trivial reductions of a range other properties is a significant mark of eliteness. (One way to think about this is to think of reduction as a kind of explanatory notion, and hence properties that appear in multiple reductions will be highly explanatory, and have something like Wright's "wide cosmological role".) The presence of a reduction in ethics shows that certain psychological (or other lower-level) properties have an additional explanatory feature beyond their lives qua psychological properties: they also explain facts about normativity, action-guidingness, and the like.

This leads to a second point about reductionist theories. Often in ethics such theories aim to reduce the normative to psychological properties that are, on the surface, fairly natural and non-gerrymandered. For example: Peter Railton (1986) looks to facts about what an agent who has undergone a certain kind of feedback-based process of desire alternation would want, and Mark Schroeder (2007) appeals to the property of promoting some desire. Both of these properties appear to be, in view of their role in psychological theory, fairly elite. It is crucial for our purposes that, supposing that one of these reductions is correct, the psychological property that features in the correct reductive theory will (other things being equal) thereby be much more elite than the other. Merely from the fact that these properties have similar lives in psychological

explanations, it does not follow that they are similarly elite, full stop. If this did follow, it would be very easy to cast doubt on the semantic stability of 'ought:' just find a community whose use of this term fit with some other very elite psychological property. Liberal, as we are developing it, does not ensure that this scenario is a metaphysical possibility.[21]

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 simplicity of canonical definition, it allows that 'rabbit' – while not a term from fundamental physics – 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 serve 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 (1988, §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, in the manner of Schroeder and Railton. The essential point is that any adequately naturalistic realism already needs a naturalistically acceptable epistemology for the normative. And our account is compatible with the central plausible proposals in this arena.

In this section, we have introduced several central assumptions of our account of reference magnetism, including the elite properties framework, Necessity of Eliteness, Uniqueness, and Liberal. These are evidently non-trivial assumptions. We have sought to explain why these assumptions should seem independently plausible to naturalistically-minded philosophers. As we emphasized, Uniqueness is a plausible commitment for the naturalistic normative realist, and the remaining assumptions are needed to explain how eliteness and reference magnetism can deliver on their general explanatory promise.[22] Together, these assumptions allow the

[21] This isn't to say that the psychological character of a property is irrelevant to its explanatory profile and hence for our evidence of its eliteness. For example: Jackson (1998) gives an argument which leaves it open that, at the psychological level of description (or any other naturalistic level), normative properties can only be picked out by infinitely long disjunctions. We are sympathetic to the thought that being massively disjunctive in this way is strong evidence against eliteness. One natural way to explain this though is the following: even if massively disjunctive properties can serve as reduction bases, they tend not to have a wide explanatory role. On our account, this is evidence that the eliteness of such properties is at best limited. For more discussion of this issue, see Dunaway (2015).)

[22] There is more work to be done in developing even the initial picture we are using in this paper for illustration. For instance, the Toy Theory of reference-determination that we sketched uses a degreed notion of relative eliteness that can be measured with real numbers. Liberal, by contrast, at best only gives an ordinal ranking of the eliteness of

normative realist to provide a strong initial answer to the two-part challenge that frames this paper. First, with these assumptions in hand, even the crude Toy Theory is able to explain the striking semantic stability of 'ought'. Second, it does so in a way that also provides a unified explanation of why some terms (like 'gold' or 'gene') are more semantically stable than others (like 'red'). The explanatory power of our account can be further dramatized by showing how it addresses one of the most important meta-semantic challenges to normative realism. We now turn to that task.

4. The Twin Earth challenge to naturalistic normative realism

In a series of papers, Terence Horgan and Mark Timmons have used 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 circumstances in which speakers can substantively disagree with each other about moral matters (1991; 1992a; 1992b; 1996; 2000; 2009). 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).

Others have already argued that reference magnetism can play a role in addressing Moral Twin Earth thought experiments (van Roojen 2006, Edwards 2013). We aim to accomplish more. In this section, we explain Horgan and Timmons' confidence that their challenge generalizes to any naturalistic normative realist. We then show that reference magnetism is more than an independently-motivated counterexample to this general challenge. Rather, it allows us to identify an error in one of the core assumptions Horgan and Timmons have made in generalizing their argument. Finally, we show that reference magnetism can explain the difference in degree of stability between 'ought' and other non-normative descriptive terms. For example, it not only explains why 'ought' is stable across a range of environments *and* usage profiles; it also explains why 'water' is stable with respect to use, but unlike 'ought' is not stable with respect to environment. We take the power of this response to further illustrate the interest of reference magnetism for any normative realist concerned about the semantic stability of normative terms.

4.1 The Normative Twin Earth challenge

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[23] *N* for 'ought', the question:

---

various properties. Whether we should revise the meta-semantics to work with an ordinal eliteness-ranking of properties, revise the Liberal epistemology to explain knowledge of real-valued degrees of eliteness, or admit some amount of skepticism about the real-valued degrees of eliteness, is a question we leave for another time.

[23] 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

I know that doing such-and-such is *N*, but ought I to do it?

could be asked sensibly, without displaying conceptual confusion or lack of semantic competence. Put differently: all naturalistic analyses of 'ought' will have an *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 H_2O. For such a speaker, the question:

I know that the stuff in this glass is H_2O, 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 (1988) 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 moral terms without knowing which kind they refer to, under a naturalistic description.

In keeping with our focus on normative rather than narrowly moral language, we can imagine adapting Boyd's proposal to 'ought'. The adapted account provides an explanation for why 'ought'-analyses (like 'water'-analyses) have an open feel. Notice, however, that such accounts also have predictable implications for semantic stability: a causal regulation account of reference-determination will tend to secure some amount of *usage stability* but not *environmental stability*. This upshot is plausible for some terms. Consider, for example, 'water.' We want to allow that a modern-day Thales, in the grip of confused metaphysical views, could believe that everything is made of water, and convince his linguistic community of this fact. That is some impressive usage stability for 'water', and Boyd's causal regulation account of reference can potentially vindicate it. On the other hand, Hilary Putnam's Twin Earth thought experiment (1975, 222-4) suggests that 'water' does *not* exhibit environmental semantic stability.

---

between the *analysans* and *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.

Briefly, that thought experiment is as follows. Imagine a Twin Earth that is identical to Earth except that the stuff in Twin Earth lakes and streams and sinks and bodies is not H_2O. It is rather a complex chemical (dubbed 'XYZ' by Putnam) with very similar macro-qualities to H_2O, but a radically different microstructure. On Earth, we typically apply the English term 'water' to samples of H_2O. 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 H_2O with our term 'water,' and our counterparts refer to XYZ with theirs.

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 (*á la* Boyd) a helpful model for the meta-semantics of normative terms, then normative terms should mimic the distinctive semantic features of natural kind terms as well. Horgan and Timmons's Moral Twin Earth thought experiments arguably show that they do not: in our terminology, central normative terms, unlike 'water' appear to exhibit environmental semantic stability.[24]

To show this, Horgan and Timmons ask us to imagine a Moral Twin Earth scenario tailored to Boyd's theory: imagine that the difference between Earth and Moral Twin Earth is that the property[25] causally regulating the use of the word 'good' on Earth is a consequentialist property (e.g., the maximization of happiness). By contrast, the property regulating the use of the word 'good' on Moral 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 'good' 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 'good' on the two planets, Boyd's theory entails that these words do not corefer. Just as use of '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, use of 'good' is likewise regulated by different properties on Earth and Moral Twin Earth. So on Boyd's theory, 'good' refers to different properties in the mouths of Earthlings and Moral Twin Earthlings.

---

[24] As an interpretive matter, we do not wish to commit to a reading of Horgan and Timmons on which they are granting that a Boyd-style theory will adequately explain what we are calling *usage stability* but fail (owing to an analogy with 'water') to explain *environmental* stability. They do not make this distinction, and their presentation of the Moral Twin Earth thought experiment might be read as suggesting that it is a difference in usage between the two communities that (according to Boyd's theory) implies that the relevant communities are talking about different things. What we are focusing on here is (i) the fact that the Moral Twin Earth community seems to be talking about the same thing as us, regardless of how their differences with us are to be categorized, and (ii) 'ought' represents a striking departure from 'water' in this respect, because in the latter case environmental differences *would* produce the talking-past-us phenomenon.

[25] Horgan 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.

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 H_2O. 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 Boyd's theory—the speakers on Earth and Moral Twin Earth don't refer to the same property, they do not substantially disagree in their utterances of superficially inconsistent sentences containing 'good'. Their disagreement is merely verbal, just like Alice's disagreement with her twin. But this consequence of Boyd's theory is implausible: when you claim that an act is 'good' while your twin steadfastly refuses to apply their functionally similar term 'good' to the same 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 Moral Twin Earth case, it is plausible that we can have real moral disagreements with our Twins, using 'good'. And for the realist this requires that our words 'good' corefer. But the case has been constructed to ensure that Boyd's theory entails that our word 'good' does not corefer with the Twins' word 'good'. So if we assume Boyd's theory of reference, moral realism appears implausible.

Notice two things here. First, it takes no great creativity to see that a Boyd-style semantics applied to the deliberative 'ought' will be at least as vulnerable to this sort of argument as his original account of the moral 'good.' The resulting argument would be a *Normative* Twin Earth challenge.[26] Second, nothing in Horgan and Timmons's argument casts doubt on the adequacy of the causal regulation account as an account of the semantic for natural kind terms. This in turn suggests two deeper worries for the normative realist. First, it seemingly shows that an important theory of the semantics of natural kind terms cannot be adapted to provide an adequate semantics for normative terms. Therefore, if the normative realist wishes to retain a

[26] The assumptions needed to get the Moral Twin Earth argument going are arguably more plausible in the context of 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 this assumption, 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, 3)), 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.

causal regulation theory for natural kind terms, she will need to adopt an unappealingly disjunctive semantics to accommodate her realist commitments about the normative. Second, because both normative terms and natural kinds terms exhibit an open feel, the semantics of natural kind terms appears to be especially promising models for the normative realist seeking a semantics for normative terms. However, Horgan and Timmons have shown that some theories that provide a promising explanation of the open feel for natural kind terms do so by appealing to commitments that will be implausible for a semantics for normative terms, casting doubt on how helpful this model is to the normative realist.

4.2 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 – and adapted to 'ought' – 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 (allegedly) 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:

> [I]t 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 earthlings bear the same kinds of relations to distinct natural properties (e.g., deontological functional properties). (2000, 146; see also their 1992b, 167.)

We conjecture that Horgan and Timmons are led to this ambitious conclusion by the following reasoning. 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' in our language rigidly refers to H_2O – that is, in our mouths it picks out the stuff in any counterfactual world that is H_2O. However, it is a contingent fact that the stuff in our environment that 'plays the water-role' is H_2O. Call this *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 can (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_2O 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.[27]

Contingency in reference-determination, however, provides exactly the materials needed to construct a Putnam-style Twin Earth case. One simply finds two communities who use a term so that it (rigidly) refers to whatever property plays the such-and-such role, but where (owing to the contingency in reference-determination) in the environment of the two communities, the property that plays the such-and-such role differs.

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. It will, plausibly, be only a contingent fact that a particular property connected to these roles. There will then, for reasons sketched above, be a Twin Earth case for ‘ought’. And this seemingly puts Horgan and Timmons in a position to construct a *reductio* of any possible naturalistic theory of reference-determination for ‘ought’. More explicitly, the *reductio* arises from the fact that 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’.

We saw the motivations for Disagreement and Coreference in Section 4.1. 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’. 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.

4.3 Methodological interlude

The preceding discussion puts us in a position to identify four broad theoretical goals for the normative realist, in light of the generalized Normative Twin Earth challenge as we have just

[27] We are indebted to Mark Schroeder for suggesting something like this hypothesis to us.

reconstructed it. First, the realist should aim to provide a *diagnosis* of where the generalized challenge goes wrong. Second, that diagnosis should be coupled with an *informative* account of reference-determination, which dovetails with the diagnosis provided. Third, this account of reference-determination should be *independently motivated*, rather than being posited ad hoc, simply to solve this problem. And finally, the account should explain (or explain away) the apparent *asymmetry* between Normative Twin Earth cases and Putnam's original twin earth cases. In the remainder of this section we argue that an approach to reference-determination that appeals to reference magnetism can meet all of these goals.

There are a number of potential ways for the normative realist to reply to the generalized challenge we have reconstructed. One might try to refute Disagreement by example, by showing that given a particular plausible reference-determination relation, any communities that refer to different properties will also intuitively disagree only verbally (Merli (2002, §III)). Alternatively, one might argue that some naturalistic theories of reference-determination are in a position to dismiss the prima facie plausibility of Disagreement on principled grounds (Dowell (2016)). 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. But 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.

Contingency is motivated by an assumption: that any fact that suffices to explain the open feel of an analysis of 'ought' will also be metaphysically contingent. But this assumption should strike us as suspicious: the open feel is a *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. As we now explain, reference magnetism fits this recipe perfectly.

4.4 Reference magnetism and Normative Twin Earth

Recall that Contingency is the assumption that only contingency in the reference-determining mechanisms can explain the open feel of putative normative and natural kind analyses. Any theory of reference-determination that includes reference magnetism constitutes a potential counterexample to Contingency. This is due to two theses that we have discussed above. The first is the Necessity of Eliteness: if a property is elite, it is necessarily elite. The second (emphasized in the example of 'goldite') is that knowledge of which properties are elite is no part of counting as a competent speaker. Together, these theses entail that the open feel of normative analyses can be explained by the possibility of ignorance, on the part of competent speakers, of the distribution of the elite normative properties. This ignorance doesn't require contingent reference-determination: given Necessity of Eliteness, *every* possible community that uses normative terms similarly might, owing to magnetism, refer to the same property.

This abstract point is best illustrated concretely. We do this by showing how the Toy Theory of reference we used to explain semantic stability phenomena allows the normative realist to address cases modeled on Horgan and Timmons' Moral Twin Earth cases. To begin, consider a Normative Twin Earth case constructed to cast doubt on a simple pure 'use' theory of reference-determination for 'ought'. It would begin by assuming (for determinacy) that the property that best fits Earthling use of 'ought' is the deontological property. It would then consider a Twin Earth where the Twins have a word 'ought*' that has the very same role in their deliberation, self-monitoring, and inter-personal criticism that 'ought' has in ours. The Twins' use of 'ought*' is otherwise different only to the extent required to make the consequentialist property best fit their use. A pure use theory predicts, given these stipulations, that 'ought' and 'ought*' do not corefer. But it is plausible that we and the Twins can express real disagreements via our uses of 'ought' and 'ought*' here. So if Coreference is correct, this Normative Twin Earth case is a serious problem for the pure use theory.

The Toy Theory of reference-determination introduced above combines use with reference magnetism, and thereby suggests a very different account of the Normative Twin Earth case just sketched. It is plausible that the deontological property and the consequentialist property have quite similar intensions. The Toy Theory of reference allows in cases like these that the referent of 'ought' may be a worse fit with use than another candidate property, provided that it is significantly more elite. (This is the characteristic significance of reference magnetism, illustrated by the 'goldite' example above.) So 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. Then the Toy Theory will entail that the Twins' word 'ought*' refers to the deontological property, and hence corefers with our word 'ought'.

We can relax the (purely illustrative) assumption that that the deontological property is comparatively elite, and replace it with the assumption that there is some single highly comparatively elite property in this vicinity. Here, thanks to reference magnetism, the Toy Theory of reference suggests that we and our twins refer to the same property with our use of the words 'ought' and 'ought*' respectively. In other words, reference magnetism can vindicate the core semantic judgment that is the heart of the Normative Twin Earth challenge.

It might seem that this illustration cheats: after all, it showed that reference magnetism can address a Normative Twin Earth case designed to refute another theory. But importantly we *cannot* construct a Normative Twin Earth case against a theory of reference-determination that includes a magnetism dimension, according to the usual recipe. Applying that recipe would require that (e.g.) we postulate that the deontological property is highly elite on earth, while it is not on the Twin Earth, where the consequentialist property is instead highly elite. Necessity of Eliteness, however, rules out this possibility: if the deontological property is highly elite on earth, it is just as elite in every possible world. The usual strategy of finding a 'Twin Earth' world which varies in the crucial reference-determining respects simply cannot be carried out.

4.5 Intensional similarity

This illustration is extremely quick. Importantly, the success of our case hangs on two assumptions that were mentioned only briefly in the examples. The first is the Uniqueness assumption explained in Section 3.3: that there is some unique property in the 'intensional vicinity' that is much more elite than other properties that fit reasonably with use. The second assumption is that the relevant candidate properties (e.g. the deontological and consequentialist properties) have quite similar intensions. Here we focus on the second assumption, and the virtues of our approach in light of this assumption.

In our illustration, we emphasized that the deontological and consequentialist properties had very similar intensions. There are two significant aspects to this similarity. First, it is generally true that in every possible circumstance in which an agent has options, there will be some (perhaps disjunctive) option that is required by the deontological principle, and also some (perhaps disjunctive) option that is required by the consequentialist principle. This has an important implication: with only some very bizarre exceptions, there are no worlds in which the consequentialist property is instantiated, while the deontological property is not. Second, in a very wide range of circumstances, plausible deontological and consequentialist principles call for exactly the same actions. So, in general, the extensions of the two properties that are candidate referents will, at a world, be relatively similar. Call this *Intensional Similarity*.

Horgan and Timmons notice this contrast, suggesting that it *helps* their case against the moral naturalist:

> 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. (2000, 145)

According to Horgan and Timmons, Intensional Similarity makes things worse for the naturalist, by exacerbating the threat of referential indeterminacy. With multiple candidate referents in the vicinity of a community's use—which is guaranteed by the intensional similarity of candidate referents—the naturalist (Horgan and Timmons claim) will not be able to explain why a community determinately refers to one of the candidate referents over the others. As we have seen, reference magnetism is tailor-made to meet this challenge. If one candidate referent is much more elite than the others, then a theory of reference-determination that includes reference magnetism will hold that the highly elite candidate is the determinate referent.

The assumption of Intentional Similarity does more than explain why meta-semantic theories that include reference magnetism are immune to standard Normative Twin Earth-style counterexamples, however. It also shows how a theory that includes reference magnetism can explain *contrasts* in 'Twin Earthability', which in our terms is evidence for environmental semantic instability. For example, 'water' is paradigmatically Twin Earthable (that is, it displays a kind of environmental *in*stability) while we have just argued that 'ought' is not. On our view, the

explanation is that while Intensional Similarity is plausible for candidate referents for ‘ought,’ it is not plausible for candidate referents for ‘water.’

Consider: when we apply the Toy Theory of reference to Putnam’s case, the Earth English word ‘water’ picks out H_2O, a highly elite kind that well-fits our use. On Twin Earth, however, the lakes and streams and showers are not filled with H_2O, but rather a different kind: XYZ. Unlike the case of the deontological and consequentialist properties discussed above, it is not plausible that H_2O and XYZ are intensionally similar: for example, a possible world can easily contain H_2O and not XYZ (or vice-versa), and the same is true of specific environments within a possible world.

Because of the failure of intensional similarity for H_2O and XYZ, we can assume that the relevant Twin Earth environments lack H_2O. In light of this, the Toy Theory of reference-determination *cannot* plausibly entail that H_2O is the referent of the Twins’ word ‘water’. The contrast in the applicability of Intensional Similarity permits the normative realist to deploy the Toy Theory (or another meta-semantic theory that includes reference-magnetism) to achieve the final theoretical goal that we mentioned in Section 4.3: to vindicate the apparent *contrast* between Normative Twin Earth and Putnam’s original Twin Earth case.

This means that a meta-semantic theory that includes reference magnetism will view the difference in environmental stability between ‘water’ and ‘ought’ as a consequence of the independent fact that Intensional Similarity is true of the latter but not the former. ‘Ought’ seems more semantically stable over environmental variation than ‘water’ does. But this is just a consequence of the fact that the magnetic referent for ‘ought’ is present in *every* relevant environment, while ‘water’ can be used in environments where H_2O is nowhere to be found. This means that the explanation in difference environmental stability is not a symptom of a disunified meta-semantic theory for the realist; rather reference magnetism is a unified meta-semantic explanation of both terms, which yields different semantic profiles by virtue of variation in Intensional Similarity.

In this section, we have applied reference magnetism to the Normative Twin Earth challenge. We have sought to show that reference magnetism can achieve four weighty theoretical goals for the normative realist. First, reference magnetism provides a diagnosis of where Horgan and Timmons’s generalized challenge to naturalistic normative realism goes wrong, by providing a clear counterexample to Contingency. Second, reference magnetism is a generally applicable (if partial) theory of reference-determination, that is part of a well-motivated approach to many of the central questions in metaphysics and semantics. In light of this, it is no ad hoc posit by the normative realist. Third, reference magnetism is informative, providing an explanation of why we and our Normative Twins corefer in our uses of ‘ought’ and ‘ought*’. And finally, reference magnetism is a unified theory that is capable of explaining the seemingly contrasting semantic appearances concerning Normative Twin Earth and Putnam’s original Twin Earth case. We take these to be highly impressive results that warrant close attention by normative realists.

5. Three objections to the reference magnetic solution to Normative Twin Earth

We have just advertised the credentials of our account in addressing the version of the core challenge exemplified by Horgan and Timmons' work. In this section we consider three objections which grant our account success against some of the specific cases raised by Horgan and Timmons, but claim that our view nonetheless faces quite general problems.

5.1 Normatively Inverted Twin Earth

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.[28]

It is worth emphasizing several preliminary points about this sort of case. First, our primary aim in this paper is to explain how reference magnetism can explain the striking semantic stability of certain normative terms. In doing so, we have offered a *toy* meta-semantic theory includes reference magnetism. We do *not* claim that the Toy Theory is correct, or generally adequate. So it may be that objections like this one invite us to develop a more serious meta-semantic theory that includes reference magnetism. Second, this case is dialectically weaker than Horgan and Timmons's cases in two ways. On the one hand, it is quite difficult to concretely imagine such a community, so we should be cautious about placing much weight on cases like this. On the other hand, it is not at all obvious that the Inverts genuinely 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 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 optimism here. First, recall the realist's commitment that the especially elite candidate normative properties are very sparsely distributed. In light of

[28] This example is a more extreme version of Hare's 'missionaries and cannibals' example (1952, §9.4), which is arguably the ur-text of Normative Twin Earth.

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 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 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 meta-semantic 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.

5.2 Alien elite properties

A second objection: there are worlds which contain "alien" fundamental properties—i.e., properties that are fundamental (in those worlds) but are not instantiated in our world. Some alien fundamental properties might have a distribution in a community's world that maps closely on to that community's use of 'ought'. Won't the alien property count as a second highly elite property in the vicinity of the community's usage, and hence a counterexample to the Uniqueness thesis? Even more worrisome, the Toy Theory will predict that the alien property, owing to its superior eliteness, is the property the community's use of 'ought' refers to.

This objection can be met with the same style of reply as Normative Inverted Earth. Alien fundamental properties will not fit well with the *full range* of a community's usage of 'ought' at all. This is because the full range of such uses will include modal and counterfactual statements, such as 'it is not possible that one performs an action simply because it causes pain, with no other effects, and thereby does something that ought to be done.' Or, speakers in the alien world will say of our world, *w*, things like 'even if one were in *w*, it would be the case that one ought not to torture others for fun'. Arguably, these modalized principles should count more for determining fit than one-off applications of 'ought' to particular actions. So on balance the imagined alien fundamental properties will be a poor fit for the full range of use of 'ought' by the alien speakers. In light of this, the possibility of alien worlds does not, after all, constitute a promising counterexample to Uniqueness.[29]

[29] We thank an anonymous referee for raising the worry posed by alien fundamental properties here.

5.3 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.[30]

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.[31]

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.

We take there to be at least two independent plausible ways of addressing this objection. First, 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 matter, and we see no reason that normative realists should not be robust in this way.

Suppose, however, that we grant that procedurally ideal normative theorizing is infallible. Indeed, suppose we grant the bolder thesis that facts about procedurally ideal theorizing ground facts about what one ought to do. If ideal theorizing grounds practical obligation, then it is very natural for the reference magnetist to think that ideality is a highly elite property. Tokens of 'ideal' in communities who use the term differently will still refer to the same property, owing to the now familiar operation of magnetism. And so (assuming there is no semantic shift in the referents of 'procedure' and 'theorizing' between the communities) the compound term 'procedurally ideal theorizing' will refer to the same property in both communities. Via the inferential connection between 'ought' and 'procedurally ideal theorizing' we spotted the objector

[30] Thanks to Pekka Väyrynen for raising a version of this objection.

[31] Sturgeon (2002, §III) defends a relevant combination of methodological naturalism and methodological autonomy.

at the outset, these communities will also refer to the same property with ‘ought’. The reference magnetist can thus get the same result as before, even when she grants a basic explanatory role to normative theorizing.

6. Are the folk reference magnetists?

We take the case developed in the preceding three sections to constitute a formidable case for reference magnetism as an explanation of the distinctive semantic stability of ‘ought’ and thereby an answer to the Normative Twin Earth challenge. However, we recognize that not everyone will be comfortable with the metaphysical and methodological assumptions required by the account as we have sketched it. An initial reply to such readers is that Moral Twin Earth cases are usually advertised 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 reference magnets. One might simultaneously deny that normative properties *actually* are highly elite (and hence reference magnets). 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—judgments 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.

The proposal we wish to explore here is that the error in judgments about Normative Twin Earth issues from an implicit 'epistemology of eliteness'. This hypothesis holds that, even if no candidate normative properties are distinctively elite, the judgements about Co-reference and Disagreement that feature in Normative Twin Earth arguments reflect speakers' implicit *beliefs* that normative properties are elite.

To turn this conjecture into a workable hypothesis that the normative realist can appeal to, we need to add several non-trivial details to the proposal. First: what makes a property believed-to-be-elite? We take it that, while the folk rarely categorize properties as 'elite' or 'not-elite' (not, in general, having kept up with the contemporary metaphysics literature), they do implicitly make judgments that can plausibly be interpreted as deployment of a proto-theory of eliteness. In particular, the folk do intuitively categorize some properties as 'gerrymandered' or not, and they do make judgments about disagreement and co-reference, as well as explanatoriness, resemblance, confirmation, and other notions that are tied to eliteness. Even if 'eliteness' is not explicitly tokened in folk theorizing, then, judgments about disagreement and reference are plausibly tied to other distinctive aspects of eliteness in a way that makes the best theory of what generates these judgments is couched in terms of folk-acceptance of a theory of eliteness. (We want to flag that what we have offered here is a non-trivial hypothesis that aims to explain empirical patterns of linguistic dispositions. Seriously defending this hypothesis would require significant empirical investigation.)[32]

A fallback proposal of this kind is revisionary, we have emphasized, as it is compatible with a rejection of the truth of some judgments about normative reference and disagreement. But it is distinct from a full-blown error theory about normative discourse. Consider an analogy: the liar paradox is arguably generated by judgments one must find compelling in order to be competent with 'true' (Eklund (2002)). 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. Such revisionary approaches reject some common-sense judgments, but not all of them. There are still, according to the theory, many true things we say about the normative. But the best theory needn't treat every piece of common sense as sacred, so long as it can provide a plausible explanation for why these judgments are compelling yet false. We think reference magnetism can be deployed in service of the latter task.

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

[32] Thanks to an anonymous referee for encouraging us to articulate what is at stake here.

Conclusions

We take the distinctive semantic stability of central normative terms to constitute the most serious challenge to the (meta-)semantic dimension of the naturalistic normative realist's research program. Such semantic stability seems inconsistent with 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. We take there to be a variety of ways to explain – or explain away – the apparent semantic stability of the normative. In this paper, we have argued that an appropriately developed form of reference magnetism can provide an especially elegant and principled answer to the challenge.

We have illustrated this answer by showing how it addresses one powerful instance of the semantic stability challenge: a general challenge to normative realism in the form of Terence Horgan and Mark Timmons' Moral Twin Earth argument. Reference magnetism is an independently well-motivated meta-semantic theory that dovetails beautifully with a diagnosis of why that generalized argument fails: the generalized argument assumes that the open feel of normative analyses must be explained by contingency in reference determination. Attention to reference magnetism as a meta-semantic mechanism shows why this assumption is at best highly controversial. We also showed that our account can explain the variability of semantic stability exhibited by different terms, exemplified (for example) by the apparent contrast between Normative 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 (this is the main theme of Edwards 2013). 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 Section 6, there is a principled way for the normative realist to offer the idea of reference magnetism for the normative 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, is capable of beautifully explaining semantic phenomena that otherwise threaten to constitute intractable challenges to normative realism, and is 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. Anyone who thinks we need to adopt additional theoretical commitments in order to account for the semantic stability of 'ought' will need to provide powerful reasons for thinking that the reference magnetism account is inadequate.

Acknowledgments

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, Mark Schroeder, Brett Sherman, Michael Smith, Tim Sundell, Pekka Väyrynen, and Sean Walsh, to several anonymous referees, to 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 Virginia Tech graduate seminar on Semantics of Normative Realism for helpful discussion of ancestors of this paper.

References

Bedke, Matthew S. (2010). Might All Normativity be Queer? *Australasian Journal of Philosophy*, 88(1), 41–58.

Björnsson, Gunnar and Steven Finlay (2010). Metaethical Contextualism Defended. *Ethics*, 121(1), 7–36.

Blackburn, Simon (1993). *Essays in Quasi-realism*. Oxford University Press.

Boyd, Richard N. (1988). How to be a Moral Realist. In Geoffrey Sayre-McCord, (Ed.), *Essays on Moral Realism*. (181-228). Cornell University Press.

Brandt, Richard B. (1979). *A Theory of the Good and the Right*. Prometheus Books.

Burgess, Alexis and Brett Sherman (2014). Introduction: A Plea for the Metaphysics of Meaning. In Alexis Burgess and Brett Sherman (Eds.), *Metasemantics: New Essays on the Foundations of Meaning* (1-16). Oxford University Press.

Cameron, Ross (2010). Vagueness and Naturalness. *Erkenntnis*, 72(2), 281–293.

Copp, David (2000). Milk, Honey, and The Good Life on Moral Twin Earth. *Synthese*, 124 (1-2), 113–137.

Dorr, Cian and John Hawthorne (2013). Naturalness. In Karen Bennett and Dean Zimmerman, (Eds.), *Oxford Studies in Metaphysics* (Vol. 8, 3-77). Oxford University Press.

Dowell, Janice (2016). The Metaethical Insignificance of Moral Twin Earth. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 11, 1-27). Oxford University Press.

Dreier, James (2004). Meta-ethics and the Problem of Creeping Minimalism. *Philosophical Perspectives*, *18*, 23–44.

Dunaway, Billy (2015). Supervenience Arguments and Normative Non-naturalism. *Philosophy and Phenomenological Research*, 91(3), 627-655.

Dunaway, Billy (2106). Expressivism and Normative Metaphysics. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 11, 241-264). Oxford University Press.

Dunaway, Billy (Ms.). The Metaphysical Conception of Realism.

Dworkin, Ronald (1996). Objectivity and Truth: You'd Better Believe It. *Philosophy and Public Affairs*, 25(2), 87–139.

Dworkin, Ronald (2011). *Justice for Hedgehogs*. Belknap Press of Harvard University Press.

Edwards, Douglas (2013). The Eligibility of Ethical Naturalism. *Pacific Philosophical Quarterly*, 94(1), 1-18.

Eklund, Matti (2002). Inconsistent Languages. *Philosophy and Phenomenological Research*, 64(2), 251–275.

Eklund, Matti (Forthcoming). *Choosing Normative Concepts*. Oxford University Press.

Enoch, David (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press.

Enoch, David and Tristram McPherson (Forthcoming). What do you mean "This isn't the Question"? *Canadian Journal of Philosophy.*

Fine, Kit (2001). The Question of Realism. *Philosophers' Imprint*, 1(1), 1–30.

FitzPatrick, William (2008). Robust Ethical Realism, Non-naturalism, and Normativity. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 3, 159–206). Oxford University Press.

Gibbard, Allan (2003). *Thinking How to Live*. Harvard University Press.

Hansson, Sven Ove (2015). Science and Pseudo-Science. *Stanford Encyclopedia of Philosophy* (Spring 2015 Edition). URL = <http://plato.stanford.edu/archives/spr2015/entries/pseudo-science/>.

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

Hawthorne, John (2006). Epistemicism and Semantic Plasticity. In *Metaphysical Essays* (185–210). Oxford University Press.

Hawthorne, John (2007). Craziness and Metasemantics. *The Philosophical Review*, 116(3), 427–441.

Horgan, Terence and Mark Timmons (1991). New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*, 16, 447–465.

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

Horgan, Terence and Mark Timmons (1992b). Troubles on Moral Twin Earth: Moral Queerness Revived. *Synthese*, 92(2), 221–260.

Horgan, Terence and Mark Timmons (1996). From Moral Realism to Moral Relativism in One Easy Step. *Critica*, 28(83), 3–39.

Horgan, Terence and Mark Timmons (2000). Copping Out on Moral Twin Earth. *Synthese*, 124(1-2), 139–152.

Horgan, Terence and Mark Timmons (2009). Analytical Moral Functionalism Meets Moral Twin Earth. In Ian Ravenscroft (Ed.), *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson* (221-236). Oxford University Press.

Huemer, Michael (2005). *Ethical Intuitionism*. Palgrave Macmillan.

Jackson, Frank (1998). *From Metaphysics to Ethics*. Oxford University Press.

Jackson, Frank and Philip Pettit (1995). Moral Functionalism and Moral Motivation. *Philosophical Quarterly*, 45(178), 20–40.

King, Jeffrey C. (1998). What is a Philosophical Analysis? *Philosophical Studies*, 90(2), 155–179.

Kramer, Matthew H. (2009). *Moral Realism as a Moral Doctrine*. Wiley-Blackwell.
Kripke, Saul (1980). *Naming and Necessity*. Harvard University Press.
Kripke, Saul (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press.
Lewis, David (1983). New Work for a Theory of Universals. *Australasian Journal of Philosophy*, 61(4), 343–377.
Lewis, David (1984). Putnam's Paradox. *Australasian Journal of Philosophy*, 62(3), 221–236.
Lewis, David (1986). *On the Plurality of Worlds*. Basil Blackwell.
Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. Penguin.
McPherson, Tristram (2015). What is at Stake in Debates Among Normative Realists? *Noûs*, 49(1), 123–146.
McPherson, Tristram (Forthcoming). Authoritatively Normative Concepts. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 12).
Manley, David (Ms.). Moral Realism and Semantic Plasticity.
Merli, David (2002). Return to Moral Twin Earth. *Canadian Journal of Philosophy*, 32(2), 207–240.
Moore, G.E. (1903). *Principia Ethica*. Cambridge University Press.
Plunkett, David and Timothy Sundell (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint*, 13(23), 1-37.
Putnam, Hilary (1975). The Meaning of 'Meaning'. In *Mind, Language, and Reality: Philosophical Papers* (*Vol. 2*, 215–271). Cambridge University Press.
Putnam, Hilary (1981). *Reason, Truth and History*. Cambridge University Press.
Quine, W. V. O. (1960). *Word & Object*. The MIT Press.
Railton, Peter (1986). Moral Realism. *The Philosophical Review*, 95(2), 163–207.
Sayre-McCord, Geoffrey (1997). 'Good' on Twin Earth. *Philosophical Issues*, 8, 267–292.
Scanlon, Thomas (2014). *Being Realistic about Reasons*. Oxford University Press.
Schaffer, Jonathan (2004). Two Conceptions of Sparse Properties. *Pacific Philosophical Quarterly*, 85 (1), 92–102.
Schoenfield, Miriam (2016). Moral Vagueness is Ontic Vagueness. *Ethics*, 126(2), 257-282.
Schroeder, Mark (2007). *Slaves of the Passions*. Oxford University Press.
Schroeder, Mark (2008). *Being For: Evaluating the Semantic Program of Expressivism*. Oxford University Press.
Shafer-Landau, Russ (2003). *Moral Realism: A Defense*. Oxford University Press.
Sider, Theodore (2012). *Writing the Book of the World*. Oxford University Press.
Skorupski, John (2010). *The Domain of Reasons*. Oxford University Press.
Smith, Michael (2013). A Constitutivist Theory of Reasons: Its Promise and Parts. *Law, Ethics, and Philosophy*, 1, 1-30.
Soames, Scott (2002). *Beyond Rigidity: the Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.
Street, Sharon (2008). Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying about. *Philosophical Perspectives*, *18*, 207–227.
Streumer, Bart (2013). Can We Believe the Error Theory? *Journal of Philosophy*, 110(4), 194-212.
Sturgeon, Nicholas L. (2002). Ethical Intuitionism and Ethical Naturalism. In Phillip Stratton-Lake, (Ed.), *Ethical Intuitionism: Re-evaluations*. (184-211). Oxford University Press.

van Roojen, Mark (2006). Knowing Enough to Disagree. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 1, 161-194). Oxford University Press.

Weatherson, Brian (2013). The Role of Naturalness in Lewis's Theory of Meaning. *Journal of the History of Analytic Philosophy*, 1(10), 1-19.

Wedgwood, Ralph (2007). *The Nature of Normativity*. Oxford University Press.

Williams, J. Robert G. (2007). Eligibility and Inscrutability. *The Philosophical Review*, 116 (3), 361–399.

Williams, J. Robert G. (2015). Lewis on Reference and Eligibility. In Ernest Lepore and Jonathan Shaffer (Eds.), *Companion to David Lewis* (367-381). Blackwell.

Williams, J. Robert G. (Ms.). Normative Reference Magnets.

Wright, Crispin (1992). *Truth and Objectivity*. Harvard University Press.