A core claim of normative realists is that normative predicates pick out properties. Traditional normative realists also hold the semantic view that central normative terms are context-invariant: the predicate ‘is morally right’, for instance, picks out the very same property when used at different times, by different speakers, or in different social contexts.¹ Our focus in this paper will be on a pair of metasemantic theses that many theorists see as central to vindicating traditional normative realism. The first thesis is that the reference of normative predicates is not epistemically constrained: just which property is picked out by ‘is morally right’ is not settled by the verdicts the speaker would endorse at the end of ideal rational reflection. The second thesis is that co-reference does not require convergence in speakers’ ideal normative judgments: competent speakers’ use of a moral predicate picks out the same property, even if speakers would arrive at divergent verdicts about which things are morally right at the end of ideal rational reflection.

It’s not hard to see why a normative realist might be attracted to a fallibilist, non-convergentist metasemantics. One common objection to normative realism is that convergence on normative questions is unlikely.²

1. Such context-invariance seems crucial to ensuring the kind of semantically guaranteed stability of topic that grounds direct logical agreement and disagreement. However, not all normative realists endorse context-invariance: contextualists, for instance, hold that normative predicates pick out different properties depending on the context of use. On such accounts, different speakers who use normative terms with the same meaning may not pick out the same property. See (Björnsson and Finlay 2010) for a recent defense of contextualist realism. In this paper, we will focus exclusively on traditional, context-invariant realism. Of course, normative realists typically endorse further substantive constraints on the nature of normative properties or the practical role of normative concepts. But the core metasemantic issue we raise in this paper does not hinge on these substantive commitments.

2. See (Enoch 2009) for a thorough survey of the different ways in which disagreement has been thought to threaten normative realism, and for a careful explanation of how realists might resist such arguments. In his monograph on disagreement arguments (Tersman 2006) focuses on three main strategies for using disagreement to undermine realism: arguing that anti-realism is the best explanation of actual disagreement, arguing that moral facts must be epistemically accessible, and arguing that the best interpretation of moral terms would construe disagreeing speakers as talking past each other. Like Tersman, we think the most serious challenge for the realist
In practice, normative disagreements often seem rationally intractable: even sophisticated ethical theorizing appears incapable of resolving deep disagreements.\footnote{For a highly influential skeptical evaluation of the role of reasoning and reflection in fostering convergent moral judgment, see (Haidt 2001). See also (Mercier and Sperber 2011) for a more general challenge to the idea that the function of reflective reasoning is to generate more accurate, intersubjectively valid beliefs. For a recent realist attempt to resolve fundamental disagreements among moral theorists, see (Parfit 2011).} It seems prudent, therefore, for a normative realist to avoid any commitment to disagreements being rationally resolvable. Better not to give hostages to epistemological fortune.

But realists cannot rest content with a simple denial of convergence. The problem is that the traditional, broadly descriptivist, approach to meaning and reference determination fixes the reference of a predicate by appealing to the speaker’s ideal reflective judgments about what that predicate applies to. On such accounts, speakers co-refer with the term ‘morally right’ only if they would converge on the same verdicts about which actions count as morally right after ideal reflection. If realists wish to avoid any commitment to ideal convergence, then they must provide a viable alternative account of how the reference of normative predicates is fixed.

Normative realists, we believe, have not been sufficiently aware of the difficulties involved in providing such an alternative metasemantics. Our aim in this paper is to clarify the nature of this reference-fixing task and the challenges that arise for a non-convergentist normative realist. Our focal point will be Richard Boyd’s seminal paper, ‘How to be a moral realist’, which has been the classical statement of a non-convergentist metasemantics in metaethics. According to Boyd, facts external to the subject’s understanding help determine which properties are picked out by moral terms. The most prominent objection to this externalist approach, the Moral Twin Earth argument, assumes that Boyd’s account secures determinate reference, but argues that it leads to counterintuitive verdicts about sameness of meaning.

\begin{quote}
\textit{...to avoid any commitment to ideal convergence, they owe us a viable alternative account of how the reference of normative predicates is fixed.}
\end{quote}

Opting instead for a fallibilist theory of reference determination, Boyd is right to see these two theoretical tasks as closely related: a theory of reference for normative predicates must interpret them as picking out properties that we can know about, and a theory of knowledge must explain how we could have reliable access to the properties normative terms do in fact pick out. Boyd’s account, however, seems to conflate these two theoretical tasks. The resulting position, we’ll argue, leads to an acute form of semantic indeterminacy, which conflicts with the fallibilist commitments that led normative realists to a non-convergentist metasemantics in the first place. A second reason for focusing on Boyd is that his theory has significantly shaped the debate over non-convergentist realism in the metaethics literature. Not only was Boyd the original target of the Moral Twin Earth argument, but many realists have responded to these arguments by proposing refinements or modifications of the externalist line developed by Boyd (e.g. Sayre-McCord 1997; Copp 2000; Brink 2001; van Roojen 2006). A final reason to examine Boyd in detail is that his account combines distinct theoretical elements. Teasing apart these different elements will help give an overview of the standard metasemantic strategies available to the non-convergentist normative realist. We will then be in a better position to understand the challenges non-convergentist realists face in securing a determinate and plausible referential assignment for normative predicates. Although non-convergentist theories will face similar challenges in explaining reference-determination in any domain, those challenges are significantly exacerbated in the case of normative predicates.
deciding which individuals are bachelors in any possible situation, actual or counterfactual. Another speaker, however, might associate the term ‘bachelor’ with a slightly different criterion: to be a bachelor, she assumes, is to be an unmarried male (whether eligible or not). Thus, after ideal, fully informed reflection, she would call the Pope a bachelor, while you would deny this claim under similar ideal conditions. According to the descriptivist, such divergences in ideal verdicts suffice to show that the term ‘bachelor’ picks out different properties in your idiolect and hers.

The traditional descriptivist account of reference determination has some important advantages. It provides a clear principle for discriminating the property picked out by your term ‘bachelor’ from all similar but non-coextensive properties: your own ideal reflective verdicts settle the question of which property your words represent. Moreover, descriptivism provides a non-arbitrary interpretive principle: the correct interpretation of your word is determined by your own ideal reflective verdicts, rather than an interpretive principle that might strike you on reflection as irrelevant or perverse. Grounding semantic interpretation in the subject’s own idealized perspective helps explain why it is an appropriate standard for assessing the correctness of her beliefs and predicting and explaining her behavior.5

It’s worth emphasizing that descriptivism is consistent with full-blooded metaphysical realism about the properties picked out. Descriptivists endorse a metasemantic thesis about reference determination, not a metaphysical thesis about the nature of the

A terminological point. Intuitively, an ordinary predicate like ‘is a bachelor’ represents the property of being a bachelor. We use the term ‘reference’ for this semantic value: the reference of ‘is a bachelor’ is the property of being a bachelor (which can be modeled in possible world semantics as a function from possible worlds to extensions). Some theorists reserve the term ‘reference’ for the actual extension of a predicate: on this usage, the reference of ‘bachelor’ is the set of objects that happen to be bachelors in the actual world. Taking properties (or extensions in all possible worlds) as the semantic value of predicates allows us to more easily characterize their intuitive representational function and to capture their contribution to the truth of modal claims.

4. A terminological point. Intuitively, an ordinary predicate like ‘is a bachelor’ represents the property of being a bachelor. We use the term ‘reference’ for this semantic value: the reference of ‘is a bachelor’ is the property of being a bachelor (which can be modeled in possible world semantics as a function from possible worlds to extensions). Some theorists reserve the term ‘reference’ for the actual extension of a predicate: on this usage, the reference of ‘bachelor’ is the set of objects that happen to be bachelors in the actual world. Taking properties (or extensions in all possible worlds) as the semantic value of predicates allows us to more easily characterize their intuitive representational function and to capture their contribution to the truth of modal claims.

5. The core idea behind the descriptivist model of reference determination — that subjects’ ideal verdicts settle the precise reference of their own words — went largely unchallenged until the work of semantic externalists in the early 1970’s. Semantic externalists forcefully pointed out that we often take ourselves to be partially ignorant or mistaken about the essential nature of the objects and properties our words represent, and that substantive empirical inquiry may be required to correct these mistakes (Donnellan 1970; Kripke 1980; Putnam 1970, 1972, 1973). Over the past twenty years, new versions of the descriptivist model have been developed to accommodate externalists’ counterexamples. On such neo-descriptivist views, reference is settled by subjects’ ideal verdicts relative to contingent information about their actual empirical circumstances. We’ll return to neo-descriptivism in §2 below.
properties themselves. The claim that the subject’s ideal judgments settle which property she is talking about involves no commitment to the claim that her ideal judgments make it the case that those properties are instantiated, or that the properties picked out are somehow constituted by her ideal dispositions.6

For our purposes, the important point is that the descriptivist approach to reference determination seems to leave no room for error at the ideal limit. The notion that subjects’ own ideal verdicts settle precisely which things fall into the extension of their words amounts to an Ideal Accessibility constraint on reference determination. In the case of normative terms, the constraint can be formulated roughly as follows:

(IA) Full empirical information together with ideal reflective powers would put the subject in a position to specify precisely which property her term ‘is right’ picks out — so that she could identify which actions, actual or possible, her term applies to.7

6. For instance, a neo-descriptivist may hold that ‘gold’ picks out the actual natural kind, if any, that satisfies your everyday criteria for applying that term. So your term ‘gold’ would pick out the chemical element Au in all possible worlds (not a property somehow constituted by your ideal judgments). For an extended defense of the compatibility of metaphysical realism with traditional accounts of reference-determination see (Pettit 1991). Proponents of two-dimensional semantics and other neo-descriptivist accounts of reference determination have also emphasized this point (see §2 below).

7. To avoid circularity, the empirical information must exclude normative, semantic, and metasemantic descriptions. (IA) is similar to a constraint on conceptual competence proposed by David Chalmers and Frank Jackson. Chalmers and Jackson have been particularly clear about the importance of such an accessibility constraint in motivating descriptivist and neo-descriptivist approaches to reference determination: conceptual competence, they suggest, puts the subject in a position to identify the extension of her words in all possible worlds on the basis of ideal reflective powers and full empirical information about the supervenience base in her actual environment (Chalmers and Jackson 2001). See also the discussion of “the scrutability of reference” in (Chalmers 2004) and (Chalmers 2012), and Christopher Peacocke’s defense of a similar accessibility constraint for moral terms in (Peacocke 2004c: ch. 7).

The thesis is not that subjects must be able to state a precise definition of what their normative terms refer to: your understanding of what it is you’re talking about may be embodied in a heterogeneous set of implicit assumptions about the reference, a ‘folk theory’, that guides your deliberation about hypothetical cases (Lewis 1972; 1994: 56; Jackson 1998a: 32). Or your understanding may be embodied in a small core of epistemically basic inferential dispositions that, together with ideal a priori reasoning, serve to constrain the referential assignment (Peacocke 1992: 22; 2004a: 96). The point is that the subject’s ideal verdicts suffice to distinguish the reference of her normative terms from all other non-coextensive referential candidates. The Ideal Accessibility constraint (IA) makes convergence a necessary precondition for co-reference and makes co-reference difficult to achieve for descriptivism, especially in the normative domain.

Considered in itself, (IA) strikes many theorists as problematic. What exactly are the relevant idealization conditions? And, more importantly, why should we suppose that ideal reflection cannot possibly be mistaken? Epistemologists no longer insist on infallibility as a precondition for knowledge — for such insistence quickly leads to skepticism. So why should we revive a form of infallibility in the theory of reference?

The guiding idea behind Richard Boyd’s externalist account of the reference of moral terms is precisely to bring the theory of reference determination into line with a fallibilist epistemology. Boyd pithily sums up the rationale behind his approach in his paper on moral realism:

The connection between causal theories of reference and naturalistic theories of knowledge and of definitions is quite intimate: reference itself is an epistemic notion and the sorts of causal connections which are relevant to the reference are just those which are involved in the reliable regulation of belief. (Boyd 1988: 195)
It’s helpful to break Boyd’s line of reasoning into three distinct steps, by drawing on his later elaboration of his position. First, Boyd ties the notion of reference to the notion of epistemic access. Second, he suggests that the sort of epistemic access relevant to characterizing the reference relation is precisely the sort that grounds knowledge. Our best theory of knowledge can therefore be recruited to articulate a theory of reference determination. Third, Boyd advocates a particular naturalistic account of knowledge, which involves causal mechanisms that reliably connect beliefs to features of the world. These three different steps are nicely expressed in his paper ‘Metaphor and theory change’. First, epistemic access:

I propose to analyze the notion of reference in terms of the notion of epistemic access. I hold that, for any particular general term, the question of reference is to be understood as the question: to which kind (or kinds), or property (or properties), or magnitude (or magnitudes), ... and so on, does our use of this term afford us epistemic access? (Boyd 1993, 505. Ellipsis in original.)

Second, the link between epistemic access and a general theory of knowledge:

The notion of reference is fundamentally an epistemological notion. Semantic Theory — insofar as it is a branch of Philosophy — is a branch of epistemology. [...] In deciding issues in the theory of reference it is, therefore, appropriate to make use of the best available epistemological theories. The true theory of reference will be a special case of the true theory of knowledge: the true theory of reference for theoretical terms in science will be a special case of the true theory of the epistemology of science. (Boyd 1993: 503. Italics in original.)

Third, since the best general account of knowledge invokes reliable causal mechanisms, the theory of reference determination must follow suit:

A causal theory of reference is true precisely because reference is an epistemological notion and causal theory of knowledge is true. (1993, 503. Italics in original.)

The first step in Boyd’s reasoning should be uncontroversial in the present context. Theorists in the descriptivist tradition obviously agree that reference to an object, property or relation entails epistemic access to these things: a correct specification of the reference can always be known through ideal reflection and full information about the relevant supervenience base. Most semantic externalists also agree that reference entails epistemic access of some sort: the reason why it is absurd to suppose your term ‘water’ refers to Julius Caesar or to strawberry jam is because your use of the term affords no epistemic access whatever to those things: the ability to refer seems to entail some knowledge of what it is one is referring to.

The second step is where Boyd and the descriptivist part ways. Boyd suggests that a fallibilist theory of knowledge can be used to characterize the epistemic access relation that constitutes reference. This allows him to reject the requirement that a perfectly accurate specification of the reference be accessible through ideal reflection.

The last step in Boyd’s line of reasoning has attracted the most critical attention. Many metaethicists have been uncomfortable with the close analogy Boyd seeks to draw between scientific and moral epistemology (Copp 1990; Loeb 1998; Svavarsdóttir 2001: VI). There has also been considerable skepticism about the role of causation in securing reference in the moral domain, fueled by the influential

8. Ruth Millikan is the exception to this rule: on her account a symbol may have a proper function of representing a certain property (because of natural selection in the past), even though the organism is currently incapable of detecting that property (e.g. due to a developmental defect in the organism or a non-standard environment) (Millikan 1984, 1993).
Moral Twin Earth thought experiment (Horgan and Timmons 1992b, 1992a). Some normative realists have responded to these worries by seeking to divorce Boyd’s approach to reference determination from his commitment to a causal theory of knowledge. By substituting a different account of moral epistemology for Boyd’s causal theory modeled on the epistemology of theoretical terms in science, these theorists hope to vindicate realism without incurring Boyd’s controversial commitment to a causal theory of reference for moral terms (Sayre-McCord 1997; Brink 2001; van Roojen 2006).

Our primary concern in this paper is whether the realist can avoid a commitment to ideal convergence. Our focus, therefore, will be on step two of Boyd’s reasoning, where he parts company with the descriptivist tradition. In the next section, we formulate a general challenge for theories of reference like Boyd’s that draw on fallibilist accounts of knowledge. In the following three sections, we examine the details of Boyd’s account and consider how it might be supplemented so as to address this challenge.

2. Indeterminacy and circularity
The fundamental problem with Boyd's approach, we think, is that it links the theory of reference determination much too tightly to the theory of knowledge. Indeed, it’s tempting to read the second step in Boyd’s reasoning as suggesting that a fallibilist theory of knowledge can do double-duty as a theory of reference. However, these two theories must be sharply distinguished. Fallible epistemic access is fine in a theory of knowledge. But treating a theory of knowledge as a theory of reference determination gives rise to two fundamental problems: indeterminacy and circularity.

9. As we saw, Boyd suggests that a theory of reference should “make use of” our best epistemological theories and that the theory of reference is a “special case” of the theory of knowledge. There may be room for debate about how exactly the relation between theory of reference and theory of knowledge should be understood on Boyd’s approach. However, such exegetical issues needn’t concern us here. Our goal in this section is to explain why these two theories must be distinguished—a point which is not, in our experience, always sufficiently appreciated by proponents of normative realism.

Normative realism: co-reference without convergence?

Consider a simple analogy. Joe is participating in an experiment at his local cognitive psychology lab. His task is to throw darts at a series of randomly shaped targets of varying sizes. Joe is asked to use different throwing methods (overhand, underhand, behind his back) and the conditions are varied (bright light, wind, strobe). The task is a challenging one, but Joe’s a pretty good dart thrower and gets many of his darts in the targets.

Two experimental philosophers are studying the results. The epistemologist, Faith, is interested in determining which dart-throwing methods are reliable. She looks at the target and records the distribution of hits inside and outside the boundaries. This allows her to calculate the relative and absolute reliability of the different methods Joe is using. Ultimately, Faith’s goal is to determine which methods are sufficiently reliable so that Joe’s successful hits using that method should be considered non-accidental.

The semantic theorist, Charity, has a very different theoretical agenda. She tries to identify the precise boundaries of the target Joe is aiming at: is he aiming at a ten-inch irregular hexagon, or a fourteen-inch misshapen ovoid, or some other even more irregular shape? Identifying the precise boundaries of the target will then allow her to determine which of Joe’s hits are successful and which are unsuccessful. How should Charity proceed? Obviously she cannot simply look at the targets to answer her theoretical question: that would be begging the question at issue.

Can Charity identify the precise boundaries of the target Joe is aiming at on the basis of information about the distribution of hits on the wall? It’s hard to see how she could. It would be reasonable for her to start with the assumption that some of Joe’s hits are on target and others are not. If she makes this assumption, she can conclude that the boundary is located somewhere within the distribution of hits; but any

10. To make this problem vivid, we can imagine that Joe has a contrarian streak: he may decide to aim at a shape that overlaps with the researchers’ target, or he may decide to throw darts mechanically without really taking aim at any particular target.
verdict about where exactly the boundary is located seems arbitrary given her evidence. This is the problem of indeterminacy.

Would it help if she were told that Joe was a reliable dart thrower, or even an expert? With this information, Charity may be able to narrow down the range of likely boundaries: she’d expect the pattern of throws to cluster more narrowly around the target and she’d look for boundaries that carve near the edges of that cluster. But the problem of indeterminacy remains: her current evidence still does not allow her to make a non-arbitrary choice that discriminates successful from non-successful hits.

Moreover, bringing in information about Joe’s reliability raises the specter of circularity. If Charity is told that Joe was using a method that was 84% reliable, she’ll have more information to use in drawing her boundaries. But where exactly does this information come from? Presumably it comes from someone like Faith, who is looking at the target and calculating Joe’s reliability with respect to that target. But the accuracy of Faith’s statistics depends on her knowing the correct answer to Charity’s question about the precise shape of Joe’s target. So relying on Faith’s statistics would introduce a vicious circularity in Charity’s reasoning: Charity couldn’t gather those statistics herself without peeking at the target whose boundaries she’s trying to deduce.

The darts analogy, we believe, highlights an important challenge for normative realists who hope to appeal to reliable belief-forming processes to explain the determination of reference. What the normative realist needs is some way of demarcating the property picked out by a term like ‘morally right’ on the basis of the subject’s use (or understanding) of that term. If we assume certain moral belief-forming methods are reliable, we can infer that the resulting beliefs are likely to be true. But which ones are true and which are false? The fact that a moral belief was generated by a reliable but fallible method provides no principled basis for drawing a line between true moral beliefs and false ones. The challenge, then, is to find a non-arbitrary way of drawing this line without relying on any question-begging information.

It may seem that there is an easy answer to the challenge. Why couldn’t a realist simply accept that the referential target is indeterminate? In the darts analogy, physical information about Joe’s dispositions and his environment will allow Charity to project a shaded pattern onto the wall that represents the probability that his darts will land in certain locations. Assumptions about Joe’s reliability will then allow Charity to draw a vague boundary between hits in the dark center that she’d count as clearly on-target versus hits in the lightly shaded extremities that she’d count as clearly off-target, leaving a large band of indeterminate cases between those two regions. Surely a normative realist could say something similar.

The problem is that this simple reliabilist account of reference fixing conflicts with the fallibilist epistemology that motivates Boyd and other non-convergentist realists. According to simple reliabilism, the hits in the lightly shaded extremities can be dismissed as performance errors. However, we can’t discard hits in the penumbra or the black center as determinately false. But the key fallibilist idea is that the subject may be disposed to get the moral facts wrong—even at the ideal limit of inquiry, we may misclassify wrong actions as ‘morally right’ and right actions as ‘wrong’. Simple reliabilism cannot vindicate this key fallibilist idea. The problem is even more vivid in the interpersonal case. If we assume, as traditional realists do, that different speakers all pick out the very same property when they use moral terms, then every controversial case—case about which there is persistent disagreement—will fall into the large band of indeterminacy. So there simply is no fact of the matter about controversial cases.

It’s worth noting that referential indeterminacy in persistently contested cases may be plausible in some domains. Perhaps we should accept that stable disagreement about which things are cozy is a sign of indeterminacy in the coziness facts. But the analogous claim seems implausible in the moral domain: even when there is persistent disagreement or uncertainty, we often assume that there is a fact of the matter whether a given action is morally right. Indeed, normative debate would lose much of its interest if we were to accept that there
are no truths about morally contested cases. This is not to say, of course, that there can’t be any residual indeterminacy in the moral domain. The claim is simply that there are true answers in many moral disputes.\(^{11}\)

Of course, the darts analogy does not show that the challenge of resolving indeterminacy cannot be met by an externalist metasemantics: the point of the analogy is merely to give a clear picture of where the difficulties lie. The case of belief is far more complex than that of darts, so there are further theoretical resources that a fallibilist epistemology might bring to bear on the question of reference determination. In the remainder of the paper, we’ll consider Boyd’s theory of reference in more detail. Although we’ll frame our discussion in terms of Boyd’s account and how it might be elaborated, our interest is not primarily exegetical: we wish to explore whether externalist theories of reference along the lines Boyd proposes have the resources to avoid the problems highlighted in this section. In a nutshell, our Fundamental Question is:

\((FQ)\) If we give up Ideal Accessibility, what further factors (beyond what’s ideally accessible to speakers) will secure reference to a sufficiently determinate property?

Non-convergentist normative realists must offer a substantive answer to this question. On their view, we’re not just speaking as if there were a single property about whose instantiation we are disagreeing, there really is a single property that is picked out by our different uses of a normative predicate. If so, there must be some way to specify precisely which property is the reference of this term. Given the realist commitment to genuine representation of properties, metasemantic quietism is not an option.\(^{12}\)

We should not assume that there is an easy answer to (FQ). In fact, persistent doubts about the adequacy of externalist metasemantics have helped to fuel the renewed popularity of the (IA) constraint on reference determination in the philosophy of mind and language. Many contemporary theorists in the descriptivist tradition explain subjects’ fallibility about the precise nature of the properties their words pick out by distinguishing between subjects’ current criteria for applying a predicate and the criteria they would accept after ideal reflection on the relevant empirical facts (specified in non-question-begging terms).\(^{13}\) Proponents of 2D semantics, for instance, suggest that the reference of your term ‘water’ is fixed by your ideal verdicts about which things count as water, formed on the basis of full microphysical information about your actual environment and ideal rational reflection. Thus your word ‘water’ actually refers to H\(_2\)O in all possible worlds because that is the verdict you would ideally endorse if you were fully informed about your actual environment; but if your actual environment were like Twin Earth, your word would refer to XYZ instead. On this account, external causal, historical or nomic relations are relevant only insofar as the subject’s ideal verdicts would

\(^{11}\) See (Loeb 1998: 290) for similar observations. The pressure to resolve contested moral cases is related to the role moral judgment plays in identifying standards for guiding action: the more determinate the normative properties are, the more guidance they provide. There may be less pressure to avoid indeterminacy in other domains. There seems to be very little at stake in resolving indeterminacy for superficial properties like being cozy. But in any epistemically ambitious domain of inquiry, such as biology, there will be pressure to identify properties as the reference that explains how everyday judgments may be ignorant or mistaken from the point of view of ideal theory. In the case of normative domains, we suggest, there is a potentially distinct interest in identifying action-guiding standards capable of adjudicating hard cases.

\(^{12}\) Quasi-realists like (Gibbard 2003) might be able to explain why it is acceptable to speak as if there were a single property picked out by normative predicates. But such metasemantic theories offer no explanation of how exactly speakers’ use of those predicates secures reference to one property rather than another: why does this individual speaker’s use of the predicate ‘is morally right’ pick out this particular extension? Without such an account of reference determination, we have no reason to suppose that speakers really do secure determinate reference with their use of normative predicates, as opposed to engaging in some non-representational linguistic practice (for a related point, see Schroeter and Schroeter 2005). We believe pragmatists like (Brandom 1994) face similar problems.

\(^{13}\) This response has been most forcefully articulated using a two-dimensional semantic framework (Lewis 1994; Chalmers 1996, 2004; Chalmers and Jackson 2001; Jackson 1998a, 1998b). For an influential alternative formulation of a neo-descriptivist position, see (Peacocke 1992, 1998, 2004b).
be sensitive to such information. Such accounts show that it’s possible to respect (IA) while accommodating externalist thought experiments that highlight subjects’ fallibility about the nature of the reference.

One advantage of respecting (IA) is that doing so promises to reduce indeterminacy to plausible levels. In particular, it provides a plausible, non-arbitrary way of distinguishing which of the subject’s current beliefs are true and which are false: the true ones are those the subject would endorse if fully informed and ideally reflective. Insofar as there remain cases about which the subject is undecided at the ideal limit, reference itself is indeterminate: either the property picked out is vague or the reference relation does not single out a unique property.

3. Reliabilism and reference determination

To see whether Boyd’s metasemantics falls prey to the indeterminacy and circularity objections highlighted by the darts analogy, we need to take a look at the details of his proposal. Although the key passages have been repeatedly quoted in the metaethics literature, there hasn’t been much analysis of how exactly the account is supposed to work. In fact, Boyd’s ‘causal regulation’ account of reference determination is surprisingly complex. We start this section by teasing out distinct strands in the account, each of which has played an important role in general metasemantic theories in the philosophy of mind and language. Distinguishing these different strands in Boyd’s account will put us in a better position to assess the viability of externalist metasemantics.

Appealing to the reliability of belief-forming mechanisms is a familiar move in externalist theories of knowledge. Reliabilism about knowledge can be formulated roughly as follows:

\[(K) \text{ A belief is knowledge iff it is true and the product of a reliable belief-forming mechanism.}\]

Boyd’s suggestion is that reference is determined by the very same mechanisms that figure in a reliabilist theory of knowledge. This core reliabilist idea can be summed up roughly as follows (taking t-beliefs to be beliefs that would be naturally expressed using the term t):

\[(R) \text{ t refers to k iff its use is governed by belief-forming mechanisms that over time tend to produce t-beliefs that are approximately true of k.}\]

Boyd provides examples of relevant belief-forming mechanisms, highlighting recognition dispositions (‘There’s a dog!’) as well as mechanisms that underpin co-reference among members of a linguistic community (acceptance of testimony, deference to experts):

Such mechanisms will typically include the existence of procedures which are approximately accurate for moral terms, since worries emerge in different ways depending on which strand is emphasized.

Boyd formulates his account of the reference of moral terms as follows:

\[\text{Roughly, and for nondegenerate cases, a term } t \text{ refers to a kind (property, relation, etc.) } k \text{ just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term } t \text{ will be approximately true of } k \text{ (excuse the blurring of the use-mention distinction).} (1988, 195)\]
recognizing members or instances of $k$ (at least for easy cases) and which relevantly govern the use of $t$, the social transmission of certain relevantly approximately true beliefs regarding $k$, formulated as claims about $t$ (again excuse the slight to the use-mention distinction), a pattern of deference to experts on $k$ with respect to the use of $t$, etc. (1988: 195)

This emphasis on coordination introduces a social dimension to Boyd’s externalist account of reference determination: what an individual means by her term depends on the ways others in the community use the term and the ways they are linked together in an information-sharing system.

When relations of this sort obtain, we may think of the properties of $k$ as regulating the use of $t$ (via such causal relations), and we may think of what is said using $t$ as providing us with socially coordinated epistemic access to $k$; $t$ refers to $k$ (in non-degenerate cases) just in case the socially coordinated use of $t$ provides significant epistemic access to $k$, and not to other kinds (properties, etc.). (1988: 195)

In this final summary of his proposal, Boyd claims that the kind picked out by a term regulates the community’s use of that term. The idea of regulation suggests some sort of causal feedback from the property $k$ ensures that the communal use of $t$ tracks the facts about $k$. However, the mechanisms Boyd invokes in this key passage (recognition dispositions, testimony, deference to experts) do not actually seem to provide any causal feedback from the world. Moreover, as Mark van

15. Testimony and deference are mechanisms of social coordination: they ensure that different speakers in a community tend to have similar patterns of $t$-beliefs, but they do not in themselves ensure that those $t$-beliefs more accurately reflect facts about a property like $k$. Our beliefs about what’s fashionable, for instance, are governed by testimony and deference without thereby being regulated by objective features of the world — at least in any normal sense of objectivity. Moreover, our beliefs about what’s fashionable are also governed by recognition dispositions: once appropriately indoctrinated, we’ll be disposed to recognize zebra stripes and navel rings as fashionable when we encounter them. But these recognition dispositions do not seem to constitute regulation of our representational practice with ‘fashionable’ by objective features of the world. What seems to be missing is some feedback mechanism that ensures that our recognition dispositions change in such a way that they come to accurately reflect facts about an independent feature of the world.

Roojen has pointed out (2006, 176), the core reliabilist idea in Boyd’s account can be detached from the causal regulation requirement. We will examine the core reliabilist idea in this section, and postpone consideration of causal feedback until the next section.

A third strand in Boyd’s position is his realism about properties (kinds, relations, objects, etc.) picked out by the reference relation. Although Boyd does not rely on the mind-independent nature of potential referential candidates to secure reference, other externalist metasemantic theories have given a large role to such metaphysical facts (Lewis 1984). We’ll examine this possibility in section 5 below. Two further aspects of Boyd’s proposal are worth mentioning, even though they bear less directly on our primary worries about his account. As we’ve seen, Boyd postulates a social element in the determination of reference. And Boyd’s account is holistic: the reference of a target expression $t$ depends on the reference of other terms in the subject’s language that figure in the subject’s system of $t$-beliefs. We’ll touch briefly on these two aspects of Boyd’s account at the end of this section.

Let’s now focus on Boyd’s core reliabilist idea summed up in (R). Because it emphasizes the reliability of belief-forming mechanisms involving $t$, rather than historical links to past encounters with $k$, (R) more closely resembles nomic or information-based theories (Stampe 1977; Dretske 1981; Fodor 1987) than it does the original causal-historical theories (Kripke 1980; Putnam 1972; Devitt 1981). Causal-historical theories assign reference on the basis of a causal chain of events linking the subject’s token use of $t$ to past encounters with $k$: e.g. your use of ‘dog’ refers to dogs in virtue of your use being linked via an historical chain of co-referential intentions to an original baptismal
event in which the term was introduced by pointing to a dog. Nomic theories of reference, in contrast, appeal to law-like counterfactual relations between the representation type and features of the world. To a first approximation, your expression ‘dog’ represents dogs because it’s associated with a mental representation that is reliably caused by the presence of dogs in your environment: e.g. you’re disposed to internally judge ‘there’s a dog!’ whenever directly confronted with a dog. The difference between historical and nomic approaches is brought out by the Twin Earth thought experiment. Historical approaches can vindicate the intuition that the chemically ignorant Oscar refers to H\textsubscript{2}O with his term ‘water’, because H\textsubscript{2}O was the kind of stuff that played the relevant historical role leading to his current use of the term. Pure nomic theories cannot vindicate this result, since they assign reference on the basis of counterfactuals like what would cause Oscar to say ‘that’s water!’: both XYZ and H\textsubscript{2}O would play the very same role. According to a pure nomic theory, Oscar’s word applies to the disjunctive kind H\textsubscript{2}O-or-XYZ. So Oscar and Twin-Oscar refer to exactly the same kind of stuff.

16. Such baptismal reference-fixing events may be plausible in the case of proper names, but the idea of a naming ceremony is harder to credit in the case of commonsense natural kind terms or moral terms (Boyd 1993: 491–3, 496–9). For recent metaethical theories that appeal to the Putnamian model of causal-historical chains of referential intentions, see (Copp 2000; Brink 2001). Such theories face strong objections on two fronts. The first type of objection focuses on contemporary users of normative terms: it seems implausible that what we mean entirely depends on the intentions of earlier generations of English speakers. The second type of objection targets the original reference-fixing events: how exactly did the first users of moral predicates manage to secure determinate reference to moral properties? In particular, how did their reference-fixing intentions when introducing a certain predicate manage to single out the property of moral rightness? Once we focus on this question, we seem to be back with our original problem of reference-determination: we need to know how speakers’ cognitive dispositions, together with facts about their environment, secure determinate reference to particular normative properties. Did our ancestors secure reference in virtue of grasping a criterion that puts them in a position to identify exactly which actions fall into the extension of the property? If not, what extra factors are involved in fixing reference? In particular, how do these extra factors secure reference in such a way that allows for ignorance and error at the ideal limit of reflection?

Can a reliabilist account of reference along the lines of (R) actually fix a determinate reference? Consider (R) once again:

\[
(R) \ t \text{ refers to } k \text{ iff its use is governed by belief-forming mechanisms that over time tend to produce } t\text{-beliefs that are approximately true of } k.
\]

(R) specifies how the reference of \( t \) is determined by facts about subjects’ belief-forming mechanisms. Identifying the reference of \( t \) will then allow us to specify truth-conditions for \( t\)-beliefs, which in turn allow us to distinguish true \( t\)-beliefs from false ones. If (R) is to help us in this project, it must not presuppose any prior knowledge of which \( t\)-beliefs are true — for that is the question (R) is supposed to answer. Just as Charity would be begging the question if she peeked at Joe’s target, any assumption about which \( t\)-beliefs are true would involve the kind of circularity highlighted in the previous section.

To avoid circularity, we must understand (R) as requiring us to favor a candidate referential assignment \( k \) insofar as that assignment allows us to construe the belief-forming mechanisms governing \( t \) as reliable. The idea is to assign a reference to \( t \) by first looking at the output of the relevant belief-forming mechanisms — the total set of beliefs involving \( t \) that those mechanisms would generate over time. We then consider different possible referential assignments \( k_1, k_2, k_3 \), etc. The reference of \( t \) is whichever assignment makes that set of \( t\)-beliefs approximately true.

To get a better understanding of how (R) functions in a theory of reference determination and of the difficulties it engenders, it will be helpful to compare it with a close relative: a truth maximization constraint. David Lewis suggests that we can specify the reference of a target term \( t \) as the best satisfier of the subject’s \( t\)-theory. Lewis’s constraint requires interpreters to favor referential assignments insofar as they maximize the truth of the subject’s \( t\)-beliefs: if there is a feature of the subject’s environment that makes the subject’s beliefs perfectly true, then it is the reference of \( t \); if there is no perfect satisfier but some
feature(s) close to perfectly satisfying the subject’s $t$-beliefs, then the reference is the one that best satisfies the $t$-theory; and if there is no close enough satisfier then $t$ fails to refer (Lewis 1970, 1972, 1994). Of course, Boyd is focused on the $t$-beliefs subjects would form rather than their current $t$-beliefs. But Lewis’s maximization constraint can be tweaked to accommodate this aspect of Boyd’s account:

$$(M) \, t \text{ refers to } k \text{ iff its use is governed by belief-forming mechanisms that tend to produce } t\text{-beliefs that are maximally true of } k.$$  

Boyd’s reliabilist constraint (R) is structurally similar to (M) in two important ways. First, both constraints are holistic: they take reference assignment for $t$ to depend on the totality of $t$-beliefs, rather than some privileged subset of those beliefs (e.g. the subset that the subject herself treats as decisive in determining $t$’s reference). Second, both constraints involve a kind of interpretive charity: they require us to construe the majority of $t$-beliefs as true. However, (R) is significantly weaker than (M): instead of requiring us to maximize the percentage of true beliefs, it merely requires that the relevant set of beliefs be approximately true. In effect, (R) is a relaxation of the maximization constraint (M).

A prima facie advantage of Boyd’s reliabilist approach is that it seems to allow much more scope for subjects’ ignorance and error about the reference of their terms. Unlike Lewis, Boyd does not require us to construe the subject as maximally successful in her epistemic endeavors, favoring interpretations that generate the highest proportion of true to false beliefs. (R) requires us to construe the subject’s belief-forming procedures as reliably leading to an approximately true body of beliefs over time — a constraint that seems to leave plenty of room for false beliefs.

The problem, however, is that (R) is much too weak to yield a determinate referential assignment. The job of a theory of reference is not merely to accommodate the possibility of false belief — a theory of reference determination must provide a principled way of assigning truth-conditions to the subject’s beliefs that allows us to discern which beliefs are true and which are false. But approximate truth is so undemanding a constraint that it can easily be met by different incompatible referential assignments: the totality of your ‘morality’-beliefs, for instance, is approximately true of a whole range of closely related but non-coextensive properties (including the properties specified by competing theories in normative ethics). Thus a simple reliabilist constraint like (R) faces the radical indeterminacy problem highlighted by the darts analogy. By itself this constraint does not provide a principled way of demarcating true from false beliefs.\(^{17}\)

Comparison with Lewis brings out two further issues for Boyd’s account. Unlike Lewis, Boyd takes epistemic coordination with a linguistic community to be relevant to fixing the reference of individual speakers’ use of moral or scientific terms. So there are important further questions for Boyd to address: a full metasemantics must specify just which social connections are relevant and how exactly they affect individuals’ reference. The important point for our purposes, however, is that introducing social factors into the account won’t help resolve the indeterminacy problem. If anything, feeding social information into a reliabilist theory based on (R) will increase the range of indeterminacy. Dispositions to apply moral terms will surely exhibit more variability between individuals than within a given

\(^{17}\) The indeterminacy worry also threatens van Roojen’s attempt to construct a successor theory of reference to that of Boyd. Amending Boyd, van Roojen writes:

\begin{quote}
Roughly, and for nondegenerate cases, a term $m$ refers to a property $p$ just in case there exist epistemically relevant procedures whose tendency is to bring it about, over time, that what is predicated of the term $t$ will be approximately true of $k$ (excuse the blurring of the use-mention distinction).\(^{176}\)
\end{quote}

Since it appeals to truth-approximation, van Roojen’s proposal inherits the problems we have highlighted for (R). To tackle indeterminacy, van Roojen introduces a naturalness constraint into his account. We will consider this type of constraint in §5 below. One additional issue van Roojen must face is how exactly the key notion of ‘epistemic relevance’ should be understood: if epistemic relevance depends on the reliability of certain procedures in promoting true beliefs about the reference, the account would be viciously circular. So epistemic relevance must be characterized independently of truth.
individual. In the case of scientific terms, we might plausibly posit relations of epistemic deference to reduce some of this indeterminacy. It would be very controversial, however, to assume that experts can play a similar role in the moral domain. Even if we do occasionally rely on others’ moral expertise, that does not mean there is a well-defined group of moral experts whose opinions are crucial to settling the reference of moral terms.

The second issue is holism. Both (R) and (M) fix the reference of the subject’s term ‘morally right’ holistically, by appealing to the truth of beliefs involving the subject’s other terms like ‘action’, ‘pain’, ‘stealing’, ‘good’, etc. But ultimately a metasemantic theory must determine the reference of all of the subject’s terms. There are two ways of fleshing out a holist approach. Lewis embraces a global holist metasemantics: reference is assigned to all terms in the subject’s vocabulary in such a way as to maximize truth among the totality of sentences she would accept (Lewis 1984). The alternative would be to provide a two-tiered metasemantics: some basic vocabulary acquires reference independently (e.g. one might appeal to verificationist or teleosemantic mechanisms to fix the reference of “observational” terms) and non-basic terms inherit their reference from basic ones via a holistic constraint like (R) or (M). It’s not clear which strategy Boyd himself would favor. Our argument has proceeded on the assumption that Boyd was a two-tiered theorist. Such an account would require further filling in. But what’s important for our purposes is that indeterminacy worries aren’t lessened by going for global holism instead. On the contrary, global holism gives rise to new indeterminacy problems, beyond the basic one we highlighted in the darts analogy (Quine 1960: ch. 2; Putnam 1980, 1981; Hale and Wright 1999).18

18. Global holism assigns reference by placing constraints on the assignment of truth-values to whole sentences, leaving us free to vary referential assignments ad libitum as long as we respect the basic constraint on truth. One worry is that it is possible to make trade-offs in the referential assignments of subsentential expressions that will preserve exactly the same patterns of truth and falsity for whole sentences. A further worry is that any consistent set of beliefs can be interpreted as perfectly true (provided that we have a large enough domain of referential candidates), so the requirement of truth-maximization will always be trivially satisfiable. To counter these worries, Lewis introduces a further metaphysical naturalness constraint on reference (Lewis 1984). In §5, we’ll consider whether supplementing (R) with such a naturalness constraint might provide a non-arbitrary way of demarcating true from false beliefs.

4. Hooking up with the world
Mere reliability, we’ve suggested, does not provide sufficient constraints on the determination of reference to distinguish true from false beliefs. But a central motivation for Boyd’s account is to show how the reference of both moral and scientific terms is dependent on our causal interaction with features of the world. In this section, we’ll consider whether causal relations can help answer our Fundamental Question: might causal relations be the extra factor that helps secure reference to a sufficiently determinate property?

This idea of causal interaction between language and the world is presumably what Boyd has in mind when he characterizes the reference relation as involving the causal regulation of a term by the property or kind picked out. Boyd also speaks of the accommodation of the use of a term to the joints of nature and of deference to nature.19 The natural way of interpreting these metaphors is that reference is fixed by causal feedback loops between our belief-forming practices with a term t and the local facts about some feature of the world k. Roughly, the idea seems to be that interaction with k causes us to modify our belief-forming mechanisms in such a way that most of our t-beliefs become approximately true of the local k-facts. Thus, over time, our belief-forming mechanisms become tuned to facts in a particular domain: our ‘dog’ beliefs tend to vary counterfactually with changes in the local facts about dogs, and they are largely true if interpreted

19. For accommodation, see for instance (Boyd 1993: §11):
The dialectical process of accommodation—the introduction of linguistic terminology, or the modification of current usage so that general terms come to afford epistemic access to causally important features of the world—is an essential component of reference. For deference to nature, see (Boyd 1991: 139).
as about dogs. What fixes the reference of *t*, Boyd suggests, are causal feedback relations that set up a systematic isomorphism between our set of *t*-beliefs and some set of local *k*-facts.\(^{20}\)

This feature of Boyd’s view—his insistence that the history of causal feedback from the world is central to the determination of reference—is what opens him up to the Moral Twin Earth objection. The objection is intended to highlight our indifference to causal interaction with features of the environment when deciding whether a foreign community uses terms like ‘morally right’ with the same meaning as ourselves. As we noted, purely nomic reliability-based accounts of reference determination are insensitive to the causal history of use of a term; so such theories won’t lead to different referential assignments on Earth and Twin Earth for ‘morally right’ (or for ‘water’). But a theory of reference that invokes historical facts about how the linguistic practice has been causally shaped by interacting with features of the local environment will vindicate the intuition that different environmental histories could lead to different assignments for ‘water’ or ‘morally right’ even when the representational practices are exactly similar.

For many metaethicists, Moral Twin Earth intuitions make any appeal to causal feedback in explaining how the reference of moral terms is fixed highly implausible.\(^{21}\) The point we want to emphasize here, however, is that simply appealing to causal regulation will not address the indeterminacy problem. The worry is that there are just too many properties that have causally influenced our linguistic practices with particular terms.

To see the force of this worry, it helps to focus on a less controversial case like ‘dog’. Some of your ‘dog’-beliefs are acquired on the basis of observing Fido. This means the property of being Fido is one of the properties that causally shape your dispositions to form ‘dog’-beliefs. Ditto for the properties of being a pet, being a hunting companion, having a low social status, being a midsize physical object, being a bloodhound, being a carnivore, being a member of *canis lupus familiaris*, being a member of the canine family, exhibiting a certain doggy perceptual Gestalt, being smelly, being frightening, etc. Even the property of being a cat on a dark night will have a systematic causal impact on your ‘dog’ usage if you’re disposed to mistake cats for dogs. The result, then, is that indefinitely many distinct properties can be said to causally regulate our belief-forming mechanisms for ‘dog’.

The moral is that having a causal impact on your ‘dog’-beliefs (or ‘morally right’-beliefs) does not suffice to single out any determinate reference. A causal theory of reference determination must specify interpretive intuitions (which were carefully manipulated in Putnam’s article), but rather the underlying epistemological interests that we presume his chemically ignorant subjects Oscar and Twin Oscar share with us and that guide rational inquiry into the real nature of the stuff they call ‘water’.

Putnam’s real point is that there is at least one sense of ‘water’ where we take our current understanding to be corrigible in the light of empirical inquiry. Could the same be said for our term ‘morally right’—is there a sense in which we take our current understanding of the nature of the property we’re talking about to be corrigible by empirical inquiry? Empirical corrigibility is more controversial for morality than for water. But if we do hold ourselves to be empirically corrigible, then this very fact will commit us to the possibility of intrinsic duplicates on a MTE who represent a different property with their homophonic term ‘morally right’. Moreover, this commitment to empirical corrigibility would then provide a sound theoretical rationale for overriding our initial seat-of-the-pants judgment that we mean the same thing as our MTE counterparts, just as we override Putnam’s space traveler’s initial seat-of-the-pants judgment that the people on TE are talking about the same topic as him when they use the term ‘water’.

20. Boyd’s notion of *regulation by a property* is not meant to suggest that we directly interact with properties themselves (which may be abstract objects or platonic forms). We causally interact with particular dogs, not with the property of doghood. Still, a regulating property must be *causally relevant* to explaining why the effect came about. At a minimum, the effect must be counterfactually sensitive to the instantiation of the regulating property, if it’s to distinguish the actual reference of a term from contingently co-extensive properties. For instance, the property of being a dog is actually co-extensive with the property of being Cesar’s favorite kind of animal; but being Cesar’s favorite doesn’t causally regulate English speakers’ practice with the term ‘dog’. Being beloved by Cesar is causally irrelevant to shaping our dispositions to form beliefs about dogs. Sayre-McCord (1997: 270) seems to float this causal relevance requirement in his proposed variation on Boyd’s causal regulation model.

21. We believe the psychological data driving the MTE argument are less probative than is commonly thought in metaethics. In a nutshell: doing the real work in Putnam’s original Twin Earth thought experiment are not our naïve
exactly which property, from among all the properties that have had a causal impact on your linguistic practice with a term, is the reference of the term. A causal regulation theory of reference, moreover, must provide some independent, non-semantic way of singling out the relevant reference-fixing relations: it would be viciously circular to specify the relevant causal relations simply as those that secure truth and reference for the terms we’re seeking to interpret. So Boyd cannot, past interaction is the most resilient cause of t-versa. So perhaps it represents k because k is the most resilient cause of your current stock of t-beliefs.

Metaethicists may be tempted to appeal to such naturalistic causal theories to help resolve indeterminacy and circularity worries. But this strategy faces serious obstacles. Although naturalistic theories aim to resolve indeterminacy, it’s far from clear that they actually do provide a satisfactory solution. Teleosemantics and asymmetric dependence are the two most popular and systematic attempts in the philosophy of mind and language to specify a naturalistic causal relation that might constitute the reference relation. But both of these approaches have been widely criticized as (a) too vague to single out a determinate property as the reference, and (b) imposing substantive causal conditions that would single out the intuitively wrong reference in many cases.

Moreover, there is reason to think that both of these worries about naturalistic causal theories will be exacerbated for non-convergentist normative realists. As we noted in section 2, normative realists want to secure more determinacy for morality than is required for many everyday or even scientific domains. In biology, for instance, it’s plausible that there’s no determinate fact of the matter about whether an individual creature is a member of the species *homo sapiens* in cases where empirically and theoretically sophisticated speakers would persistently disagree. But morality seems different: even in cases of persistent moral disagreement among informed and sophisticated speakers, we assume there may be a fact of the matter about what’s morally right. Non-convergentist normative realists aim to vindicate this intuition — indeed, that’s the primary reason why they’re interested in causal theories of reference in the first place. Like the indeterminacy problem, the wrong reference problem is also exacerbated in the moral domain. Normative realists generally recognize a distinction between moral properties and sociological properties — like conventional morality — that causally explain social interactions within a given community. For such normative realists, then, it’s a constraint on an adequate theory of reference that it not assign sociological properties as the reference of moral terms. But naturalistic causal theories have

22. Teleosemantics has been developed in a number of different ways, and may invoke causal selection processes like learning or cultural selection rather than simply biological selection (see e.g. Millikan 1984; Dretske 1988; Papineau 1987; Neander 1995).

23. On Fodor’s account, asymmetric dependence is a purely synchronic relation among causal laws, but the notion could be adapted to choose from among past causes (Fodor 1987, 1990).

24. For helpful surveys of the difficulties faced by naturalistic accounts, see (Loewer 1999; Neander 2006). There are a host of more specific indeterminacy challenges, which arise in different ways for different naturalistic theories: e.g. the problem of error, the ‘qua’ problem, the distal problem, the disjunction problem.
difficulty in securing this result. On such theories the reference of a term is whichever property plays a certain generic causal role in reinforcing the subject’s past use of the target expression. Sociological properties seem ideal candidates to play this causal role for moral predicates. Even if genuinely normative properties can figure in some true causal claims, sociological properties seem more important as external factors that causally shape and reinforce moral beliefs. So there is a very significant risk that naturalistic theories of reference will privilege conventional over non-conventional morality as the reference of moral predicates. The problem is not just that this would result in a highly revisionary theory of morality; the problem is that it would undercut the very fallibility that provides the rationale for non-convergentist normative realism in the first place.

5. What the world can do for you
A different strategy for reducing indeterminacy is to appeal to facts about the referential candidates themselves, rather than to causal relations linking referential candidates to our representational practices. Perhaps there are objective metaphysical facts that make some objects, kinds, and properties better, more natural candidates for reference than others. We could then appeal to the relative naturalness of referential candidates to reduce the indeterminacy of reference. Consider the darts analogy once again. The task of identifying which target Joe is aiming at would be much easier if Charity knew in advance that the target must be a perfect 18-inch circle placed at one of six different locations. Given this rich information about possible targets, Charity’s initial presumption that Joe is a reliable dart thrower may suffice to single out one of the eligible candidates as the target Joe is aiming at.

David Lewis proposed such an objective naturalness constraint as a way of avoiding radical indeterminacy of reference:

Among all the countless things and classes that there are, most are miscellaneous, gerrymandered, ill-demarcated.

Normative realism: co-reference without convergence?

Only an elite minority are carved at the joints, so that their boundaries are established by the objective sameness and difference in nature. Only these elite things and classes are eligible to serve as referents. The world—any world—has the makings of many interpretations that satisfy many theories; but most of these interpretations are disqualified because they employ ineligible referents. (Lewis 1984: 227)

Lewis takes the metaphysical joints of nature to be the objects and properties characterized by fundamental microphysics. But obviously most people don’t talk about quarks, so the eligibility constraint on referential candidates must be loosened if it’s to provide a plausible interpretation of our everyday language.

Indeed, physics discovers which things and classes are the most elite of all; but others are elite also, though to a lesser degree. The less elite are so because they are connected to the most elite by chains of definability. Long chains, by the time we reach the moderately elite classes of cats and pencils and puddles; but the chains required to reach the utterly ineligible would be far longer still. (1984: 228)

Lewis’s strategy is to posit an objective hierarchy of metaphysical naturalness, which ranks objects and properties according to how many logical connectives it would take to define them in the language of microphysics. This naturalness constraint rules out gerrymandered referential assignments by requiring us to prefer more natural referential candidates to less natural ones. Lewis’s ‘global descriptivist’ theory of reference combines this naturalness constraint with a truth maximization constraint. Roughly, the idea is that we should favor referential assignments that maximize the truth of the subject’s beliefs, while also maximizing the naturalness of the referential candidates.
Lewis introduces his metaphysical hierarchy of naturalness to address the radical indeterminacy worries highlighted by Putnam’s paradox. But a naturalness constraint may also help to address the more mundane type of indeterminacy we highlighted in the previous two sections. Our challenge was to specify a constraint on referential assignment that would allow us to draw a principled line between true and false beliefs. If we give up Ideal Accessibility, the normative realist needs an account of the further factors that fix reference to a sufficiently determinate normative property. Perhaps adding a metaphysical naturalness constraint on referential assignments can help to answer this challenge.

The problem, however, is that Lewis’s proposal provides an implausible account of less-than-perfectly natural properties. In particular, the idea that the reference of ‘morally right’ could depend on how many logical connectives it takes to define a referential candidate in the language of microphysics seems incredible: intuitively, moral rightness is not beholden to microphysics in this way. In response to this type of worry, Mark van Roojen has suggested that we should modify Lewis’s constraint: instead of taking definability in microphysics as the standard of naturalness, we can appeal to a “discipline-relative” notion of naturalness:

Naturalness should be seen as discipline-relative. The kinds of properties which are more natural for the purposes of physics may not be the same as those which are more natural for the purposes of biology. The more eligible semantic values for one’s terms when engaged in the former may or may not be the same as the more eligible semantic values for one’s terms when one is engaged in the latter. Probably the naturalness of a property or kind for a given discipline is relative to the subject matter and purposes of the discipline. [...] Tables seem perfectly natural for anthropological purposes, but not for the purposes of physics. (van Roojen 2006: 180–1)

Van Roojen combines this discipline-relative naturalness constraint with a reliability constraint of the type we discussed in §3. Adding his naturalness constraint to a Boyd-inspired reliability constraint, van Roojen suggests, will suffice to specify a determinate reference for moral terms without appealing to causal regulation by moral properties (2006: 181–2).

We think that appealing to discipline-relative naturalness is much more problematic than it may seem at first blush. According to Lewis, naturalness is an objective metaphysical feature of the world as it is in itself, independently of our subjective perspective. The idea of discipline-relative naturalness, in contrast, seems relational. Whether a property counts as a natural relative to a discipline is not a matter of its objective metaphysical character, but of how it “fits” with the set of beliefs, background interests, epistemic standards, explanatory styles, etc. associated with our linguistic practice with particular terms: which properties count as maximally natural will be different depending on whether the target term plays a central role in chemistry, biology, etc.

25. Putnam argued that a truth maximization constraint will be trivially satisfiable: given any system of beliefs, there will always be a way of assigning extensions to the subject’s words that will make all the beliefs come out true — in fact there will be indefinitely many different assignments that will do the trick (Putnam 1977, 1980, 1981). Although many of these assignments may strike us as perverse, Putnam argues that no aspect of the subject’s understanding can distinguish between these different referential assignments: the subject’s understanding will just add more beliefs to the set whose truth we are seeking to maximize, and we will always be able to find many different referential assignments that will make the entire augmented set true. Lewis’s solution is to appeal to wholly objective metaphysical facts about the world itself, rather than facts about the subject’s mental states, to rule out perverse referential assignments. It’s thus crucial to Lewis’s solution that naturalness be an objective metaphysical fact about the world, rather than an aspect of our subjective perspective on the world.

26. For van Roojen’s reliability constraint, see above fn. 17. Van Roojen rejects the idea that causal regulation plays a privileged role in determining the reference of moral terms because this condition opens the door to the Moral Twin Earth objection.
economics, musicology, history, morality, cooking, or cleaning. Such a relational notion of naturalness does not reflect the metaphysical nature of the world in itself, but rather the world as seen from the perspective of certain projects and practices.

But this relational naturalness cannot play the same constraining role in a theory of reference as Lewis’s objective metaphysical joints of nature. Once again, the point can be brought out with the darts analogy. In giving up an objective metaphysical naturalness, we give up the idea that Charity can know in advance that Joe’s target must be one of six perfect 18-inch circles (i.e. one of a limited set of objective “joints of nature”). Instead, Charity must identify the shape of Joe’s target on the basis of facts about Joe’s dart-throwing practices (e.g. in one practice session he throws in such-and-such a pattern, in another practice session he throws in a different pattern). It should be clear that asking Charity to decide which candidate shapes are “most natural” relative to these two practices is equivalent to our original question of asking her to decide what shape Joe is aiming at on a given occasion. Unlike positing metaphysical naturalness constraints in the world, positing practice-relative naturalness leaves us right back where we started: to say that a referential candidate counts as “natural relative to our practice with a word” is just to say that it is plausible as a referential assignment for that word. The upshot, then, is that this notion of discipline-relative naturalness does not provide a substantive constraint that we can invoke in order to explain reference: discipline-relative naturalness is explained in terms of our theory of reference, not vice versa. 27

27. It’s worth pointing out that Boyd himself rejects Lewisian objective metaphysical natural kinds in favor of a discipline- or interest-relative notion of natural kinds (1993: 531–2). In his philosophy of science, Boyd analyzes natural kinds as homeostatic property clusters: i.e. sets of properties that have a stable tendency to be instantiated together in virtue of self-reinforcing causal mechanisms. On this account, the ‘gemological’ kind of jade (roughly, being a green stone apt for certain decorative roles and apt for certain techniques of carving) will count as a natural kind, a homeostatic property cluster, even though it’s not metaphysically important “from the point of view of the universe”. As this example makes clear, there are potentially infinitely many different ways of carving out discipline-relative homeostatic property clusters. So Boyd appeals to facts about our social linguistic practice to single out one such cluster as the reference of a particular term. In this sense, Boyd characterizes natural kinds as ‘social constructions’:

Biological taxonomists sometimes speak of the ‘erection’ of higher taxa, treating such taxa as, in a sense, human constructions. They are right—and the same thing is true of natural kinds in general. Natural kinds are social constructions: they are the workmanship of women and men. (Boyd 2010: 220)

It should come as no surprise, then, that Boyd does not invoke naturalness as an independent constraint on reference determination. In the passage immediately preceding this quote, Boyd explicitly explains discipline-relative naturalness in terms of his theory of reference, not vice versa.

28. Recent advocates of multipolar reference magnets include (McPherson 2012; Sider 2011).
other properties supervene; and microphysical properties seem to play this role in the actual world (Lewis 1983: 356–7). Proponents of multipolar naturalness need a replacement story about what it takes to be a perfectly natural property: what is the principle for deciding which properties are part of the basic furniture of the world? If moral properties belong in this class, presumably it’s not in virtue of playing a crucial, irreducible causal role — as the properties posited by the special sciences seem to do. So why exactly do moral properties deserve to be considered basic joints of nature any more than the properties of being a chair or being cozy? (Or perhaps those properties are themselves basic joints of nature — without a demarcation principle it’s impossible to say.)

Moreover, why suppose that the basic joints of nature are sufficiently sparse that they can help resolve the semantic indeterminacy problem for moral terms? Once we allow that there are many different kinds of basic metaphysical joints, there is a risk that there will be too many: if conventional morality, preference maximization, maximization of human flourishing, and conforming with the categorical imperative are all basic joints of nature, then being a joint of nature won’t help to reduce semantic indeterminacy of moral terms to acceptable levels. If normative realists hope to use metaphysical naturalness to explain how the reference of moral properties is fixed, they must have plausible, independently motivated answers to these metaphysical questions.

On the metasemantic front, we need a coherent theory that can explain how basic metaphysical joints of nature affect the reference of our terms. A number of questions must be addressed. First, why are mind-independent joints of nature relevant to determining the precise reference of our terms? It’s not obvious that a purely metaphysical constraint will deliver plausible referential verdicts, especially for terms that are beholden to anthropocentric concerns. It will be of little comfort to the non-convergentist realist if the metasemantics solves the disagreement problem by construing speakers as picking out properties that are important to the universe but irrelevant to speakers themselves. Second, how exactly do reference-magnets affect the semantic interpretation of an individual’s language as a whole? The metaphor of a “reference magnet” can be misleading in this respect, since it invites us to imagine a particular word attracted by a particular magnet. But reference magnetism isn’t a one-one relation, it’s a many-many relation: we have lots of words, and external reference magnets operate on all of them at once. So a proponent of a metaphysical constraint on reference determination must explain exactly how objective joints of nature affect the interpretation of different words within a subject’s idiolect differently: e.g. (1) why do some terms pick out basic joints of nature, while others do not? And (2) why do these terms pick out those joints rather than some other joints?

Prima facie, Lewis’s account seems to meet this challenge of connecting metaphysics with metasemantic intuitions. Lewis holds that natural properties are objectively important from the point of view of the universe. Moreover, it’s intuitively plausible that some notion of importance plays a role in determining the reference of our terms: we take our words to pick out genuinely interesting and important features of the world. But the notion of interest behind commonsense intuition is an anthropocentric one, which is very different from Lewis’s metaphysical notions of naturalness (e.g. definability from microphysics, or multipolar fundamentality). This conflations of metaphysical and metasemantic importance, we suspect, explains the initial plausibility of a Lewisian naturalness constraint: if metaphysical naturalness really could capture our commonsense verdicts about which referential candidates are most relevant, then the metaphysical constraint would be metasemantically plausible. See (Williams 2010) for a related observation.

On the multipolar approach, for instance, many basic joints of nature will be extensionally quite close: e.g. the normative property of being morally right and the sociological property of being conventionally right have similar extensions across different possible worlds. If moral properties are perfectly natural joints of nature, it’s plausible that sociological properties will be perfectly natural as well. So what is it about competent speakers’ understanding or use of the predicate ‘is morally right’ that guarantees that they all lock onto moral rightness? Why won’t at least some of them lock onto conventional rightness instead? After all, many individuals’ dispositions to apply moral predicates to cases more closely track facts about conventional morality than they track...
If one is primarily focused on normative expressions in particular, it’s easy to overlook a central constraint on an adequate answer to such metasemantic questions. The very same metasemantical principles that explain why your use of ‘is morally right’ picks out this perfectly natural property must also be capable of explaining how other words like ‘is cozy’ or ‘is a chair’ or ‘is grue’ pick out those less-than-perfectly-natural properties. The answers we give for moral predicates must fit into a general theory of reference determination: if your metasemantical theory delivers the right answers for moral terms but the wrong answers for some other terms, then that theory cannot be correct.\textsuperscript{32}

Lewis provides a systematic way of answering questions such as (1) and (2). On the one hand, the subject’s contribution to the determination of reference is the same for all terms: an acceptable interpretation must maximize the truth (or rationality) of the attitudes the subject would express in her language. On the other hand, the world provides a uniform hierarchy of naturalness: an acceptable interpretation must maximize naturalness in the reference assigned to the terms of the subject’s language. By trading off “fit” against “eligibility” for the subject’s whole language, Lewis hopes to single out a determinate interpretation for all terms in the language. But a multipolar metaphysics cannot be directly inserted into Lewis’s metasemantical theory. If physics is no more natural than biology, sociology or ethics, then there is no uniform hierarchy of relative naturalness; and without such a hierarchy, it’s not clear how to assess the relative naturalness of imperfectly natural properties. So moving to a multipolar metaphysics will erase significant structure from Lewis’s account. What metasemantic machinery will take up the slack?

\begin{itemize}
\item the normative properties about which we are supposed to be fallible and irreconcilably divided (according to the non-convergentist normative realist).
\end{itemize}

\textsuperscript{32.} For holistic theories like Lewis’s, the generality constraint is even stronger: the reference to any expression in the language depends on the correct assignment for all the expressions in the subject’s language. So an assignment that yields incorrect semantic values for non-moral terms will eo ipso invalidate the semantic assignments for moral terms.

Providing the requisite metasemantic machinery will be an especially delicate task for non-convergentist normative realists. For non-convergentists, the main role of the naturalness constraint is to help secure co-reference despite variation in moral understanding. So reference magnets must be \textit{strong enough} to override differences in understanding among speakers. But making perfectly natural properties into sufficiently strong reference magnets threatens to make it impossible to refer to any less-than-perfectly-natural properties in the vicinity of moral rightness. Say you introduce names for the properties picked out by two non-coextensive moral theories, one Kantian, one utilitarian. A strong naturalness constraint may lock both predicates (‘is right\textsubscript{k}’ and ‘is right\textsubscript{u}’) onto the very same perfectly natural property rather than two distinct less-than-perfectly-natural properties. The problem here is the inverse of that faced by descriptivist theories of reference determination: whereas descriptivism threatens to make genuine moral disagreement impossible, reference magnetism threatens to make it impossible to represent slightly different properties. Simply positing perfectly natural moral properties does nothing to address these problems. Whether there is a satisfactory account of the strength of reference magnets will depend entirely on the details of the metasemantic theory: which aspects of subjective practice are taken into account, and what exactly the naturalness constraint consists in.

To be clear: we are not claiming that the objective joints of nature have no role to play in a theory of reference determination. Our point is that there are important explanatory burdens one must undertake before this approach is assessable: (i) one must explain the metaphysics and (ii) one must say exactly how the metaphysical facts help determine referential assignments. Modifying Lewis’s approach to accommodate many distinct types of objective naturalness is not a minor modification: it involves unexplained metaphysical commitments and removes crucial structure from Lewis’s account of reference determination.
6. Conclusion

Taking Boyd’s theory as our starting point, we’ve explored different ways of explaining how reference is determined if one wants to vindicate traditional normative realism while avoiding any commitment to convergence in speakers’ ideal judgments. The core reliabilist proposal was that the mechanisms that underwrite everyday normative knowledge also suffice to fix the reference of normative predicates. However, a simple reliabilist metasemantics leads to radical referential indeterminacy that clashes with the fallibilist commitments of non-convergentist realists. We then considered two standard ways of supplementing this core proposal: (a) by appealing to causal relations to the world or (b) by appealing to metaphysical “joints of nature”. Both proposals face significant objections as general metasemantic theories. But the difficulties are compounded when the goal is to vindicate the distinctive theoretical commitments of normative realism: e.g. ensuring that normative terms do not pick out mere sociological properties, vindicating commonsense intuitions about the co-reference of normative terms, and securing enough determinacy of reference to adjudicate rationally irresolvable normative disagreements. A clear view of these challenges does not encourage optimism that standard externalist metasemantics can succeed in vindicating traditional normative realism.

This conclusion is significant for non-convergentist normative realists because the externalist metasemantic strategies we have been exploring provide the best worked-out answers (by far) to our Fundamental Question: If subjects’ ideal, fully informed verdicts do not determine the reference of normative terms, what other factors will do the job? It’s not obvious that there are other viable metasemantic options for a non-convergentist realist. The constraints on an acceptable metasemantics for the non-convergentist realist include: (i) avoiding circularity, (ii) securing a plausible level of referential determinacy, (iii) securing plausible referential assignments, (iv) integrating the metasemantics for normative terms within a general metasemantics, and (v) avoiding any commitment to Ideal Accessibility (IA). In our view, no non-convergentist metasemantic theory currently on offer meets these constraints.\(^{33}\)

The lesson we draw is that normative realists should give the Ideal Accessibility constraint (IA) a second chance. A metasemantics that respects (IA) will avoid many of the difficulties externalist metasemantics face in securing plausible referential assignments for normative predicates. For instance, theories that are grounded in subjects’ ideal verdicts will avoid the ‘wrong reference’ problems highlighted by constraint (iii). Your normative predicate ‘is morally right’ won’t end up representing conventional rightness simply in virtue of the property’s causal role in shaping your normative beliefs; and non-moral predicates won’t end up representing moral rightness simply in virtue of the importance of this property from the point of view of the universe. The determinacy constraint (ii) also becomes much less challenging. By assigning reference on the basis of an individual’s own ideal reflective dispositions, (IA) ensures that semantic indeterminacy will remain within intuitively acceptable bounds: penumbral cases will reflect the epistemic commitments of the subject herself, not the independent commitments of an externalist metasemantic theory. And since ideal verdict methodology can be applied for any type of expression, a metasemantics based on (IA) promises to integrate the

---

\(^{33}\) Other non-convergentist metasemantic theories have been proposed, but we are not sanguine about their prospects. According to Ralph Wedgwood (2001, 2007), for instance, competent speakers master conceptual roles (‘rules of rationality’ governing the use of a term) that suffice to fix reference of normative terms independently of speakers’ ideal verdicts about the extension. However, we believe Wedgwood’s account of reference determination faces a dilemma: either the account locks normative terms onto the wrong reference, or it relies on question-begging information in specifying the conceptual role mastered by competent subjects (Schroeter and Schroeter 2003). If so, constraints (i) or (iii) will be violated. Other theorists have proposed a return to Putnam’s and Kripke’s ‘causal chain’ approach to reference fixing for normative terms (Copp 2000; Brink 2001). See fn. 16 for our objections to this approach. If we’re right, the ‘causal chain’ approach is at risk of violating constraints (ii) and (iii).
account of normative terms smoothly within a general theory for the subject’s language as a whole.

The primary objection to (IA) in the eyes of many normative realists is that convergence in ideal normative judgment seems unlikely. This pessimism about ideal convergence, we believe, is partly fueled by an overly individualistic approach to metasemantics: contemporary metaethics tends to ignore the role of the linguistic community in helping determine the reference of an individual’s words and thoughts. We've argued elsewhere that the primary unit for semantic interpretation is not an individual speaker's current use of an expression considered in isolation, but rather the communal representational practice in which that individual participates. 34

Individual speakers’ epistemic commitment to holding their own understanding of normative properties accountable to such a communal representational practice, we believe, provides new resources for vindicating ideal convergence in normative judgments. But that is a topic for another occasion. 35

References


34. For the application of this anti-individualist approach to metaethics, see (Schroeter and Schroeter 2009, 2013). For a more general defense and elaboration of an anti-individualist approach to meanings and concept individuation, see (Schroeter 2008, 2012). The original anti-individualist thesis that a subject's social environment affects the reference and content of her thoughts is due to (Burge 1979, 1986).

35. This paper originated in a series of discussions with Dick Boyd while he was a Miegunyah Distinguished Fellow at the University of Melbourne. We’d like to thank Dick for that very stimulating exchange, from which we learned a great deal. Thanks also to David Copp, Ant Eagle, Krista Lawlor, Tristram McPherson, Connie Rosati, Michael Smith, Mark van Roojen, and two anonymous referees for this journal for helpful comments on earlier drafts of the paper.


