Realism and Fundamentality in Ethics and Elsewhere

by

William R. Dunaway

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2013

Doctoral Committee:

Professor Allan F. Gibbard, Chair
Assistant Professor Ezra Russell Keshet
Assistant Professor David Manley
Professor Peter A. Railton
Professor Mark Schroeder, University of Southern California

Dedication

For Rachel, Gabe, Ellie, and Jamie–
who have showered me with love.

Acknowledgments

I have always found it difficult to limit my research topics in philosophy to a single sub-discipline, and this dissertation is no exception. It brings together ideas from ethics, metaphysics, and philosophy of language, among other areas. While I am sure that in present form it leaves much to be desired, a wide-ranging project like this would not even be passable without my advisors and colleagues who have had the patience and knowledge to work through these ideas alongside me. They have graciously taught me, encouraged me, and been overall philosophical role models for me throughout my time as a graduate student (and, in some cases, as an undergraduate too). While it is difficult to name all of their contributions in a short space like this one, I would at the very least like to give an all-too-brief description of who these people are and how this dissertation came to be.

Each of the four chapters in this dissertation is a product of Mark Schroeder's engaging and rigorous introductions to topics in metaethics. Chapter 3, along with my earlier paper "Minimalist Semantics in Meta-ethical Expressivism", arose out of discussion of the "problem of creeping minimalism" in Mark's seminar on Expressivism in 2006. Chapters 1, 2 and 4 are products of topics Mark presented in a 2007 on realism and reduction. Mark's papers "Realism and Reduction: the Quest for Robustness" and "Supervenience Arguments under Relaxed Assumptions" (with Johannes V. Schmitt) also provided valuable resources. Mark's introductions and work on these topics not only sparked my interest in these issues; I would not have pursued them further if it were not for his subsequent encouragement and direction as I formed my own thoughts on them. It is not an exaggeration to say that this dissertation would not exist if it were not for Mark Schroeder.

Allan Gibbard gave his characteristic clear and careful feedback as these ideas came into present form, meeting weekly for the entirety of Fall term of 2010. While this dissertation takes a robustly metaphysical starting point that in some ways could not be further from Allan's preferred theoretical home, he graciously worked to understand what I was trying to say, even when I wasn't saying it very well. His questions forced me to think through foundational issues, made me find more articulate formulations of these ideas, and helped the formulations to be amenable to a wider range of philosophers. Allan also taught me to appreciate the power of the Expressivist approach to language, both during our meetings and in his 2011 seminar on what would become the book *Meaning and Normativity*. I hope to have done justice to his thinking in Chapter 3, and to have made progress on understanding what is distinctive about the important ideas he has developed over the last two decades (and counting).

Peter Railton, whom I have long admired for his broad and creative philosophizing, gave substantial feedback on each of the chapters in this dissertation. His expertise across many sub-disciplines of philosophy made him the perfect interlocutor as I wrote. Not only was the content of his comments invaluable; they were also given in the spirit of genuine philosophical curiosity. I found this spirit to be most refreshing, and hope that some of it has found its way into this dissertation.

From David Manley I learned swaths of metaphysics through hours of conversation (when he, as a tenure-track assistant professor, likely had much more pressing work to be doing). But more importantly I also had the opportunity to witness firsthand someone who doesn't take the presuppositions of any philosophical debate for granted, who is confident that there is always an intellectual reward in working through the options and arguments on most any issue, and who refuses to to allow his use of technical jargon to lose touch with what is substantive and important. I hope to have successfully emulated, to some small degree, each of these traits in this dissertation. Most of all, however, David has been a great philosophical friend, and I am very fortunate to have been around him during my time as a graduate student.

I owe a significant intellectual and personal debt to many others as well, and any attempt at a list here is bound at best to be a partial list. Among those whom I have learned from and enjoyed friendships with are: Gordon Belot, Annette Bryson, Sarah Buss, Steven Campbell, Nate Charlow, Jeremy Dickinson, Matt Evans, Daniel Fogal, Stephen Finaly, J. Dmitri Gallow, John Hawthorne, Jim Joyce, Ezra Keshet, Jeff King, Jason Konek, Maria Lasonen-Aarnio, Kathryn Lindeman, Todd Long, Errol Lord, Paul Miklowitz, Sarah Moss, David Plunkett, Lewis Powell, Scott Soames, Jake Ross, Johannes V. Schmitt, Alex Silk, Dan Singer, Justin Snedegar, Julia Staffel, Rohan Sud, Jim van Cleve, Ken Walker, David Wiens, and Nina Windgätter. There are many others as well, and I am sorry to have left them out.

I owe a special thanks to my family—my sister Amelia, my mother Nancy, my father Bill, and in-laws Bill and Rogene—for their continual support. And finally, the biggest thanks of all go to wife Rachel, who brings never-ending joy and optimism to our time together, and to our children Gabe, Ellie, and Jamie. Their enthusiasm and smiles are a lesson for me, more important than any to be found in a book or classroom.

Preface

This is a dissertation about how the notion of metaphysical fundamentality can contribute to several outstanding problems in metaethics and elsewhere. The notion of what is fundamental opened numerous avenues for fruitful research in metaphysics itself. I hope that the chapters in this dissertation together constitute a kind of proof-by-example that similar progress can be made by attending to the relevance of the notion to familiar metaethical issues. In particular, I hope to show that it affords a philosophically rich understanding of realism and reference in ethics, the distinctive features of quasi-realist Expressivism, and the commitments of Moorean non-naturalism. The payoff of developing these applications is not limited to metaethics, however: we also learn more about the structure of the underlying notion of fundamentality.

It is not uncontroversial that metaphysical fundamentality is of use to theorizing in metaphysics itself—some have even doubted its intelligibility. I won't aim to directly defend the notion in what follows, and will simply assume it as a starting point. The conclusions here should then be read conditionally, contingent on fundamentality being a legitimate resource in the first place. But these conclusions can also provide something of an indirect defense of the notion. This is for two reasons: first, they describe concrete cases where fundamentality plays a crucial role in solving problems that have proved resistant to treatment in other terms. Second, those who doubt that they can *understand* what metaphysical fundamentality is might take the roles I outline as helpful for coming to a working grasp of the notion. Insofar as we have an adequate grasp on realism, reference, and non-naturalism, they can provide an entry way to understanding—if, as I claim, they are best characterized in terms of fundamentality.

Chapter 1 argues that the distinction between realism and irrealism, both in ethics and in other domains, is best understood in terms of a distinction in fundamentality. The crucial point is that fundamentality has a structural feature that other proposed characterizations of realism lack: it can come in *degrees*. I claim that this is required for both crafting a unified treatment of realism across domains, and accounting for to the gradability of judgments about realism within a single domain.

Chapter 2 brings this understanding of realism to bear on a prominent argument against realism in some of its forms. The argument in question is the Moral Twin Earth argument, which alleges that realists cannot provide an adequate theory of what it is in virtue of which moral terms acquire their referents. Here I make use of an auxiliary claim that is found in the work of David Lewis, Theodore Sider and others: highly fundamental properties are reference magnets. I claim that an understanding of realism on which moral properties are highly fundamental and are therefore reference magnets allows realists to provide an independently motivated account of what it is in virtue of which moral terms acquire their referents. The Moral Twin Earth argument fails to appreciate all of the resources available to a realist theory of reference.

Chapter 3 considers a difficult case for distinguishing realism from other metaethical theories. *Expressivist* theories of normative language have proven adept at giving an account of why claims that are traditionally associated with realism are true. This is *quasi-realist* Expressivism. It is intuitive that this version of Expressivism is not a species of realism, but given the quasi-realist maneuvers, it is remarkably difficult to say why this is so. Here again fundamentality comes to the rescue: the realist, for reasons familiar from Chapter 2, needs the fundamentality of moral properties to explain why moral terms refer to those properties. Quasi-realists can accept the same first-order claims about reference. But by explaining these claims in the distinctively Expressivist fashion, I argue, the quasi-realist thereby takes on commitments for the fundamentality of the moral that differ from those of the realist. The difference between realism and quasi-realism is a difference in fundamentality.

Chapter 4 closes by applying these ideas in a limited defense of the traditional non-naturalist view of the normative. Frank Jackson has argued that the falsity of non-naturalism follows from the fact that the normative supervenes on the natural, plus other minimal auxiliary assumptions. I construct a *reductio* of these auxiliary assumptions, and then argue that the form of the *reductio* makes it very natural to reject the auxiliary assumptions by understanding non-naturalism as the view that the normative is fundamental.

These are all ways in which metaphysical fundamentality has something to offer to problems that are primarily metaethical in nature. But what emerges from the attempt to get traction on these problems with the notion of fundamentality is that it must have some very specific features, not all of which have been appreciated in the literature on fundamentality. Chief among these is the naturalness (in the present context of investigation) of a *relative* notion of fundamentality, which allows us to talk about what is *more* fundamental without being *fully* fundamental, and which is not easily definable in terms of the absolute, binary notion of the fully fundamental. While my focus in these chapters is primarily concerned with issues in metaethics, the reasons for utilizing a notion of primitive notion of relative fundamentality will quite naturally be applicable to questions of realism and reference outside of metaethics as well.

Table of Contents

Chapter

List of Figures

Chapter 1: The Metaphysical Conception of Realism

1.1 Preliminaries

'Realism' is a word that is used frequently to describe views in philosophy, and often its application in a particular context is unproblematic. For instance, most will not hesitate to label Berkeley's idealism an irrealist view of material objects, to be contrasted with our ordinary, pre-theoretic realist view. Or again, the realist about the unobservable posits of our scientific theories is easily distinguished from the irrealist instrumentalist. And in ethics, most will not hesitate to label Expressivists and Subjectivists as irrealists, distinguishing them from the realist Moorean non-naturalist.

These are just a few examples. The point I want to make about them now is just that 'realism' is a philosophical term that can be applied across a wide range of domains, with a fairly intuitive sense. (This is not to say there are no hard cases.) We can then pose an interesting question: is there a natural, joint-cutting kind that all of these uses of the term pick out? The alternative is that either these uses are not univocal across domains (that is, we use 'realism' to talk about something different when applied to material objects and theoretical posits), or that there is some gerrymandered, not-very-interesting property that all of these uses pick out. I have no direct argument for or against any of these options. But an indirect approach is available. We can proceed by asking what the best candidate characterizations of the kind underlying our talk of realism *are*, and asking whether any of them offer *plausible* accounts. If our theorizing turns up a plausible candidate, then there is good evidence that there is a natural kind our term 'realism' picks out. If no such candidate is forthcoming, we then need to revisit the alternatives.[1]

Before proceeding, a few caveats are in order.

[1] There would still be an interesting philosophical question to ask: one option would be to conclude that 'realism' marks a gerrymandered and uninteresting distinction, and should therefore be discarded. Another option would be to find a variety of different distinctions that 'realism' marks out in different contexts (though no single non-gerrymandered distinction that is marked in every context). In this case, the best conclusion to draw is that uses of 'realism' potentially refer to interesting distinctions, but that it would be best going forward to discard the ambiguous term in favor of names that unambiguously pick out the kind at issue in each context.

First, we need to be careful to distinguish the present project from that of performing conceptual analysis on a term of art. ‘Realism’ is a philosophers’ term, bearing no straightforward connection to our pre-theoretic vocabulary. It hasn’t been introduced by explicit definition in terms of other theoretical vocabulary. Its use is unconstrained by easily identifiable external linguistic and communal standards—all we have to guide our use of the term is our own dispositions and intuitions. But an account of realism which simply tries to describe these community-wide uses and intuitions is nothing more than a redescription of the linguistic habits of a small community. This might be of sociological interest, but is of no direct philosophical significance.

The project here is importantly different in a number of respects. As I have already mentioned, there is a possibility of failure—there might not *be* a theoretically interesting natural kind that plausibly counts as the referent of ‘realism’. We aren’t prejudging the question of whether we will be successful in finding an analysis by embarking on the investigation. Conceptual analysis of the term, on the other hand, is much less susceptible to failure: concepts can be highly gerrymandered and disjunctive. A natural kind analysis also allows for a final characterization of realism that substantially diverges from our intuitive judgments about realism, while such divergence is a mark of failure for a conceptual analysis.

The second caveat is that, in what follows, I will be assuming that a view is realist (or not) primarily in virtue of its *metaphysical* consequences. This is a natural idea—Berkeley’s idealism seems irrealist precisely because of its consequences about the nature of material objects. (They are, according to Berkeley, merely collections of ideas.) Similarly for realism about unobservables; the realist, it is often felt, takes an unwarranted or unintelligible metaphysical stance. Moorean non-naturalism is similarly said to be metaphysically too extravagant. So it is quite natural to take our project to be one of asking which metaphysical consequences of a view are necessary and sufficient for realism.

Such an understanding of realism is not, however, universally accepted. Many claim to find additional, non-metaphysical aspects to realism. Some are *epistemic*: Boyd (1989, 181-2) takes realism to imply that our cognitive faculties afford us a means of “obtaining and improving” knowledge in the relevant domain. (Dummett (1982, 55), on the other hand, claims to find in realism a distinctive commitment to the truth of claims “independently of whether we know, or are even able to discover” their truth.[2]) Other

[2] Strictly speaking, Boyd and Dummett’s claims are consistent: we might have good cognitive resources

characterizations of realism involve *semantic* properties like truth, literalness, etc.

I will proceed in what follows by ignoring these non-metaphysical dimensions to realism. The most straightforward motivation for this is methodological—insofar as it is clear there is some metaphysical component to realism, a characterization that posits additional epistemic or semantic dimensions to realism will thereby be less natural and more gerrymandered. A search for the natural kind underlying talk of 'realism' then does best by beginning with a purely metaphysical characterization; other dimensions should be added only if a purely metaphysical conception of realism is unavailable.

The final caveat is that there may well be specific domains where use of 'realism' has been spun off from its general philosophical use, and in these contexts has a specialized meaning. Plausible cases include: scientific realism, mathematical realism, and legal realism. (I don't want to assert that these *are* distinct uses of the term; just that these *might* constitute special cases, or that some uses of 'realism' in these domains have a special domain-specific sense.) We shouldn't expect any general account to accommodate specialized meanings of 'realism'.[3]

The aim of the present chapter, then, is to find a metaphysical characterization of the natural kind picked out by our general use of 'realism' (if any such kind exists). We will proceed from here as follows. §2 outlines three popular accounts of realism in the literature: these are the *existence*, *mind-independence*, and *fundamentality* conceptions of realism. §3 argues that they face a pervasive problem: they cannot account for the fact that some but not all analyses of a domain are compatible with realism. §4 piles on, adding two further arguments against these prominent conceptions: they cannot account for the *gradability* of claims about realism, and they cannot make intelligible the disposition of some philosophers to set a very high standard for which views count as realist. These objections together are, I think, decisive against the existence and mind-independence conceptions of realism. But the situation with respect to the fundamentality-based view is different: these objections only cause trouble for views which take give an account of realism in terms of *absolute* fundamentality. There is another approach in the neighborhood, which instead characterizes realism in terms

for arriving at knowledge of a domain, while some of its claims are nonetheless in principle unknowable. Regardless, the divergence over the epistemic dimension of realism is telling.

[3] I don't have any test for when 'realism' is used in a specialized way. But there are clear-cut cases: for instance, in some of these domains there are explicit references to stipulative definitions for 'realism'; in these cases, it is very natural to take the continued debate to be over the stipulated thesis. Since I do not wish to commit to the claim that this is the only way 'realism' in a domain can take on specialized meaning, I will remain open on whether (and how) it takes on different meanings in different contexts.

of *degrees* of fundamentality. §5 sketches how an account along these lines yields a promising account of our subject matter.

1.2 Three conceptions of realism

There are three main metaphysical conceptions of realism: existence views, mind-independence views, and fundamentality-based views.

1.2.1 Existence views

Existence views hold that a theory is realist just in case it entails that entities of an appropriate kind exist. What *kind* of entity is required is variable: some versions hold that realist theories entail that *properties* of the relevant kind exist; other versions hold the same for the relevant kind of *facts*. Existence-based views are prominent in the literature on ethical realism.

One instance is found in J. L. Mackie (1977), where he intends his metaethical view, which he calls "moral skepticism", to be the denial of ethical realism. He characterizes this view in the following way:

> [W]hat I have called moral skepticism is a negative doctrine, not a positive one: it says what there isn't, not what there is. It says that there do not exist entities or relations of a certain kind, objective values or requirements, which many people have believed to exist.[4]

Mackie's "negative thesis" is a denial of an existence claim—namely, the claim that certain "entities or relations" exist, and is supposedly in conflict with standard realist conceptions of ethics on this count. This presupposes that the realist view entails the existence of certain things—"values," as Mackie says.

Shafer-Landau (2003) offers a broad taxonomy of available metaethical positions. The first position is the eliminativist view, which "is represented by error theorists and non-cognitivists. Such philosophers do not believe that there are any moral properties, and believe that all appearances to the contrary are either founded on error, or can be otherwise explained away." (p. 66) The other options are reductionism, which holds that "moral properties, if they are to exist, must be (in the sense of be identical to) one of these kinds of natural property" (pp. 66-7), and non-naturalism, which rejects "the identity of moral and descriptive properties." (p. 72)

[4] Mackie (1977, p. 17)

On a standard classification, only the last two views—the reductionist and non-naturalist views—are the views that are consistent with realism. The eliminativist view, represented by error theorists and noncognitivists, is not. What separates these realist views from others in Shafer-Landau's taxonomy is that they entail the existence of moral properties. This strongly suggests that Shafer-Landau takes the existence of these properties to be the key ingredient for realist views about ethics.[5]

These existence-based approaches to realism about ethics can be thought of as generalizations on a standard characterization of realism about unobservables in scientific theories. In van Fraassen (1980), the characteristic claim of realism is that *there are* electrons and other unobservables posited by scientific theories. This makes sense in the context of realism about unobservables: the primary motivation of the irrealist is to avoid what she believes to be an unwarranted ontological commitment to an unobservable world of electrons, and the way to avoid this commitment is to decline to believe that they exist.[6] Existence conceptions of realism in other areas are then natural extensions of this idea to other domains.

If this is the motivation of the existence approach, it is already suspect. Not all metaphysical commitments are *ontological* commitments in the sense that they are commitments concerning which objects or entities exist. Theories can contain unwanted metaphysical commitments by including an unnecessarily complex primitive *ideology* as well. Much of Lewis (1986), for instance, is motivated by the desire to eliminate any primitive modal ideology in the form of terms like 'possible', even if it comes at a high ontological cost. The Moorean non-naturalist about ethics is most naturally accused of this sort of ideological promiscuity. She agrees with her irrealist counterparts over matters of ontology, countenancing exactly the same actions, events,

[5] Elsewhere, he says that what is definitive of realist views is that they entail the existence of moral *facts* (see for instance Shafer-Landau (2003, p. 15)). Shafer-Landau may either be undecided between one of two existence-based views, or may think that they amount to the same thing. The latter view would make since if one thought that facts are structured set-theoretic entities with properties (among other things) for constituents. Then, the failure of ethical properties to exist would by itself give rise to a lack of existence in ethical facts. I won't work with these distinctions in the main text, since it will not matter for much of what I say whether the existence view is primarily concerned with properties, facts, or similar entities.

[6] In the case of van Fraassen's irrealist alternative, the irrealist doesn't take on the contrary commitment by denying that there are electrons. Rather, she withholds belief and (in van Fraassen's terms) merely *accepts*, rather than *believes*, scientific theories for the purposes of carrying out scientific investigation. I take this to be an instance of the existence conception of realism, even though van Fraassen nonetheless recommends *acceptance* of an existence claim. This is because the difference between acceptance and belief concerns whether bearing the relevant attitude to the claim that electrons exist brings along an ontological commitment to the existence of electrons. I take van Fraassen to recommend mere acceptance over belief precisely because it does not bring about this kind of commitment.

persons, etc. They differ ideologically. What the non-naturalist requires is the further ideology of wrongness to apply to certain actions; the irrealist attempts to do without such primitive ideology.[7] The existence view is then naturally seen as an attempt to expand an account of realism about ontological matters into an account that covers realism about ideology as well. I think it is not at all obvious that the this kind of assimilation is warranted, but will postpone criticism of the view until §3.

1.2.2 Mind-independence views

Another common way to characterize realism about a domain is to claim that all realist views hold the domain to be *independent* of the mental. Examples of mind-dependence (and, consequently, irrealism) are familiar from the history of philosophy: think of Berkeley's claim that ordinary objects are collections of ideas, or Hume's belief that causation is nothing more than constant conjunction plus expectation. This thought seems especially apt when considering irrealism the ethical domain, as many paradigmatic instances of irrealist ethical theories enlist mental states of approval, disapproval, and the like, to play important explanatory roles.

One way to articulate this approach is found in Sharon Street (2006). She says:

> The defining claim of realism about value, as I will be understanding it, is that there are at least some evaluative facts or truths that hold independently of all our evaluative attitudes.[8]

For Street, then, metaethical theories are realist just in case they entail that ethical facts are independent of evaluative attitudes.[9]

Before moving on to a generalization of mind-independence accounts to other domains, there is an important nuance to Street's characterization of realism that needs to be addressed. (Similar issues arise with some other characterizations of realism, and the point I make below will apply to them as well.) Readers familiar with Street's work will notice that she insists that her definition is purely stipulative, and might contend that her characterization therefore cannot be criticized as failing to adequately capture some pre-theoretically grasped notion of realism. It is, of course, possible to stipulate

[7] As I discuss later in §3, not every attempt to do without the primitive ideology of wrongness is thereby an instance of irrealism. The only point I am making here is that some avoidance of certain metaphysical commitments are the primary motivation for some versions of irrealism.

[8] Street (2006, p. 110).

[9] For more discussion of mind-independence and realism, see Jenkins (2005), Brink (1984).

meanings for terms, and ‘realism’ is no different in this respect. A stipulation of this kind, moreover, isn’t the kind of thing that can be *argued* against. I agree with these claims, and so, strictly speaking, will not subject the stipulation that ‘realism’ means mind-independence to criticism below.

But there is more going on with characterizations of realism which, like Street’s, are officially stipulative. The amount of recent interest in Street’s arguments does not derive simply from the fact that she has labeled *some* thesis ‘realism’ and criticized it. (This would be remarkably easy to do and in many cases would be of absolutely no philosophical interest.) Rather, interest in Street’s argument derives from a sense that she has defined ‘realism’ in terms of a philosophically important notion. Many philosophers are, moreover, inclined to think that the question of mind-independence is a philosophically important notion.[10] And this isn’t just a sense that mind-independence is *somehow* a philosophically interesting question; this is a sense that it is the question that matters *for realism.* So, even though we won’t be arguing against the explicit stipulation Street makes in what follows, we will be assessing the presupposition that the stipulation is warranted because mind-independence is, in general, what is at issue in disputes over realism.[11]

A mind-independence characterization of realism can be extended to other domains in various ways. The basic idea is that just as facts about value are mind-dependent if they depend on our evaluative attitudes, so likewise other domains are mind-dependent if they depend on our attitudes in some way.[12]

1.2.3 Fundamentality-based views

A final approach to realism proceeds in terms of the notion of *metaphysical fundamentality.* This notion goes under different headings in the literature; Fine (2001) calls it “Reality”; Sider (2012) calls it “Structure”, and Lewis (1983) uses the term “perfect naturalness”. Ralph Wedgwood (2007) articulates the relationship between this idea and realism in the following passage:

[10] For more, see Rosen (1994).

[11] This is not to say that mind-independence in ethics is not of interest on its own, regardless of its connection to realism. It may well be, and then Street’s arguments can be evaluated in terms of whether they refute views on which the ethical is mind-independent. The present point is not that mind-independence is a philosophically uninteresting notion; only that we should wonder whether it is the philosophically interesting notion picked out by ‘realism’ on its ordinary usage.

[12] There are many different ways to flesh out both *which* attitudes are at issue, and exactly how the domain *depends* on the attitudes. (See Jenkins (2005) for more on the various dependence relations that might be employed here.) I will ignore these details, because I believe that there are cases which are counterexamples to the mind-independence characterization understood in any of these ways.

> What exactly is realism? Following Kit Fine (2001) I shall suppose that a realist about the normative is a theorist who says that there are normative facts or truths—such as the fact that certain things ought to be the case, or that it is not the case that certain things ought to be the case—and that at least some of these normative facts are part of reality itself.
>
> The notion of *reality* invoked here is a notion that has its home within a certain sort of metaphysical project—namely, the project of giving a metaphysical account or explanation of everything that is the case in terms of what is real [. . .] [I]f certain normative facts are real, then [. . .] these normative facts, properties or relations may also form part of the fundamental account or explanation of certain things that are the case.[13]

Wedgwood—and his predecessor Fine—primarily use the term 'Reality' to signify the metaphysically privileged layer at which gives "a metaphysical account or explanation of everything that is the case". For terminological uniformity, I will instead use the term 'fundamental'. In the sense in which I intend it, then, it is a blanket term for what the notions employed by Fine, Sider, and Lewis have in common. It describes whatever it is that features in the most basic metaphysical explanations.

Realism about the ethical on the Fine-Wedgwood conception is the view that the most fundamental explanation of everything that is the case makes reference, in part, to ethical facts or properties. This conception of realism generalizes easily to other domains: realism is similarly the view that the domain in question is fundamental.

1.3 Realism and analysis

§2 outlined three prominent metaphysical approaches to realism. They all face a common problem: they fail to properly account for the relationship between realism and reductive analyses.

While uncontroversial reductions are rare in philosophy, we are familiar with a range of purported reductions, even if they are known to fail. Examples include: the Russellian reduction of physical objects to logical constructions of sense-data;[14] the Lewisian reduction of modality to quantification over maximally complete chunks of concrete spacetime,[15] and the Logicist's reduction of mathematics to logic.[16] Our concern here will not be with whether these (or other) reductions are correct; we will only be concerned with whether these reductions—correct or not—are consistent with realism.

[13] Wedgwood (2007, pp. 1-2)

[14] Russell (1912)

[15] Lewis (1986)

[16] Whitehead and Russell (1910)

There is no simple answer to this question: some reductions are, and others are not, consistent with realism about the reduced domain. But none of the conceptions of realism outlined in §2 can explain why this is so.

1.3.1 Existence and reduction

It is a familiar theme that some reductions are "vindicating" and others are "eliminative". At a first pass, the difference is something along the following lines. Vindicating reductive accounts give an informative characterization of the reduced property or domain—they tell us something about the nature of the reduced thing. Other reductions show that our purported talk about the property or domain was really talk about something else—they show us that what we thought we were talking about isn't really there. Railton (1989, 161) gives an example:

> The successful reduction of H_2O reinforces, rather than impugns, our sense that there really is water. By contrast, the reduction of "polywater"—a peculiar form of water thought to have been observed in scientific laboratories in the late 1960's—to ordinary water-containing-some-impurities-from-improperly-washed-glassware contributed to the conclusion that there really is no such substance as polywater. Whether a reduction is vindicative or eliminative will depend on the specific character of what is being reduced and what the reduction basis looks like.

This is an intuitive difference—it really does seem like, upon learning of the relevant reductions, beliefs about polywater are discovered to be mistaken, while no widespread error is revealed for beliefs about water. But *why* is there such a difference? The natural answer is that the terms 'water' and 'polywater', like many theoretical terms, are associated with a "theoretical role" that determines as referent the property that best satisfies a set of theoretical constraints. These constraints include the observed properties of the relevant substance (that it is wet, clear, drinkable, etc.), the role it plays in explanations (that salt dissolves in it), among other things. The important part here is just that while these theoretical constraints tolerate *some* divergence in a candidate referent, if the best candidate strays too far from the intended role, the term fails to refer.

The difference between the water and polywater reductions, then, is that the water reduction supplies a property (H_2O) that sufficiently approximates the theoretical role associated with 'water'; the polywater reduction supplies a property (water-plus-impurities) that does not. As Railton says, not even the water reduction provided a *perfect* satisfier for the relevant theoretical role: "[e]ven the reduction of water to H_2O was in part revisionist ... of both common-sense notions and previous chemistry".[17] But the reduction isn't *radically* revisionary, and so we conclude after the relevant empirical investigation that H_2O adequately satisfies that theoretical role associated with 'water'. The case is different with polywater: presumably those who originally introduced the term thought they discovered a new, interesting form of water with a molecular basis similar to that of water. This supplies a theoretical role for 'polywater'. But upon discovering that the "substance" in question was really just water-plus-impurities, we learn that the theoretical role isn't even close to being satisfied. A substance that is water-plus-impurities does not have molecular basis similar to that of water. (Compare, for instance, the difference between the gerrymandered molecular basis for polywater, and ^2H_2O, or "heavy water". Only the latter has a molecular basis similar to that of water, but scientists presumably thought they were discovering another molecular variant of this kind when they coined 'polywater'.) Hence we learn that the theoretical role associated with 'polywater' goes unsatisfied—the water-plus-impurities referent is *too revisionary* to satisfy the associated theoretical role. Polywater does not exist.[18]

This is a semantic story about the difference between vindicating and eliminating reductions. It is just a sketch, but that something like it explains the difference between the reductions in Railton's example, and in the difference between vindicating and eliminating reductions more generally, seems highly plausible. It can easily be extended to other domains: Mackie, for instance, is naturally interpreted as claiming that there are no properties that come close to satisfying the theoretical role for 'wrong'. This is because the theoretical role for 'wrong' requires that its satisfier be objectively prescriptive, and nothing (according to Mackie) comes close to satisfying that role. The existence-based view of realism then has the outlines of an account of when a reduction is incompatible with realism: this occurs when the posited reduction basis is too revisionary, in the sense that it substantially fails to satisfy the theoretical role associated with the term. In such a case, the reduction is eliminative, which means that from the reduction it follows that there is no such thing as the reduced domain. Given the existence conception of realism, the lack of existence in the reduced domain amounts to irrealism about the reduced domain. This is encoded in the following thesis:

REVISIONARY REDUCTION BASIS (RRB) A reduction expressed by ⌜t reduces to b⌝ is

[17] Railton (1989, 161)

[18] Thanks to David Manley for suggesting this reading of the existence view.

> incompatible with realism just in case *b* substantially fails to satisfy the theoretical role associated with t.

Whether the existence conception of realism is plausible boils down to the question of whether RRB is true.

RRB requires that every irrealism-entailing reduction offers a reduction base that strays too far from the theoretical role associated with the reduced domain. This sets a high bar. In making the case that certain mental states cannot be given a vindicative reduction, Paul Churchland (1981, 67) suggests that this is because nothing even comes close to satisfying the folk theoretical role associated with mental states, in the following way:

> Eliminative materialism is the thesis that our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience.

If a reduction must show the associated theoretical role to be a "radically false theory" in order to be eliminative, eliminating reductions will be hard to come by. A reductive Behaviorist, who reduces mental states to disjunctions of behaviors, need not on this conception be an eliminativist. This is because she finds the ontology of mental states, and principles to connecting mental states to behavior, to be by and large true and useful. (The distinctive feature of the Behaviorist is that the connection between mental states and behavior is more or less trivial.) Behaviorism as described here nonetheless seems to be an irrealist theory. RRB, and the existence conception of realism more generally, fail if they follow Churchland in insisting that an extremely high degree of departure from theoretical role is required for elimination.

Perhaps Churchland is, for rhetorical purposes, making a stronger claim than required: eliminating reductions need not posit a basis that shows the theoretical role at issue to be a "radically false theory" and "fundamentally defective". A lesser degree of divergence between a proposed reduction basis and relevant theoretical role might nonetheless be sufficient for elimination.

One problem with this proposal is that there are accounts which are to some extent revisionary yet which appear to be staunchly realist. A vitalist view of living organisms such as that found in Bichat (1801, §1) is substantially revisionary in view of the theoretical role we associate with 'life'. (I assume that, whatever the details of the role associated with 'life', it includes the claim that life is not explained by a primitive life-force, but rather can be explained by other biological and chemical processes.) The vitalist view certainly has vices that warrant its rejection, but among these is not its *irrealism* about life. In fact (and this is a subject I will return to below), its vices stem in part from the fact that it is unnecessarily realist about life, giving it a basic explanatory role when none is needed. This isn't an isolated example, either: any time the revision goes in the direction of giving a greater explanatory role to the reduced property, we will have a case of a revisionary view which is nonetheless realist. Such views are counterexamples to RRB.[19]

1.3.2 Irrealist mind-independence

On the mind-independence conception of realism, only views that entail a domain to be mind-independent are realist. But reflection on the reductive Behaviorist view from §3.1 suggests that this is not so. According to that view, mental states are just disjunctions of behaviors, or dispositions to behave. Thus, the mental state *pain* reduces to either clutching one's arm, or screaming, or .... Similarly for other mental states. The resulting view is, as we mentioned before, an irrealist one.

But it also satisfies the conditions imposed by any reasonable construal of mind-independence. That Sam is exhibiting the behavior of (say) clutching her arm doesn't depend in any natural sense on the mental—neither the mental states of Sam nor those of any ascribee. That she exhibits such a behavior is a fact as objective and mind-independent as any. Moreover, the Behaviorist view seems to be irrealist precisely

[19] Another objection to the RRB-based view comes from the idea that the distinction between vindicating and eliminating reductions itself fails to be substantive. Quine (1960, 265) suggests something along these lines when he says:

> For a further parallel consider the molecular theory. Does it repudiate our familiar solids and declare for swarms of molecules in their stead, or does it keep the solids and explain them as subvisibly swarming with molecules? [...] The option, again, is unreal.

We might explicate Quine's thought as follows: it is indeterminate, or perhaps a matter of pure linguistic convention, whether we associate with 'solid' a theoretical role that is adequately satisfied by subvisible swarms of molecules. What exists isn't in question: it is swarms of molecules. But whether this is sufficient for the truth of the sentence 'solids exist' is just a matter of whether we choose to associate a more or less strict theoretical role with 'solid'. The truth (or falsity) of this sentence doesn't reflect a deep metaphysical fact; just a choice about our language.

On this way of explicating Quine, his comments suggest that talk of existence isn't sufficiently metaphysically robust to capture the metaphysical dimension to realism.

because a non-mental reduction basis of behaviors isn't the right kind of thing for the mental to reduce to. The mind-independence conception of realism gets things completely backwards in this case. (Consider also the converse case: a Cartesian view of mental states explicitly denies their reducibility, and instead identifies them with modifications of a fundamental mental substance. Thus mental states on this view aren't mind-independent, but the Cartesian is a realist *par excellence* about the mental.)

The same issue will arise for any theory that is about broadly psychological phenomena. The account of syntactic principles as "psychologically real" in Noam Chomsky's *Knowledge of Language* is strikingly realist in contrast with its competitors. But it is precisely because he postulates a psychological realization of the principles that his view has this feature. By loosening the tie with the psychological—say, by interpreting syntactic principles as merely the most convenient representations a complex set of human linguistic behavior—one thereby distances oneself from Chomskian realism.[20]

1.3.3 Realism and the absolutely fundamental

The fundamentality view of realism holds that realist views about a domain are just those that take the domain to be *most fundamental*. This view nicely accommodates some of our earlier examples of irrealism. By analyzing mental states in terms of behaviors, the Behaviorist view entails that mental states are not most fundamental. And, by analyzing wrongness in terms of speakers' attitudes of disapproval, the Subjectivist view entails that wrongness is not most fundamental. They are both irrealist views, as the fundamentality-based view predicts.

But the fundamentality view gets the right results in these cases for the wrong reasons: *any* analysis of a domain will entail that it is not fully fundamental, and hence will be an analysis that entails irrealism about the analyzed domain. There are plenty of examples of analyses that are consistent with realism.

Here are two. An Identity theorist reduces mental states by identifying them with neurophysiological states. Hence, according to the Identity theorist, pain is not most fundamental; some neurophysiological state is more fundamental than it. Likewise, the view of moral properties like wrongness presented in Railton (1986) is one on which they

[20] See Chomsky (1986, p. 39). Some authors such as Miller (2010, sec. 1) claim that the mind-dependence of psychological claims is not a counterexample to a mind-independence conception of realism, on the grounds that the dependence is *trivial*, and the conception of realism on offer claims that all realist theories entail that their domain is not *non*-trivially mind-dependent. But Chomsky's claim about the mind-dependence of a grammar is not a trivial one, as it is supported only by sophisticated arguments about language acquisition. See also Reynolds (2006, 481) for more discussion.

reduce to facts about what promotes human interests from the "social point of view". Hence human interests are more fundamental than moral properties. But both views are intuitively consistent with realism— Railton even presents this view in a paper called "Moral Realism".[21] The fundamentality-based approach fails to count these views as realist *merely because* they are reductive views about the relevant domains.

1.4 Further problems: gradability and shifting standards

The previous section focused on the relationship between realism and reduction to show that none of the §2 views are plausible metaphysical characterizations of realism. The present section adds two separate difficulties: the gradability of ascriptions of realism, and the apparent flexibility of the standards for realism in philosophical conversation. While (I will argue) these phenomena are inconsistent with the views of realism from §2, they also serve a positive purpose. In §5, I will suggest that they point us toward a specific refinement of the fundamentality view which can accommodate the issues raised in §§3-4. The crucial modification is that it treats realism primarily in terms of *relative*, and not absolute, fundamentality.

1.4.1 Gradability

Some judgments about realism can be *graded* in form: that is, it makes sense to say that view *A* is *more realist* than *B*, without committing to the ungraded claim that *A* is realist and *B* is not. (Compare: Bill might be *taller than* Bob, without its being the case that Bill is tall and Bob is not.) Modality provides an example: consider first a fictionalist view about modality, an example of which can be found in Rosen (1990). Such a view holds that

> Possibly I have red hair

is false on its literal interpretation, since it says that there is a possible world where I have red hair, and possible worlds do not exist. But as a consolation, the fictionalist says that it can be appropriate to assert such sentences; the related claim

> According to the fiction of possible worlds, there is a possible world where I have red hair

[21] Readers familiar with that paper will note that Railton acknowledges on pp. 200-1 that his view lacks some of the characteristic features of realism (though he nevertheless claims that it resembles realism enough to deserve the name). I will return to the question of *how* realist Railton's view is in later sections; for now, I will only make the point that Railton's view is much *more* realist than the standard Subjectivist view of wrongness.

is true, and unembedded modal utterances convey the same information. The situation is analogous to ordinary statements about fiction: the sentence

Holmes lives at 221b Baker Street

is literally false (Holmes doesn't exist); but it can be asserted, as it conveys the information encoded by the true sentence

In the Sherlock Holmes fiction, Holmes lives at 221b Baker Street.

This view is an irrealist view when compared with the view of modality put forward in Lewis (1986), which analyzes modal claims in terms of quantification over maximal chunks of concrete spacetime. On the Lewisian view, 'possibly I have red hair' is literally true, as there is, given the Lewisian ontology, a concrete chunk of spacetime where a counterpart of mine has red hair. The point I want to focus on is that this view seems strikingly *realist* in comparison with the fictionalist view described above.

The gradability of judgments concerning realism about modality comes into focus when we consider a third view of modality on which it is unreducible. Views of this kind are found in Adams (1974), Plantinga (1978) and Stalnaker (1976), the details of which need not concern us here. All that matters is that where the Lewisian view quantifies over a concrete chunk of spacetime in analyzing 'possibly I have red hair', the present views use a modal term to specify the domain of quantification, and deny that it can be paraphrased away. Proponents of this view have claimed that they, and not Lewis, deserve the honorific 'realist' when it comes to modality. Here is Plantinga:

> Lewis is a modal realist and/or a realist about possible worlds in approximately the sense in which William of Ockham is a realist about universals: namely, not at all.[22]
>
> Of course there is *something* in the neighborhood with respect to which Lewis is a realist, and a pretty unusual and interesting thing at that: a plurality of maximal objects [...] Lewis is certainly a realist of an interesting kind, but what he isn't is a *modal* realist. On his theory, as I see it, there are no propositions, states of affairs, possible worlds, essences or objects with essential and accidental properties; what there are instead are concrete objects and set theoretical constructions on them, some of which play roles formally similar to the roles in fact played by the phenomena of modality if the modal realist is right.[23]

[22] Plantinga (1987, p. 189)

[23] Plantinga (1987, p. 213, his italics)

Plantinga may or may not be right to say that Lewis is definitively not a realist about modality. (I will return to this question in the next subsection.) Regardless, what Plantinga's quote draws attention to is that the nonreductivist about modality can plausibly claim that her view is *more* realist about the modal—after all, where Lewis finds modal locutions to be eliminable in favor of quantification over concrete spacetime, the nonreductivist insists that the modal locutions are not eliminable at all.[24]

But with the fictionalist view in the picture, we can't account for this difference between the Lewisian and nonreductivist by saying that the nonreductivist view is realist, while the Lewisian view is not. Calling the Lewisian view irrealist would fail to distinguish it from fictionalism—but surely there is a difference, since only one view allows that possible worlds exist, and that modal claims are literally true. And calling the Lewisian view realist would blur the distinction with Plantinga-style non-reductive views, as only the latter posits primitive, unreduced modality. The situation is markedly improved by giving up on a simple binary distinction between what is real and what is not; we can then say that the Plantingian non-reductivist is *more realist* about possibility and necessity than her Lewisian counterpart. And the Lewisian is more realist about the relevant notions than the fictionalist is. Thus, insofar as we have a grip on the features picked out by our talk of "realism", it would seem that they come in degrees.

One can easily generate similar cases in other domains where both reductivist and nonreductivist versions of realism are live options. For instance, Moorean non-naturalism is clearly a realist view, and differs in this respect from the versions of Subjectivism we outlined earlier. But reductivist views such as those found in Railton (1986) and Schroeder (2007) appear to differ in *some* relevant respects from both views, though for similar reasons this distinction is missed if we use only a binary distinction between realism and irrealism.

Gradations of realism are apparently inconsistent with any characterization of realism that proceeds in terms of existence, mind-independence, or an absolute notion of fundamentality. These are all absolute or binary notions—something can't exist *more* than another, neither can one thing be *more* mind-independent than others. It makes

[24] If one doesn't feel the intuition strongly—perhaps one is convinced that 'modal realist' *is* an appropriate label for the Lewisian view—one might try instead to imagine being someone who is *unsure* whether the Lewisian view is realist or not. (Certainly it is not obvious in every case whether a particular view is realist or not.) When one is unsure in this way whether the Lewisian view is realist, one won't find oneself completely at a loss for words to describe the situation: for one should still say, while remaining agnostic about the realism of Lewis, that his view is *more* realist than the fictionalist's, and *less* realist than the nonreductivist's.

sense to say that something is more fundamental than another, but the fundamentality-based view, as stated, makes no use of this gradable notion. So, as stated, the fundaentality-based view is incompatible with gradations in realism. (We will, in the next section, return to the possibility of using a graded notion of fundamentality in place of the absolute notion.) A gradable notion requires an analysis in gradable terms.

There is more to be said about this argument, but I want to address only a few attempts to supplement the existence view with resources to account for gradability. Similar strategies could, perhaps, be pursued with the binary 'mind-independent' or 'fundamental' instead, but these seem most naturally wedded to the existence view, so I will ignore these other views for the sake of exposition.

Given the semantic characterization for the existence view in §2, it seems reasonable to claim that the Lewisian view is more realist than the fictionalist view, but less realist than the nonreductivist view, owing to the following facts. First, the theoretical role attached to 'possible', even in its strictest forms, is satisfied according to the nonreductivist view. Second, the same is not true for the Lewisian view: some strict theoretical roles for 'possible' are not satisfied by quantification over concrete chunks of spacetime. These strict theoretical roles require that whatever satisfies them not be non-modal concreta. (This might be what Plantinga is getting at when he says that some elements of the Lewisian ontology "play roles formally similar to the roles in fact played by the phenomena of modality".) But there are some more relaxed theoretical roles for 'possible' which the Lewisian ontology does satisfy. And this distinguishes the Lewisian view from the fictionalist view, which doesn't offer a satisfier even for these relaxed theoretical roles. This suggests a proposal on which gradations of realism are simply numbers of theoretical roles on which the relevant domain can be said to exist.[25]

One point to make about this proposal is just that this semantic phenomena doesn't by itself explain gradability. The multiplicity of available meanings for 'possible' does not by itself entail gradable claims of the right kind. Consider the polysemous term 'keep': it has various related meanings, and the following sentences are true on some but not necessarily all of these readings:

[25] Here is a friendly analogy: 'pregnant' is often said to be, like existence, non-gradable. But we can use the gradable 'more pregnant' to compare groups of people, members of which might or might not be non-gradably pregnant, as follows:

On the whole, the 20-to-30-year-old population is more pregnant than the 30-to-40-year-old population.

Thanks to Ezra Keshet for the example and discussion.

1. John kept the score sheet;
2. John kept a penny;
3. John kept time.

1 can be true on both the readings of ‘keep’ in 2 and 3, while 2 and 3 are true on at most one such reading.[26] But we wouldn’t be willing to say (for instance) that John kept the score sheet *more* than the penny. We could, in principle, introduce a term that relates the score sheet to the penny and time in the right way. But such a term would have a clearly meta-linguistic meaning, which would have roughly the sense of ‘satisfies more meanings for ‘keep’ than’. It is clear that ‘more realist than’, to the extent that it makes sense, is not meta-linguistic in this way.[27]

All the case of ‘keep’ shows is that satisfying a greater number of related meanings for a term isn’t sufficient for the truth of a gradable claim expressed by that term. This leads to a second point: perhaps something more can be said about the way in which the theoretical roles are relaxed in discussions of realism; in particular, there might be some quality or dimension such that related theoretical roles are obtained by selecting different thresholds along that dimension. (For ‘loud’, for instance, gradable claims about what is louder than what can be thought of following from facts about how many candidate meanings for ‘loud’ the things satisfy, when different meanings are obtained by selecting different thresholds for loudness along the dimension of volume.[28]) But once such a privileged dimension is singled out, it will be very natural to give the account of realism in terms of *that* dimension directly. In such a case, the detour through the language of existence is unneeded.[29]

Another approach to an existence-based account of gradability is to treat the multitude of theoretical roles associated with a term as varying on a dimension of *literalness*:

[26] John can keep a penny by putting it in his pocket; he can keep time by ensuring that the game proceeds in accordance with the relevant device. He can do both of these with the scoresheet, but not a penny or time.

[27] Thanks to David Manley for discussion here.

[28] Similarly, the pregnancy comparison in the earlier footnote relies on a single comparable metric: number (or rate) of pregnancies within a population. The challenge for the existence view is to find a single comparable dimension in theoretical roles that licenses this kind of comparison.

[29] For example: I will propose in §5 that realist an irrealist views differ in the degree of fundamentality they claim for a domain. One could wed this to the existence view, claiming that a view is more realist than another when the first satisfies more theoretical roles that differ only along the dimension of how fundamental they require the reduction basis to be. Such a view, however, is most naturally thought of as a fundamentality-based view, and not an existence view.

the strict theoretical roles are, we might think, more literal renderings of the term than their less strict counterparts. Grades of realism are, on this view, grades in the literalness with which a domain can be said to exist. Plantinga allows for the sentence ‘possible worlds exist’ to be truly asserted with a greater degree of literalness than Lewis does.[30]

Literalness won’t, however, always be tied to intuitive grades of realism. The vitalist about the life-force in organisms has a highly realistic view about life; one which is much *more* realistic than our still-realist current view, on which life and related properties are grounded in chemical and biological functioning. But (assuming we have an adequate grasp on literalness in the first place) the most literal interpretation of our term ‘life’ associates it with a theoretical role that presupposes its satisfier is grounded in the chemical, biological, and so on. The scientific rejection of vitalism is sufficiently well-known and accepted to color the standard interpretation of ‘life’. If anything, it is the vitalist who requires a less-than-fully literal use of the term, but the vitalist has the *more* realist view.

There no doubt could be other proposals for reconciling an account of realism in terms of a superficially binary term like ‘existence’ with the apparent gradability of judgments about realism. Instead of canvassing more proposals, however, I will instead move to consider related objection to these views. This will in turn serve to help motivate and clarify the positive proposal of §5.

1.4.2 Shifting standards

When gradable terms are in use, speakers can shift the standards for application of a term by making salient a particular standard. For instance, an utterance of

That music is loud

might be perfectly felicitous in a context where the music makes conversation at a normal volume difficult. But such an utterance could be followed by an interlocutor adding:

That music isn’t loud; we couldn’t hear it at all when the train went by.

It is clear that what is going on in this case is something like the following: ‘loud’ is used to indicate that something exceeds a contextually supplied threshold for volume. The

[30] Thanks to Peter Railton for the suggestion.

first utterance takes the threshold to be the volume which would makes normal conversation difficult; anything is loud which exceeds this threshold. The second utterance attempts to shift this threshold: by making salient the volume of the recent train passing by, it sets a higher standard for the volume something has to be in order to count as loud. Many gradable terms are like this, permitting speakers in a context to (attempt to) set the relevant threshold at different levels.[31]

If 'realism' is gradable, we should find a similar phenomenon. And, in fact, we do find something like this happening. §2 was concerned with how, intuitively, some reductions are consistent with realism. But not all philosophers speak this way: some insist that only views on which the domain is unreducible are consistent with realism.[32] Plantinga's claims about modality provide one instance. While Lewisian reductivism is an account on which many ordinary modal claims come out as true (and hence is to be distinguished from fictionalism), it nonetheless leaves modality out of the fundamental level. Plantinga insists that the Lewisian view is irrealist about modality on these grounds.

A similar phenomenon occurs in ethics: some reductivist views plausibly count as realist, but philosophers have attempted to deny this. Thomas Nagel (1989), for instance, says

> [I]f values are objective, they must be so in their own right and not through reducibility to some other kind of of objective fact. They have to be objective *values*, not objective anything else.[33]

In a similar vein, David Enoch (2010) describes ethical realism in a way that rules out the reducibility of the ethical, characterizing it as follows:

> [T]here are irreducibly, non-naturalist normative truths, response-independent truths that are perfectly objective and that are not reducible to—not even identical with—natural, not-obviously-normative truths'.[34]

What are we to make of these claims? One option would be to pass these uses of 'realism' off as a different use of the term—there is, we could say, the sense of 'realism'

[31] See, for instance, Kennedy (1997) for discussion.

[32] The following examples are due to Mark Schroeder; see his Schroeder (2005). He puts them to use in support of a different (though not incompatible) account of realism. I discuss his account in more detail in the concluding section.

[33] Nagel (1989, p. 139)

[34] Enoch (2010, p. 414)

which allows some reductive views to be realism, and the sense of 'realism' (used by Plantinga, Nagel, et al.) which does not. But this ignores an evident connection between the two uses: while these theorists are certainly making different claims using the term, they are nonetheless intuitively speaking about the same kind of thing as those who allow reductivist views to be realist. To see this, imagine someone who used 'realism' in a way that, in general, allows reductive views to be realist. This person might falsely say of Moorean non-naturalism that it is not a realist view. But such a person would be *contradicted* by Nagel, who would insist that Moorean non-naturalism is a realist view. This difference in use of 'realism' is not so great that it precludes its users from disagreeing with those who use the term in the ordinary way.

Another approach is to say that in cases of disagreement, one party is simply making a mistake: their application (or lack thereof) of 'realism' simply fails to match the facts about which views are, and which views are not, realist. Here the problem is different. While we have no issue with explaining apparent disagreement between apparent uses of the term, the mistakes in question are *unintelligible*. In particular, the mistake-based account provides no explanation for why (say) Plantinga, Nagel et al. are inclined to call only nonreductive views about certain domains realist, yet still count as competent users of a common term.

The gradable aspect to realism provides a neat solution that faces neither of these problems. When there is some dimension or scale along which views which are realist to different degrees differ, the threshold for *where* on this scale the cut-off point for realism lies can be a fluid matter, depending in many ways on features of context. Hence there is a natural account on which Plantinga, Nagel et al. are speaking as if the threshold for realism is a very strict one—or, alternatively, are trying to change the setting of the threshold by speaking as if it is a strict one. Thus they are using the term with broadly the same meaning as other speakers. 'Realism' is a gradable term in everyone's mouth; differences, in effect, amount only to differences over where on the relevant scale the threshold is set. This makes divergent use of the term by competent speakers intelligible since it avoids chalking the differences up to brute misapplication by some parties.

One point of clarification is in order here. I am not claiming that Plantinga, Nagel, and Enoch are *correct* to speak as if strict standards for realism are in place. Even if the semantic structure of gradable adjectives like 'loud' permit a shift to a strict threshold, it is not always correct to assertively utter a sentence whose truth requires a strict threshold. (I do not succeed in setting a higher standard for loudness if, when discussing whether a certain cricket is loud in comparison to other crickets, I point that the volume of cricket chirps is generally much lower than that of a Boeing 747.) The present claim is then weaker, in that it doesn't take a stance on exactly when standard-shifting with 'realism' succeeds. The claim is instead that, when someone speaks as if a stricter standard for a gradable term is in place, they are using the term with the same meaning as those who don't speak as if the strict standard is in place, and are no making a semantically unintelligible mistake in doing so. If they speak falsely, they are making some mistake, but it is a failure to successfully navigate conversational rules, and is consistent with full semantic mastery.

These are plausible claims, but none of them can be made on the conceptions of realism outlined in §2. Just as these conceptions do not provide an account in gradable terms, they likewise do not provide an account with a scale along which different thresholds for realism can be set. For example, if one thing can't exist more than another, then there is no room on a scale along which a contextually sensitive threshold can be set. Similarly for mind-independence and absolute fundamentality: since these are ungraded notions, they afford no room for a variable threshold.

The next section explores the most promising option for the relevant degreed notion. While it doesn't even make sense to talk about degrees of existence or mind-independence, there is a serviceable notion of degrees of fundamentality. I will sketch below what a degree of fundamentality might be, and how it might be put to work in addressing the problems canvassed in §§3-4.

1.5 Relative fundamentality and realism

One thing (or property, fact, etc.) might be *more* fundamental than another without being *fully* fundamental. For instance: it is natural to say that if acids are electron-pair acceptors, then there is something that is more fundamental than acidity, namely electrons. Of course, electrons are not fully fundamental; they are in turn grounded in further subatomic particles. Similar examples are easily multiplied: if gravity is curvature in spacetime, then spacetime points are more fundamental than gravity. And if galaxies are collections of stars and other celestial objects surrounded by an interstellar medium, then stars are more fundamental than galaxies.

I will call these claims of the form '*A* is more fundamental than *B*' claims about *relative* fundamentality, or claims about *degrees* of fundamentality.[35] Much of what

[35] Strictly speaking, these do not amount to the same thing: it could be that *A* is more fundamental than

we said by way of introducing the notion of absolute fundamentality in §2.3 applies to relative fundamentality as well: electrons, for example, provide a kind of "metaphysical explanation" for facts about electron-pairs; stars provide the same kind of explanation for facts about galaxies, and spacetime points provide the same kind of explanation for facts about gravity.

Two clarifications are in order here. The first concerns our understanding of the fundamental as that which provides the "most satisfying" metaphysical explanation (in the words of Fine (2001)). Obviously this kind of gloss applies to that which is absolutely fundamental, and cannot be applied directly to explain relative fundamentality. Since electrons have further explanations in terms of the subatomic, electron-pair acceptors do not provide the most satisfying metaphysical explanation of acidity. Sill, we can say that what is *more* fundamental provides the same kind of metaphysical explanation; it simply need not provide the *most* satisfying version of this kind of explanation. Thus, the electron-based explanation of acidity is still a metaphysical explanation of the same kind, even if it isn't the final such explanation. An analogy with causal explanation may be helpful here: one can causally explain the breaking of a window in terms of the ball that was thrown, its trajectory, the fragility of the glass, etc. This is a perfectly legitimate causal explanation if filled out appropriately. But it isn't the *final* causal explanation: that would make reference to the causal precursors of the throwing of the ball, and the causal precursors of the precursors, and so on, perhaps only terminating in a description of the Big Bang. This would be a most satisfying causal explanation, but that doesn't preclude the existence of more proximate non-final causal explanations. That which is *more* fundamental similarly provides more proximate non-final metaphysical explanations.

The second clarification is that the examples of differences in relative fundamentality mentioned above all represent discoveries from the physical sciences (in particular, chemistry, physics, and astronomy). This might be thought to distinguish relative fundamentality, as I have described it here, from the notion of absolute fundamentality as developed by Fine and others. The reason is roughly as follows: absolute fundamentality is a purely metaphysical notion which has its use (if any) in metaphysical theorizing. But relative fundamentality as described here isn't metaphysical in the same way: it is, in the first instance, discovered by empirical theorizing in the sciences. This difference

B, while there are no specific *degrees* of fundamentality, d_A and d_B, such that A is fundamental to degree d_A, B is fundamental to degree d_B, and $d_A > d_B$. At times, I will speak as if these degrees exist, but much of what I say below can be rephrased (albeit in somewhat more complicated language) using only the comparative 'more fundamental than' and without reference to degrees.

would seem to make relative fundamentality particularly inapt for use in a *metaphysical* characterization of realism as described in §1.

I would like to claim that relative fundamentality as I have described it here is no less "metaphysical" than standard conceptions of the absolutely fundamental. Lewis's conception of absolute fundamentality (or "perfect naturalness") also assigns a central role to empirical science. He says:

> To a physicalist like myself, the most plausible inegalitarianism seems to be one that gives a special elite status to the 'fundamental physical properties': mass, charge, quark colour and flavour .... (It is up to physics to discover these properties, and name them; physicalists will think that present-day physics at least comes close to providing a correct and complete list.)[36]

Lewis thus gives physics (or something close to it) a close relationship to the absolutely fundamental. But this approach to absolute fundamentality does nothing to undermine its metaphysical character. Instead, this picture is one on which physics provides an *epistemic* window into the metaphysical absolute fundamentality facts. That physics makes reference to quarks doesn't *make* quarks most fundamental; rather, it is simply the means by which we know that they are. Similarly, then, for other sciences and relative fundamentality: these sciences provide an epistemic window into the facts about relative fundamentality, but do not constitute them.

With these clarifications in place, we can investigate a positive proposal concerning the natural kind that underlies out talk of 'realism'. The positive case rests on the following observation: relative fundamentality provides the kind of scale needed to explain gradability and standard shifting in judgments about realism. It also gives a sufficiently metaphysical explanation for which reductions are, and which reductions are not, irrealism-entailing. The positive case presented here will, however, be limited to some extent. I will only claim that an account based on the notion of relative fundamentality does not face the structural difficulties from §§3-4 that plague other accounts. This should represent an accomplishment in itself: as we have seen, the most prominent views of realism on the table fail to even account for the basic data we have discussed so far. But I will not go on to offer a full-fledged defense of the account sketched below.

[36] Lewis (1984, 228)

1.5.1 Gradability

The first point to make is that if an account of realism is given in terms of relative fundamentality, then it can in principle account for the gradability of judgments about realism. Recall our earlier examples: first, a nonreductivist view about modality is quintessentially realist; the Lewisian reductivist view is less so, but still fares better on the same scale than a fictionalist view, which is not realist at all. In short, the Lewisian view is more realist than the fictionalist but not the reductivist view. This is naturally explained in terms of relative fundamentality, in the form of the following two claims:

R1 Modality on the nonreductivist view is more fundamental than it is on the Lewisian view;

R2 Modality on the Lewisian view is more fundamental than it is on the fictionalist view.[37]

We might then say that the nonreductivist view is more realist about modality than the Lewisian view is in virtue of **R1**. Similarly, the Lewisian view is more realist about modality than the fictionalist view is in virtue of **R2**. Degrees of fundamentality are naturally tied to degrees of realism.

An analogous point can be made about realism in ethics: Moorean non-naturalism is more realist than the reductivist views in Railton (1986) and Schroeder (2007), and this is in virtue of the Moorean view taking the ethical to be *more fundamental* than Railton and Schroeder do. These reductivists in turn fare much better on the scale of realism than a Subjectivist: a Subjectivist view implies that the ethical is much less fundamental than the reductivist views of Railton and Schroeder.

This is a virtue of a relative fundamentality-based account of realism, but it also raises a flag. What is it about the Subjectivist view that makes it a view according to which the ethical is much less fundamental than it is according to other reductivist views? One answer, found in Lewis (1986) is that the definition in some privileged language of wrongness on the Subjectivist view is much *longer* than the definition of wrongness according to Railton and Schroeder.[38] This has some plausibility: the Subjectivist requires a complicated definition of wrongness that assigns intensions to

[37] **R1** holds because the nonreductivist takes modality to be *most* fundamental—it reduces to nothing further—while the Lewisian takes it to be less fundamental as she explains modal facts in terms of facts about concrete chunks of spacetime. **R2** holds because modal facts as facts about concrete chunks of spacetime are more fundamental than the facts about the fiction of possible worlds that appear in the fictionalist's account.

[38] Here is Lewis:

utterances in contexts; other reductivists can do without this added layer of complexity. But such a difference in length of characterization is just a difference of complexity at *some* level of description—and it isn't obvious that these differences will remain if we describe these properties in more fundamental terms. Indeed, we might worry that there will be *no* good reason to claim that the Subjectivist has a more complicated definition at the most natural level of evaluation of such matters.[39]

An account of realism in terms of relative fundamentality then owes an account of what makes for differences in degree of fundamentality that is not simply tied to complexity of definition. I will not try to show here how it can repay the debt, and will instead work with an intuitive understanding of where questions of degree of fundamentality arise. But this shouldn't obscure the need for something to be said on the matter—it is a non-trivial question whether an adequate account of relative fundamentality is available for present purposes.

1.5.2 Reduction

While relative fundamentality can provide us with an underpinning for degrees of realism, this does not by itself provide an account of the referent of the non-degreed term 'realism'. Compare: simply knowing that my coffee grinder is louder than a cricket is not sufficient for knowing whether my coffee grinder is loud *simpliciter*. Our discussion of reductions that are (and reductions that are not) consistent with realism provides some further direction concerning what such an account should look like.

If we begin by considering the difference between an Identity theorist about mental states (a realist) and a reductive Behaviorist (an irrealist), it is very natural to say that the difference between the two views lies in how fundamental they take mental states to be. If pain is a particular neurophysiological state, it is more fundamental than it is if it is a disjunction of behaviors. This makes it quite natural to adopt a threshold account of realism about mental states along the following lines:

MENTAL STATE REALISM (MSR) There is a degree of fundamentality *d* such that a

> Some few properties are *perfectly* natural. Others, even though they may be somewhat disjunctive or extrinsic, are at least somewhat natural in a derivative way, to the extent that they can be reached by not-too-complicated chains of definability from the perfectly natural properties. (Lewis (1986, 61); see also Lewis (1983, 347)).

[39] See Hawthorne (2006, 2007) for versions of this problem and a sketch of a positive proposal about the nature of relative fundamentality. See also Chapter 2.

> theory *T* is realist about mental states just in case *T* entails that mental states are fundamental to degree *d*.

The assumption behind MSR, then, is that the Identity theorist's view entails pain to be fundamental to the relevant degree, whereas the Behaviorist's does not. Views analogous to MSR which posit a threshold for realism might then be adopted for other domains.

MSR and its analogues raise the question of whether the threshold for realism is the *same* across different domains. That is: is it the case that there is a single degree of fundamentality *d* such that realist views about *any* domain entail it to be fundamental to degree *d*?

Here is a simple argument that the answer is 'no'. An Identity theory of mental states of the kind we have been discussing holds that mental states are neurophysiological states. A number-theorist might identify numbers with similar entities—perhaps the synaptic firings that correspond to counting operations in normal human minds. Thus, the number 2 on this view reduces to the neurophysiological state that occurs when normal humans count to the second item in a normal counting sequence. The reduction base for pain and the number 2 are then very similar in kind according to these views; hence pain and numbers are, on these views, fundamental to the same degree. But Identity theory seems clearly to be a realist view of pain, while our number theorist holds an irrealist view about numbers. So the threshold for realism about numbers and mental states must be set at different points on the scale of degrees of fundamentality.

If there is variability in where the threshold for realism is set, one approach to accommodating it is to take another aspect of the analogy with gradable adjectives seriously. For 'loud' and other gradables, the threshold is set by conversational context. Exactly what features of context are relevant, and how they conspire to set a standard for loudness is a tricky matter. But it is clear that my coffee grinder counts as loud in some contexts and not others, and that the difference between these contexts in part has to do with the *comparison class* at issue. The comparison class contains contextually and conversationally salient objects, and determines in some way where on the scale of volume the threshold for loudness is to be set. In contexts where the comparison class contains only chirping crickets, my coffee grinder counts as loud; in contexts where the comparison class contains only train whistles, it does not.

The comparison class in a discussion of realism is naturally taken to include other salient views about the domain in question. Thus, when realism about mental states is at issue, the comparison class includes theories of mental states that conversational participants take to be relevant. This comparison class then (somehow) sets a threshold for fundamentality. Quite plausibly, the salient views about mental states will constitute a comparison class that determines a degree of fundamentality that is higher than the degree to which mental states are fundamental on the Behaviorist view. This is what sets the relevant threshold for MSR.

The story about realism in other domains is then a variation on this theme, and may well require a different threshold. This is because when a realism about a different domain is at issue, the comparison class is different as well: if we shift to a discussion of realism about numbers, then salient theories of numbers populate the comparison class, not theories of mental states. This yields a general account of realism which goes as follows:

REALISM For any domain *D*, the comparison class for *D* determines a degree of fundamentality *d* such that a theory *T* is realist about *D* just in case *T* entails that *D* is fundamental to degree *d*.

1.5.3 Standard shifting

REALISM leaves open exactly which views might feature in the comparison class for a domain, and exactly how these determine a threshold for realism. There very well could be additional flexibility on both of these fronts. This makes available the outlines of a natural picture of what nonreductivists like Plantinga, Nagel, and Enoch are up to when they insist that only nonreductivists are realists.

First, they might be attempting to narrow the comparison class to include only views which take the domain at issue to be most fundamental. Then, *any* reductivist view will be a view that entails the domain to be much less fundamental than the views in the comparison class, and hence will not meet the threshold for realism. Second, they might be taking a stance on how the comparison class determines a threshold for realism, assuming that even when some reductivist views are in the comparison class, they fail to determine a threshold that sets a very relaxed standard. Absent a theory of how, in general, conversational context determines a comparison class, and how that comparison class determines a threshold, we have no means of assessing whether these theorists are *correct*, in their contexts, about where the threshold for realism lies. (It is, to my mind, neither obvious that they are correct, nor obvious that they are not.) After all, some attempts to change contextual parameters can fail. Nevertheless, we at least have a clear account of what these nonreductivists of why these speeches by nonreductivists are not making unintelligible mistakes. The variable threshold for realism makes their speech intelligible; whether it makes such speeches *true* is a separate (and unanswered) question.

1.6 Conclusion

REALISM, I have argued, represents a much-improved attempt to give a metaphysical account of the natural kind that underlies philosophical talk about 'realism'. Whether it provides a fully *satisfactory* account, given the parameters set out in §1, can be debated. There are two directions in which this question might be further pursued. First, its context-sensitive character belies a sense in which there is no *one* property use of 'realism' denotes. Rather, uses of the term with reference to different domains potentially refer to different degrees of fundamentality. All that prevents us from saying that 'realism' is equivocal across contexts is that which property 'realism' denotes in different contexts varies in semantically predictable ways. Whether this constitutes an account of *the* natural kind that underlies talk of 'realism' is an open question.[40]

A second question concerns the metaphysical character of the analysis. Given our gloss on the notion of a comparison class which sets the threshold for 'realism' in a context, facts about speakers' intentions and psychological states will play a large role in determining whether claims about realism are, in a context, true. This is a decidedly non-metaphysical aspect to the present analysis. But there is a robustly metaphysical fact underlying it: as we have presented the notion, facts about relative fundamentality are pure metaphysical facts, and in particular are not determined by anything like the contingent psychological states of speakers in a context. The point can be put in the following way (if we permit use of our earlier neologism): whether the claim that *A* is more realist than *B* is true is a pure, robust metaphysical fact. Nothing about the

[40] Note that, in answering this question, it is likely that we will not be able to *simply* ask whether context-sensitivity is a barrier to naturalness in an analysis. Surely *some* context-sensitive terms have highly unnatural analyses: if the Subjectivist view discussed in this chapter are construed along contextualist lines, the items 'wrong' applies to in different contexts will have very little in common, as what Sally disapproves of may be very different from what Sue disapproves of. But this isn't necessarily the case for every context-sensitive term: for instance, contextualists about 'knows that *p*' hold that the term expressed different properties in different contexts, but can still give a very natural account of what the instantiators of 'knows that *p*' in different contexts have in common. Or, perhaps 'species' in biology should be treated as context-sensitive, owing to the fact that different properties in the neighborhood are useful to different areas of biological investigation. Even if this is so, we shouldn't conclude that species is a highly unnatural kind. The question for REALISM, then, is whether it is similar in the relevant respects to the latter type of context-sensitive analysis.

attitudes or intentions of speakers even in part determines whether it is true.[41] But whether *A* in addition meets the threshold to be called realist *simpliciter* may depend, in part, on contingent psychological features of the context in which the ascription of realism is made. Again, I will leave whether this constitutes success as an open question.

To close, I will mention one more point in favor of an account along the lines of REALISM. Many philosophers have found the accounts in terms of existence, mind-independence, and absolute fundamentality to be very compelling accounts of realism. But, if the present objections are correct, these views fail for very straightforward reasons. What can explain their appeal? The REALISM-based account has a simple answer: existence, mind-independence, and absolute fundamentality often stand proxy for a greater degree of fundamentality.

Take existence views first. There are some cases where we, with good reason, restrict the theoretical roles we associate with a term to require that its referent be highly fundamental. One example is found in Schroeder (2005): with theological terms like 'God', one doesn't count as talking about the theological if one accepts a reductive account of their referents. Thus, if one identifies the referent of 'God' (as in Schroeder's example) with the strong nuclear force that holds positively charged protons in atomic nuclei together, one does not thereby count as someone who accepts that God exists.[42]

A very natural explanation of why this is so presents itself: there might be *some* truth-preserving interpretation of terms that feature in true God-claims like 'God is a person', 'God created the universe', 'God loves humankind', etc. But for the reductive

[41] Of course, this abstracts away from whether the truth of *A* or *B* depend on such facts. All I am claiming here is that the 'is more realist than' relation by itself introduces no dependency of this kind.

[42] See also Manley (2009). For Schroeder, this is because the theoretical role for 'God' requires of its referent that it bear certain properties—the referent of 'God' must be a person, must be capable of creating the world, must love humankind, etc. So we need to find referents for 'person', 'create', 'loves', etc. that apply to the strong nuclear force, but also preserve most of the ordinary truths about persons, creation, and love. Schroeder claims that identifying the referent of 'God' will eventually fail to preserve the truth of related claims involving 'person', 'create', 'loves', etc., and that the attempted reduction of the theological fails because of this. (See Schroeder (2005, 6).)

There is a worry about whether this is really what explains the failure of reductive accounts of the theological to be consistent with 'God exists' on its ordinary meaning. There are arguments in the literature inspired by Hilary Putnam (1980) that for any interpretation, there is a "permutation" of that interpretation which is truth-preserving. It isn't obvious that such a permutation isn't available to the theological reductivist who starts off by identifying the referent of 'God' with the strong nuclear force. Whether a relevant permutation is in fact available, however, seems irrelevant to the question of whether the strong nuclear force-theorist has a view on which God exists—of course she doesn't. It would then be desirable to find an explanation for why an interpretation that starts by assigning the strong nuclear force as the referent of 'God' will fail to yield a candidate English meaning of the term regardless of whether a Putnam-style permutation is available.

theorist who begins by assigning the strong nuclear force as the referent of 'God', such an interpretation will be highly gerrymandered and contrived. (Consider the interpretation of 'person': it must be a property that applies to not only the strong nuclear force, but also to the referent of 'human', but not to the referent of 'rock', 'planet', or 'number'. Such maneuvers will inevitably require a significant amount of gerrymandering in order to preserve the truth of many ordinary claims.) This suggests that theological terms are not only connected via theoretical role to claims expressed by 'person', 'create', 'love', etc.; the theoretical role attached to theological terms in addition requires that the properties of personhood, creation, love, etc. be *highly fundamental.* Since any interpretation that starts by assigning the strong nuclear force as the referent of 'God' will end up with not-very-fundamental referents for other terms that are closely connected to the theoretical role for 'God', such interpretations are ineligible. They cannot count as interpretations on which the English sentence 'God exists' is true.[43]

The question of theological realism then goes hand-in-hand with the question of the existence of the theological. But this is only because the theological can be expected to be fundamental, if it exists at all. Theological sentences like 'God exists' *require* that theological terms be assigned highly fundamental referents in order to come out true. This won't, however, be true for every domain for which the question of realism can arise. Discovering the constitution of polywater needn't show that polywater doesn't exist, as there need not be an expectation that any referent for 'polywater' is a highly fundamental one.[44] More generally, once the domain in question isn't one that can

[43] Whether this claim can be assimilated to a general fundamentality-maximization constraint is a question I leave open. Lewis (1984) and Sider (2012) both claim that, in light of Putnam-style permutations, we need to posit a general constraint on assignments of reference which requires that such assignments maximize fundamentality in assigned referents, in addition to other virtues. It is consistent with this that certain theoretical roles explicitly claim of their referents that they are highly fundamental—after all, it is not absurd to think that the strong nuclear force just isn't the right kind of thing to be the referent of 'God' in English, and that this is because it can't instantiate any very fundamental personhood property. This would place added weight on considerations of fundamentality in the specific case of the theological. Even if there isn't any general constraint on reference that requires fundamentality-maximization, it might be encoded in the theoretical roles for specific terms. I will not try to resolve these issues here, but rather point them out to illustrate how the connection between theological realism and fundamentality might be explicated in various ways.

[44] Note that there *might* be such an expectation: as we filled out Railton's example earlier, one might, prior to the relevant discovery, associate with 'polywater' a theoretical role that requires it to be of the same kind of molecular constitution as ordinary water. But, we emphasized, this isn't required: one might also continue to use the term 'polywater' with the same meaning after the discovery. One can consistently do this so long as one associates a less strict theoretical role with the term. This kind of case shows how, at least in principle, discovery of a not-very-fundamental reduction basis need not require a denial of the existence of the reduced domain or property.

be expected to be highly fundamental if it exists at all, it can still be properly said to *exist* even if it turns out to be highly gerrymandered and unnatural. In these cases, mere existence won't be sufficient for realism about the relevant domain. From the perspective of REALISM, the existence conception provides a sometimes (but not always) useful heuristic for when a view is realist.

REALISM also explains why the mind-independence and fundamentality accounts are tempting. Often, claiming something to be mind-dependent is to claim that it isn't very fundamental: after all, the mind-dependence claim itself is a claim that there is something more fundamental, namely the mind. For this reason, mind-dependence will in many cases give rise to a failure of REALISM; so long as there are other salient views which do not posit mind-dependence in the relevant domain, the view that does will come out as irrealist. Of course there are exceptions: when the domain in question is explicitly mental, mind-dependence no longer serves as a useful proxy for a lack of comparative fundamentality.

Finally, absolute fundamentality will always imply realism—views on which a domain is absolutely fundamental are guaranteed to be views on which the domain meets the contextually set degree of fundamentality required for realism. But the converse need not hold: some views which meet the contextually set degree of fundamentality required for realism need not be views on which the domain is absolutely fundamental. The absolute fundamentality-based view, like the existence and mind-independence-based views, provides in some cases a useful proxy for what is at issue in discussions of realism. But none of these views provide a complete picture; for this we need to appeal to the notion of relative fundamentality.[45]

[45] Additional thanks go to Ezra Keshet, Eliot Michaelson, David Plunkett, Timothy Rosenkoetter, Kenny Walden, and participants in the "Realism, Objectivity, and Meta-metaphysics" seminar at Dartmouth College for helpful comments and discussion.

Chapter 2: The Reference-Magnetic Solution to the Moral Twin Earth Problem

Many metaphysicians have thought that properties which are highly fundamental serve as "reference magnets". Roughly, this is the idea that highly fundamental properties are easier, in some sense, to refer to than their less fundamental neighbors. According to the "Moral Twin Earth" argument from Horgan and Timmons (1992), certain realist theories of ethics fail to account for the full range of semantic disagreement between possible users of moral language. The present chapter explores the prospects for a realist response to the Moral Twin Earth argument by appeal to reference magnetism.

2.1 Preliminaries

The Moral Twin Earth argument from Horgan and Timmons (1992) presents a serious challenge to a class of metaethical views that many philosophers have otherwise found attractive. The argument takes the form of a semantic challenge: it alleges that, from within a naturalist metaethical framework, any view on which moral terms are both (i) descriptive, and (ii) synthetic (in the sense that they do not have their natural referents by analytic necessity), is guaranteed to make unsatisfactory predictions. The predictions at issue take the form of disagreement relations between possible communities who use moral terms differently. Horgan and Timmons claim that any view satisfying (i) and (ii) will fail to predict that certain communities disagree when, intuitively, the communities in question *do* disagree.

There are a number of options for responding to an argument of this kind: one can, in the first instance, accept the consequence of the argument, but deny that this is a significant cost. One might do this by denying the existence of the disagreement intuitions in question. Or one could accept that failure to accord with the intuitions constitutes a theoretical cost, but maintain that it is a cost that is outweighed by other theoretical benefits of synthetic naturalism. A second approach one could take is to accept that there is genuine disagreement between the communities in question, and accept that it must be accounted for in some way. What one denies on this approach is that the disagreement must be accounted for with purely semantic tools. That is, if Horgan and Timmons are right, a synthetic naturalist semantics for moral terms cannot explain all of the relevant instances of disagreement. But this doesn't rule out the possibility of other explanations—one could, for example, appeal to the notion of "disagreement in attitude" from Stevenson (1937) to carry out the explanation. There is, finally, a more ambitious type of response: one could claim that Horgan and Timmons fail to properly assess the resources synthetic naturalism can bring to bear in explaining the relevant disagreements. It is last type of response to the Moral Twin Earth argument that I will be exploring in this chapter.

More specifically, I will be focusing on a broad kind of strategy for responding to the argument which claims that, because certain properties are *highly eligible* for reference, there is a semantic explanation for the disagreement. That is to say: communities who use moral terms differently nonetheless manage to refer to the same natural property with their moral terms, and as a result are capable of disagreeing about its pattern of instantiation. Call this an *eligibility response.* A number of papers in the literature on the Moral Twin Earth argument propose eligibility responses; these include Sayre-McCord (1997) and van Roojen (2006).[46] They all face the same challenge: they must give a plausible explanation of the mechanism by which certain natural properties highly eligible for reference by moral terms.

To articulate the eligibility response I will be exploring, I will in §2 explain the Moral Twin Earth argument in more detail. Then in §3, I will argue that none of the existing eligibility responses are promising attempts meet the challenge: in particular, they fail to present a general, independently motivated account of reference that the synthetic naturalist can then apply in a straightforward way to the case of moral terms. The remaining sections of this chapter aim to outline an eligibility response that lacks this defect. §4 takes the first steps toward such an account by introducing a general account of when properties are highly eligible for reference: on this account, properties are highly eligible when they are *highly fundamental* in a sense familiar from the metaphysics literature. I will call properties that are highly eligible for reference in virtue of being highly fundamental *reference magnets.* Reference magnetism is therefore a particular kind of eligibility which is fully general—it can, in principle, feature in an account of which property *any* kind of term refers to; its scope isn't limited to just moral terms. The task for the synthetic naturalist is then to explain why it is that moral properties are highly fundamental and therefore reference magnets. This is the aim of §5: there, I outline

[46] Copp (2000) also provides a version of what appears to be an eligibility response, but for reasons of space I do not discuss it here.

a particular conception of degrees of fundamentality from which it follows that on the synthetic naturalist view moral properties are reference magnets. While this account has the virtue of giving a fully motivated eligibility response to the Moral Twin Earth argument, it also highlights the costs of such a view: first, it requires that a phenomenon like reference magnetism exists—a claim that many have doubted. And second, it requires a specific kind of account of the notion of metaphysical fundamentality that underlies reference magnetism. These aspects of the account are not uncontroversial, and to the extent that they are objectionable, the viability of eligibility responses to the Moral Twin Earth argument should be called into question.

Before proceeding, we should pause to consider the relationship between the various strategies for responding to the Moral Twin Earth argument. As I have indicated already, my focus will be on developing a kind of eligibility response. But "giving an eligibility response" is not, strictly speaking, an on-or-off matter: it is possible that eligibility explains *some* but not *all* of the supposed disagreement between possible communities who diverge in their use of moral terms. In this case, the remaining instances of supposed disagreement might be dealt with by other means—e.g., by rejecting the relevant intuitions in some way, or by explaining them with non-semantic tools. Thus an eligibility response might have an important role to play in addressing the Moral Twin Earth argument, even if it doesn't constitute a complete response to the argument all by itself. I will, in what follows, leave this possibility in the background. The reference-magnetic account I develop should then be understood as a promising proposal for those cases which should be explained in terms of an eligibility response. But this should not be taken to be a commitment to the stronger claim that reference magnetism explains *all* of the supposed data surrounding the Moral Twin Earth argument. I will not be taking a stance on this latter question because I will not be answering the larger question of the extent to which synthetic naturalists need eligibility to explain the supposed disagreement data.

2.2 The Moral Twin Earth argument

Horgan and Timmons take the Moral Twin Earth argument to show that any metaethical view according to which the semantics for moral terms is (i) descriptivist, (ii) naturalistic, and (iii) synthetic, is untenable.

We can call a metaethical view *descriptivist* just in case it treats the semantics for moral sentences to be of the same kind as the semantics for ordinary sentences like ‘grass is green’.[47] I won’t say much more about what requirements this places on the semantics for moral terms, mostly because it is controversial how the proper semantics for ordinary descriptive sentences should go. All I will assume is that, in the case of moral predicates like ‘wrong’, ‘good’, etc., it makes sense to talk of them having a “referent”, which (I will assume) is a property that is instantiated by all of the possible things the predicate is true of.

A theory is *naturalistic*, we can say, just in case it holds that the referents of moral predicates are natural properties. There is a large literature on the question of what, exactly, naturalism is. This condition might not be *sufficient* for ethical naturalism either: it isn’t obvious that non-naturalists in the style of Moore (1903) are committed to rejecting that wrongness is identical to a natural property, so long as they deny the further claim that wrongness is *most fundamentally* a natural property.[48] I will bypass these complicating issues in what follows, settling instead for an intuitive working grasp of which properties are paradigmatically “natural”, and how moral properties must be related to them if synthetic naturalism is true.

Finally, a naturalist metaethical theory is *synthetic* just in case it does not entail that there is something every speaker competent with ‘wrong’ knows about the relationship between the moral to the natural in virtue of her competence with ‘wrong’. To put it concisely, we can say that according to synthetic theories, it is not the case that every competent speaker knows which natural property wrongness is. If wrongness is failing to maximize happiness (say), then someone who is competent with ‘wrong’ need not know that wrongness is the property of maximizing happiness. These ‘knows what property’ expressions are tricky, and are probably not the best language for expressing a fundamental commitments of a major metaethical view. We can, however, get a better grasp of the intended sense by analogy: speakers competent with ‘wrong’ need not know which natural property wrongness is in the way that speakers competent with ‘bachelor’ know which property bachelorhood is (namely, the property of being an unmarried man). In what follows, I will summarize this (roughly characterized) commitment of synthetic theories by saying that a competent speaker need not know what ‘wrong’ in her mouth refers to.

[47] Horgan and Timmons use the term ‘cognitivist’ here.

[48] See the introduction of Chalmers (1996) for one (controversial) proposal about the commitments of naturalism; and see Chapter 4 for more on non-naturalism about the normative.

2.2.1 A simple case: causal theories of reference

In their first statement of the argument, Horgan and Timmons (1992) describe the Moral Twin Earth argument as an argument against a particular instance of synthetic naturalism. The target theory is that of Boyd (1989) which holds that 'wrong' as used by a linguistic community refers to the natural property which "causally regulates" its use in that community. Without going into the details of what causal regulation amounts to, it is clear why such a theory is synthetic: speakers competent with 'wrong' might nonetheless not know *which* natural property causally regulates the use of 'wrong' in their community. The facts about causal regulation are discovered empirically, and a speaker can continue in her competent use of 'wrong' without acquiring the relevant empirical information.

The features of the Boyd-style causal theory which make it a synthetic view are also the precise features which make it vulnerable to the Moral Twin Earth argument. Horgan and Timmons make this case using the following description of a pair of communities:

> Suppose [...] Earthlings' moral judgments and moral statements are causally regulated by some unique family of functional properties, whose essence is functionally characterizable via the generalizations of a single substantive moral theory [...] For specificity, let this be some sort of consequentialist theory, which we will designate T^c.
>
> Now for Moral Twin Earth. Its inhabitants have a vocabulary that works very much like human moral vocabulary; they use the terms 'good' and 'bad', 'right' and 'wrong', to evaluate actions, persons, institutions, and so forth [...] But on Moral Twin Earth, people's use of twin-moral terms are causally regulated by certain natural properties distinct from those that (as we are already supposing) regulate English moral discourse. The properties tracked by twin English moral terms are also functional properties, whose essence is functionally characterizable by means of a normative moral theory. But these are *non-consequentialist* moral properties, whose functional essence is captured by some specific deontological theory; call this theory T^d.[49]

Horgan and Timmons go on to note that, according to Boyd's theory, the inhabitants of Earth and Moral Twin Earth refer to *different* natural properties with their use of moral terms. Hence, when one community asserts **Wrong** and the other asserts **Not Wrong**, they assert propositions that are *consistent* with each other:

[49] Horgan and Timmons (1992, 245)

Wrong ϕ-ing is wrong;

Not Wrong It is not the case that ϕ-ing is wrong.

The reason is straightforward: supposing P^c is the property characterized as wrongness by T^c, and that P^d is the property characterized as wrongness by T^d, $P^c \neq P^d$. According to the Boyd-style theory, then, since P^c causally regulates Earthlings' use of 'wrong', they refer to P^c. Hence, what Earthlings say by asserting **Wrong** is true just in case ϕ-ing instantiates the property P^c. P^d causally regulates Twin Earthlings' use of 'wrong', so they refer to P^d. Hence, what Twin Earthlings say by asserting **Not Wrong** is true just in case it is not the case the ϕ-ing instantiates P^d. But ϕ-ing might both instantiate P^c *and* not instantiate P^d—they are, after all, distinct properties. In this case, the assertion of **Wrong** by the T^c-community is *consistent* with the assertion of **Not Wrong** by the T^d-community. The two communities, we can say, fail to *semantically disagree*.

Call such a pair of communities—that is, any pair of communities that, according to at given theory of reference, fail to semantically disagree—an *MTE pair* for the theory. The existence of an MTE pair for the Boyd-style theory of reference is a mark against it. At least, the existence of the pair is a mark against the theory *if* we find it intuitive that Earthlings and Twin Earthlings do disagree, and *if* we decline to explain the disagreement in non-semantic terms. I will grant this assumption for the sake of exploring eligibility-responses, as I explained in §1.

It is not surprising that we were able to find an MTE pair for the Boyd-style theory: as we noted earlier, it is a synthetic theory precisely because a competent speaker can fail to know which natural property causally regulates her use of 'wrong'. She can fail to know this because knowing which property her use of 'wrong' refers to requires empirical information—in this case, it is information about the causal relations between various properties in the world and her use of 'wrong'. And since information about what property causally regulates her use of 'wrong' is empirical information, it is *false* with respect to other possible worlds. A different property might have causally regulated her use of 'wrong'. In short, the Boyd theory is synthetic because there is another world where speakers are such that their use of 'wrong' is causally related to a different property. This embodies a substantive (though in this case plausible) assumption, which we can call the *Ignorance-Existence Assumption*, or IE for short:

IE If in $w_@$ one needs to learn some information in order to know which property causally regulates one's use of 'wrong', then there is a possible world $w_\dagger$, compat-

> ible with what one knows in $w_@$ in virtue of being competent with 'wrong', where speakers' use of 'wrong' is regulated in $w_\dagger$ by a property that is *distinct* from the property that regulates use of 'wrong' in $w_@$.

The community in $w_\dagger$ is then the community which Horgan and Timmons will use as the second community in an MTE pair. In the example cited above, the Twin Earthlings constitute the second community in the MTE pair for the Boyd-style theory. IE makes it clear that it was no accident that Horgan and Timmons could find such a community: given the empirical nature of the facts about causal regulation, there *must* be a possible second community whose use of 'wrong' is causally regulated by a different property. This community can then feature in an MTE pair for Boyd's theory.

This second community of the MTE pair, Horgan and Timmons are willing to bet, will be a community that speakers in $\mathrm{w}_@$ intuitively *disagree* with.[50] The Boyd-style theory thus gains the virtues of synthetic naturalism at the cost of failing to account for the intuitive disagreement data between possible linguistic communities.

2.2.2 Generalizing the argument

But the Moral Twin Earth argument is supposed to be a *general* argument, refuting any version of synthetic naturalism. So far, we have only shown how the argument goes for one instance of this kind of view. What we have said so far, however, gives us the resources to see why Horgan and Timmons think the argument generalizes, allowing us to show that for any version of synthetic naturalism, there is an MTE pair. The generalization rests on the assumption that every version of synthetic naturalism will count as synthetic *for the same reason* that the causal regulation theory is synthetic.

Horgan and Timmons (2000, 147) say:

> We challenge the moral naturalist to tell us a story (or at least sketch one) about how reference is fixed for moral terms as employed by such speakers (which will involve specifying some relation *R* which supposedly fixes determinate reference-relations between moral terms and certain natural properties) [...] [W]e will concoct a Twin Earth scenario featuring *R* in

[50] As Mark Schroeder has pointed out to me, this is not a trivial assumption: Horgan and Timmons picked a case where we have antecedent, independently motivated reasons for thinking there is a disagreement by calling the Earthlings consequentialists and the Moral Twin Earthlings deontologists. But the general principle is less compelling: it requires that, in any case where *s* must acquire empirical information to rule out certain possibilities before *s* can know what property *s*'s use of 'wrong' refers to, *at least one* of the ruled-out possibilities *r* must be such that speakers in *r* intuitively disagree with *s* with their use of 'wrong'.

> which another group's use of moral terminology is R-related to a distinct set of natural properties. This story will reveal that the reference-fixing story on offer leads to objectionable relativism [i.e., there is a pair of possible communities which constitute an MTE pair for the relevant theory].

Horgan and Timmons thus are not assuming any special view about what the reference-determining relation R is: they are not, in particular, assuming that it is a *causal* relation (which the argument against the Boyd-style theory does assume.) What they are assuming is that the target view to be a version of synthetic naturalism. And this, they assume, means that we can then construct an MTE pair for the view, by an argument that takes exactly the same steps that we took in §2.1.

If the theory on which R is the reference-relation is synthetic, a competent speaker can fail to know which natural property bears R to her use of 'wrong'. She can fail to know this because knowing which property her use of 'wrong' refers to requires information—in this case, it is information about the R-relations between various properties in the world and her use of 'wrong'. Since information about what property bears R to her use of 'wrong' is (Horgan and Timmons assume) empirical information, there are other possible worlds where a different property bears R to her use of 'wrong'. In short, the synthetic theory is synthetic *because* there is a world where speakers are such that their use of 'wrong' bears R to a different property; in order to know which natural property her use of 'wrong' refers to, she must acquire empirical information which rules out worlds where different properties bear R to her use of 'wrong'. This embodies an assumption which is a generalization of IE, and which we can call IE^{+}:

IE^{+} If in $w_{@}$ one needs to learn some information in order to know which property one's use of 'wrong' bears R to, then there is a possible world $w_{\ddagger}$, compatible with what one knows in $w_{@}$ in virtue of being competent with 'wrong', where speakers' use of 'wrong' in $w_{\ddagger}$ bears R to a property that is *distinct* from the property use of 'wrong' bears R to in $w_{@}$.

Given IE^{+}, there must be a possible community that refers to a different natural property with their use of 'wrong'. This is the community in $w_{\ddagger}$, and they will form the second community of an MTE pair for the synthetic theory in question. Horgan and Timmons are willing to bet, as before, that such a community will be a community speakers in $w_@$ intuitively disagree with. Thus any version of synthetic naturalism will be vulnerable to the Moral Twin Earth argument.[51]

2.2.3 Vulnerabilities in the Moral Twin Earth argument

We noted at the outset a number of ways in which one might respond to the Moral Twin Earth argument. *Eligibility responses*, which I will focus on for the remainder of this chapter, reject the Ignorance-Existence Assumption in its general form, IE^+. Here I will sketch *why* this assumption might be vulnerable by making explicit what, exactly, it is committed to.

IE^+ embodies a claim about the relationship between (i) needing to learn something in order to know what a term (in this case, 'wrong') refers to, and (ii) the distribution of possible linguistic communities across modal space. The connection with metaphysical modality in (ii) is crucial: MTE pairs are pairs of *possible* communities.[52]

But IE^+ is vulnerable for precisely this reason: from the fact that speakers competent with 'wrong' must learn *something* before they can know which property their use of 'wrong' refers to, it is a large and controversial jump to the claim that they must learn something *which rules out a metaphysically possible world*. There is room for ground in between: in particular, a speaker might need to learn something more in order to know what 'wrong' in her mouth refers to, but need not learn something that rules out possible worlds that were previously consistent with what she knows. There are two ways in which this can happen.

First, a speaker might learn something by deducing it a priori. There are many truths that can be deduced without relying on empirical information, and many of these truths rule out *no* possible worlds.[53] Second, a speaker might deduce truths from

[51] I owe this way of understanding the crucial components of the Moral Twin Earth argument to Mark Schroeder.

[52] To see why this point is crucial to the Moral Twin Earth argument, suppose the argument only showed that there are metaphysically impossible but conceptually possible communities who form the second element of an MTE pair for a given theory. This amounts to the following claim: the use of 'wrong' by these merely conceptually possible communities bears the reference-fixing relation R to a different property than the property our use bears R to. But (the hypothetical objection continues) such communities intuitively *disagree* with each other in their use of 'wrong'. Given the supposition that the community in question is conceptually, but not metaphysically possible, this argument would only show that the concepts of semantic disagreement and the reference-fixing relation R are different. But this is something that synthetic naturalist would be ready to grant: she isn't offering a conceptual analysis of disagreement, but rather a substantive theory of the relation that makes semantic disagreement possible in the relevant metaphysically possible worlds. Only an MTE pair where both communities are metaphysically possible is a threat to this thesis.

[53] There may be exceptions if contingent a priori knowledge is possible. (See Hawthorne (2002) for more

things she already knows in virtue of having some empirical information. In this case, what she learns by the deduction does not itself rule out any further possible worlds, as such worlds were already ruled out by the body of knowledge from which the inference started. One body of empirical information that might be especially important in this respect is the information available to speakers in virtue of their being competent with 'wrong'.

The existing eligibility responses I will discuss in the next section provide positive proposals about the reference of 'wrong' that are inconsistent with IE^+. These proposals are all accounts on which speakers' lack of knowledge about the reference of their use of 'wrong' is not the kind of ignorance that can be eliminated only by acquiring further information which rules out possible worlds where use of 'wrong' refers to a different property. The trick for these responses is to give an independently motivated account of why it is that coming to know what property 'wrong' refers to doesn't require learning this kind of information.

2.3 Eligibility and analyticity

It is common for eligibility responses to draw on some analogy with the ways in which natural kind terms acquire their referents to explain why moral vocabulary as the synthetic naturalist construes it is not vulnerable to the Moral Twin Earth argument. This provides the needed motivation for the synthetic naturalist's semantic claims. The underlying thought is that natural kind terms like 'water' have referents that are, in an intuitive sense, *highly eligible*. In particular, speakers manage to refer to the kind H_2O even if they don't know that it does, and even if they mis-use the term to a substantial degree. Thus: even if I, along with everyone in my linguistic community, don't have any knowledge of contemporary molecular theory, and hence don't even know what H_2O is, we might still succeed in using 'water' to refer to it. Likewise, even if we sometimes apply 'water' to some clear liquid substances that are *not* H_2O, these uses count as *mis*applications of the term—'water', in our mouths, still refers to H_2O. Natural kind terms have the peculiar feature that they can refer to a kind in the absence of an intention of the part of speakers to refer to that specific kind, and even in the presence of misapplications of the term. We can treat features like these as the marks of properties that are highly eligible for reference.

on this kind of view.) Even if this is so, some a priori knowledge certainly fails to rule out a possible world, in those cases where it is not knowledge of a contingent truth.

Natural kind terms, however, are something of a double-edged sword for the synthetic naturalist: terms like ‘water’ are precisely the motivation behind Horgan and Timmons’ Moral Twin Earth argument (including its name), and in particular provides the paradigmatic example behind the general premise IE^+. Which property our use of ‘water’ refers to is not an analytic fact: knowing it involves, on the standard story, the empirical discovery that water is H_2O. But this very discovery rules out another possible world: in particular, it rules out the world where the similar substance XYZ is the natural kind in the vicinity of speakers’ use of ‘water’. The XYZ-world is a “Twin Earth” where a use of ‘water’ similar to ours refers to a different property. Speakers competent with ‘water’ don’t necessarily know that ‘water’ refers to H_2O because they don’t necessarily have information which rules out the XYZ-world. Horgan and Timmons assume that by treating moral terms on analogy with ‘water’ and not analytically referring to a particular property, similar claims apply: there should be a “Moral Twin Earth” where a use of ‘wrong’ similar to ours refers to a different property. Speakers competent with ‘wrong’ don’t necessarily know what property ‘wrong’ refers to because they don’t necessarily have information which rules out this world. The assumption of IE^+ embodies the assumption that any attempt to treat a class of expressions as synthetic requires the existence of a Twin Earth possibility of this kind.

There are then two aspects to natural kind terms: one is the high degree of eligibility in their referents, which allows reference to succeed in the presence of substantial ignorance and misapplication. The other is that claims about which property a natural kind term refers to are synthetic; this is often explained by the existence of possible worlds where a community who uses the term with the same meaning refers to a different property. These two aspects of natural kind terms are, in principle, separable. The trick for eligibility responses is to explain eligibility in a way that motivates not taking on commitment to IE^+. Existing versions have failed to successfully navigate these waters.

2.3.1 Sayre-McCord: moral kinds

Sayre-McCord (1997) provides one attempt to make use of the eligibility of the referents of natural kind terms without taking on board objectionable claims about the modal properties of the reference-relation. Which natural properties constitute natural kinds, he says, is a matter that is settled by scientific investigation. (Thus, we arrive at the conclusion that H_2O is a kind, and hence eligible for reference by our use of ‘water’, because ‘H_2O’ has a place in our best chemical theory.) Scientific investigation tells us which properties are the highly eligible natural kinds. Since scientific investigation is empirical and a posteriori, there are worlds where its claims are false—hence, discovering which properties are the natural kinds requires acquiring information which rules out other possible worlds. But it is a mistake to think, Sayre-McCord says, that natural kinds are the only properties that are highly eligible:

> The kinds that matter to morals are not those of natural science but those that are countenanced by the best moral theory. Whether the Twin Earther's term 'good' refers to what our term does depends on whether their use of the term is causally regulated, in the appropriate way, by what is good, and it may well be so regulated even if all the instances of goodness with which they come into contact are quite different in other respects.[54]

Thus Sayre-McCord wants an account on which the referents of moral terms are eligible in the same way natural kinds are—he still calls them "kinds"—but on which the non-analytic facts about reference are, unlike the case of natural kinds, discovered by means other than empirical scientific discovery. The last component is crucial: the eligibility response needs it in order to avoid the modal implication that there are possible worlds where speakers refer to a different kind with 'wrong'.

The relevant kinds—which are appropriately eligible, but not empirically discovered—are *moral* kinds in Sayre-McCord's terminology. To the question of whether two instances might be of the same moral kind, he says, "it is irrelevant whether they are, in other respects (say, from the point of view of science), all of the same kind"—hence, on this view, there is very little to be said about what natural and moral kinds have in common.[55] So the crucial claim of Sayre-McCord's eligibility response is the following:

The properties that are singled out by moral theorizing as the properties that constitute wrongness, goodness, etc. are, like natural kinds, highly eligible for reference.

Supposing, then, that Horgan and Timmons' T^c-community uses the correct moral theory, the property singled out by T^c as wrongness constitutes a moral kind. The T^d-community then uses 'wrong' in a way that is regulated by the *same* moral kind, since that is the highly eligible *moral* kind in the vicinity of their use of 'wrong'. Hence they count as referring to the same property as the T^c-community, and are capable of semantically disagreeing with them. This is an instance of an eligibility response, as it in

[54] Sayre-McCord (1997, 284)

[55] Sayre-McCord (1997, 285)

principle allows that every pair of communities which intuitively disagree in their use of 'wrong' all refer to the same highly eligible moral kind and hence genuinely disagree.

This account has the makings of an account of semantic disagreement between communities that use moral language differently. What it lacks is a *motivation* for the claim that moral properties, despite *not* being natural kinds, should still share the feature of being highly eligible for reference with natural kinds. For all that Sayre-McCord says, there is nothing moral properties and natural kinds have in common which explains why they are both highly eligible for reference. (Note that the fact that he has chose to designate both with the term 'kind' provides nothing in the way of explanation on this front.) The moral kinds account is convenient as a description of the kind of account the eligibility response needs to give—it needs to explain the eligibility of moral properties without taking on board every feature of paradigmatic natural kinds—but does so without providing anything in the way of constructive explanation.

2.3.2 Van Roojen: naturalness

A related eligibility response in van Roojen (2006) improves on the modal kinds approach in one respect. The reference of 'wrong' is determined, according to van Roojen, by the *naturalness* of candidate referents, in the sense of Lewis (1983). Moral Twin Earth cases are, in van Roojen's words, cases where "a particular judgment will be best interpreted as involving mistakes about a natural kind rather than knowledge of an unnatural kind".[56] This provides a partial answer to a question Sayre-McCord leaves unanswered: where Sayre-McCord says nothing about what it is about moral kinds that makes them highly eligible for reference in the same way as natural kinds, van Roojen does. Both moral properties and natural kinds are *highly natural* in the Lewisian sense.

Lewis (1983) introduces naturalness as a metaphysical notion which captures the thought that some properties—perhaps those instantiated by fundamental microphys-

[56] van Roojen (2006, 185). Van Roojen's account is actually two-pronged, though I am focusing only on the second prong. The referent of 'wrong' as used by a linguistic community, he says, is determined in part by what would maximize the *knowledge* of speakers in the community. Supposing that wrongness on the correct moral theory coincides with property *P*, and that this theory guides the use of the T^c-community, the T^d-community will not know various facts about wrongness if their use of 'wrong' is interpreted as referring to a different distinct from *P*.

The knowledge-maximization constraint doesn't all by itself explain why the T^c and T^d communities disagree, and this is why van Roojen introduces the naturalness component to his account. Given a sufficiently liberal conception of properties, there are *many* properties that are plausibly knowledge-maximizing. (Consider all of the properties that have in their extension all of the wrong actions *except* for one instance that no one can, or at least will, form any beliefs about whatsoever.) What distinguishes the moral property from these nearby gerrymanders is the superior naturalness of the moral property.

ical entities—are highly elite, basic properties. They are to be distinguished from non-elite, derivative properties: some things have the property of being a tree, for instance, but have this property only in virtue of some distribution of microphysical particles and their fully elite properties. So being a tree is not a (perfectly) natural property. (There may, however, be *degrees* of less-than-full naturalness, and Lewis explicitly explains how his theory allows for this. I will return to this issue in the next section.)

For the synthetic naturalist, wrongness is not plausibly among the perfectly natural properties in Lewis's sense. Neither are the natural kinds that constitute the paradigmatic examples of properties that are highly eligible for reference. So how are considerations of naturalness supposed to play a role in determining the reference of a linguistic community's use of 'wrong'? Here van Roojen departs from Lewis by introducing a *discipline-relative* notion of naturalness. He says:

> Naturalness should be seen as discipline-relative. The kinds or properties which are more natural for the purposes of physics may not be the same as those which are more natural for purposes of biology. The more eligible semantic values for one's terms when engaged in the former may or may not be the same as the more eligible semantic values for one's terms when one is engaged in the latter [. . .] Tables seem perfectly natural for anthropological purposes, but not for the purposes of physics.[57]

The most salient problem for discipline-relative naturalness in the present setting is that it is most natural not to think of it as one single notion which is shared by many different properties. Instead, it is most naturally treated as a label for many different features—one for each of the disciplines naturalness might be relative to. Hence there is *naturalness relative to physics*, *naturalness relative to anthropology*, and *naturalness relative to ethics*.[58] This is particularly problematic in the present setting, because we are after an account on which moral properties are analogous to natural kinds in a very particular respect. This would be an account on which natural kinds are highly eligible for reference because they have a certain feature, which moral properties also have. Van Roojen's initial appropriation of Lewis's metaphysical notion of naturalness was more

[57] van Roojen (2006, 181)

[58] Strictly speaking, I am ambivalent over whether we say discipline-relative naturalness is a "single" feature shared by many properties or not. There are different ways to use the word 'feature' (and its cognates), and one of them allows for there to be "features" with disjunctive, highly gerrymandered extensions. I am content to call discipline-relative naturalness a "feature" in this sense, the fact that the various instances of discipline-relative naturalness share a feature *in this sense* won't do any explanatory work.

promising, in this respect: if natural kinds are highly eligible for reference in virtue of being perfectly natural, then moral properties are highly eligible if they too are perfectly natural. Such an account is, as we noted, unpromising for other reasons, but a move to discipline-relative naturalness renounces its main virtue. Even if water is highly eligible because it is natural *relative to chemistry*, this by itself tells us nothing about whether wrongness is highly eligible for reference. Wrongness is not natural *relative to chemistry*, but rather natural *relative to ethics*.

Absent a further theory of what properties that are natural relative to different disciplines have in common, the discipline-relative naturalness of ethics is unpromising.[59] Like Sayre-McCord's use of 'moral kind', it at best provides a unified practice of naming for what might be a highly disparate set of phenomena.[60] This is a significant failing in the account, but it points the way forward. I will explore, in the next section, the prospects for a theory that takes something like Lewis's metaphysical (and non-discipline-relative) notion of naturalness to explain eligibility of both natural kinds and moral properties. If this can be made to work, it promises a constructive (and more-than-merely-verbal) account of what makes both natural kinds and moral properties highly eligible for reference.

2.4 Fundamentality and reference magnetism

2.4.1 Metaphysical fundamentality: what

Perfect naturalness, in Lewis's sense, is a version of the notion of metaphysical fundamentality that is at the center of much recent research in metaphysics (or

[59] One thing such an account would have to do is to explain why the physics- and ethics-relative notions have something in common and both make for eligibility, but why naturalness relative to a possible discipline like anthropology*, which doesn't talk about tables, artifacts, etc., but rather the intuitively gerrymandered tables*, artifacts*, etc. does not give rise to eligibility. (I assume that tables* and artifacts* are not highly eligible for reference.) Thanks to David Manley for suggesting this point.

[60] One way to pursue this complaint further would be to point to the different *kinds* of explanation that different disciplines are concerned with. Some disciplines are explicitly concerned with *causal* explanation: for instance, economics might appeal to a property to explain the causes of inflation. Others employ descriptive generalizations: half-life is an explanatory property in atomic theory that summarizes probabilistic properties of atoms. And for ethics, the explanatory properties will neither feature in causal explanations nor be probabilistic. Once we focus on these differences, the notion of discipline-relative naturalness would appear to be a highly disjunctive kind, where being natural relative to one discipline is a very different kind of thing from being natural relative to another. (This should especially give us pause when considering the thesis that highly eligible natural kinds are natural relative to (say) chemistry, moral properties are therefore highly eligible because they are natural relative to ethics.)

I do not want to claim here that there is *no* possible account to be given here, but it is plausible that if one is forthcoming, the account of eligibility can proceed directly in terms of the account. We can then avoid the detour through the notion of discipline-relative naturalness.

*meta*metaphysics). In addition to the use of “naturalness” in Lewis (1983), Fine (2001, 2009) employs the notion of “Reality”, Sider (2012) uses “Structure” and Schaffer (2009) develops the notion of “ground”. I will gloss over the differences between these accounts and use the term ‘metaphysical fundamentality’ (or ‘fundamentality’ for short) to refer to the explanatory role that is common between them. Fine (2001, 22) captures the explanatory character the most fundamental by saying “in providing the [fundamental] ground for a given proposition, one is explaining, in the most metaphysically satisfying manner, what it is that makes it true”. There are important issues about what the primary bearers of fundamentality are (whether they are facts, properties, entities, etc.), and the character of the explanatory relationship. I will attempt to remain neutral on these questions for present purposes.

2.4.2 Relative fundamentality and reference magnetism

It is common for those working within a fundamentality-based framework to hold that those properties which are highly fundamental are thereby highly eligible for reference. I will, following common usage, label this phenomenon *reference magnetism*.[61]

Here is a first pass at a characterization of reference magnetism. Given a pattern of use of a term t by a community and a set of candidate referents $e_1 \ldots e_n$, the referent of t is partly determined by whichever of $e_1 \ldots e_n$ scores best on considerations of both (i) fit with the community’s use of t, and (ii) fundamentality. The basic idea is that if e_i is fundamental but e_j is not, e_i might be the referent of t even if e_j better fits with its use. Reference magnetism so described holds that fundamentality *might* suffice for making e_i the referent; whether it *does* will depend on many details: in particular, which entity in the end qualifies as the referent of t depends on *how much* better e_j fits with use, the precise scoring rules for balancing fundamentality against degree of fit, and the extent to which an assignment of e_i over e_j as referent for t maximizes balance of fit and fundamentality for terms across the entire language. These details matter for the purposes of arriving at a final theory, but even without delving into them we can see that this first-pass characterization of reference magnetism is not adequate.

The phenomenon of reference magnetism has been claimed to resolve purported the purported massive indeterminacy in reference that has been alleged, for various

[61] See Sider (2012, Ch.3) and Hawthorne (2007). On the way I am using the terms, reference magnetism is a species of eligibility for reference, but is to be distinguished from the general notion of eligibility. The general notion implies no account of *which* entities, properties, etc. are highly eligible. Reference magnetism, by way of contrast, is committed to an account on this front as it is committed to metaphysical fundamentality explaining eligibility.

reasons, in Putnam (1981), Kripke (1982) and Quine (1960). But it is obviously a non-starter if, in keeping with our original formulation, we limit our attention to the binary distinction between what is *fully* fundamental and what is not. For the allegations of indeterminacy depend in no way on indeterminacy in terms standing for fully fundamental entities, properties, etc. Putnam's "model theoretic argument", for instance, is stated for terms like 'cat' and 'tree', as it proceeds by devising permuted interpretations which fit with use sufficiently well, but which assign highly bizarre interpretations to 'cat' and 'tree'.[62] So there is no hope of avoiding massive indeterminacy in reference by appeal to reference magnetism as characterized above; even the non-bizarre referents for 'cat' and 'tree' are not fully fundamental.

What I take this to show is that the motivations for reference magnetism are worth considering only if they appeal not to fundamentality *simpliciter*, but rather to *degrees* of fundamentality. The solution to Putnam's argument in Lewis (1984) explicitly assumes this, as there Lewis appeals a *degreed* conception of naturalness. (He calls this "less-than-perfect naturalness".) I will discuss the Lewisian account in more detail below, but for now it will suffice to point out that a degreed conception will allow us to say, for instance, that some things can be *more* fundamental than others without being fully fundamental. I will call this degreed conception *relative fundamentality*. The use of "relative" here, however, is to be sharply distinguished from van Roojen's talk of "discipline-relative" naturalness. The latter is a binary distinction that holds relative to specific disciplines. Relative fundamentality, as I will use it, is a gradable notion that holds absolutely (i.e., is *not* to be relativized to discipline).

Relative fundamentality is more promising than its absolute counterpart in an account of reference magnetism. Even if trees and cats are not fully fundamental, they are plausibly *more* fundamental than the referents assigned by a Putnam-permuted interpretation. This motivates a more refined conception of reference magnetism, as follows:

REFERENCE MAGNETISM Given a term t and candidate referents $e_1 \dots e_n$, each of which (i) fits the use of t to some degree, and (ii) is fundamental to some degree, the reference of t is that of $e_1 \dots e_n$ which best balances degree of fit and degree of fundamentality.[63]

[62] See Putnam (1981, Ch. 2).

[63] As before, this is a simplification that needs to be refined to account for the rules to balance degrees of fit and fundamentality, and needs to be generalized to take into account the way in which reference is determined by maximizing fit and fundamentality across an entire language, and not on a case-by-case

That is: a greater degree of fundamentality privileges one extension over others in an assignment of referents.[64]

2.4.3 The nature of relative fundamentality

Recall the problem that led van Roojen to move to a "discipline-relative" conception of naturalness: ethical properties, according to the naturalist view, are not elite, perfectly natural properties (or, in the terms of the present section, are not fully fundamental). Consequently, they cannot owe their eligibility for reference to the fact that they are perfectly natural. Van Roojen then tries to explain the eligibility in terms of discipline-relative perfect naturalness: both natural kinds and ethical properties are natural relative to their respective disciplines. We criticized this move on the grounds that it does nothing to explain why properties that are natural-relative-to-ethics are highly eligible. (Paradigmatic natural kinds are, after all, not natural-relative-to-ethics, but rather natural-relative-to-chemistry or some other empirical discipline.) REFERENCE MAGNETISM suggests a well-motivated and promising alternative: if eligibility isn't a matter of being *fully* fundamental but rather only fundamental to a *significant* degree, natural kinds and moral properties might turn out to share some feature in which they are both highly eligible for reference. If both kinds of property are highly fundamental (without, of course, being fully fundamental), then a sufficiently motivated eligibility response which appeals to REFERENCE MAGNETISM would appear to be in the offing.

While this is promising, it is by no means obvious that we should say that both natural kinds and moral properties are highly fundamental. One way to see this is to look at the explicit account in Lewis (1983) of how the scale of relative fundamentality (or "relative naturalness") is fixed. For Lewis, one property is more fundamental than another just in case the definition in fully fundamental terms of the former property is *shorter* than that of the latter. Call a definition in fully fundamental terms a *canonical definition.* Then, the property *being a hydrogen atom* is more fundamental than *being a cat,* according to Lewis, because the former has a much shorter canonical definition.

On the canonical definition approach, it is implausible that moral properties will turn out to be reference magnets. One way to see this is to compare the candidate moral properties in the original Moral Twin Earth argument: there, one community applies the term 'wrong' to the consequentialist property P^c and the other community

basis.

[64] This idea is developed in Lewis (1983, 1984) and Sider (2012, Ch. 3), and is discussed in the concluding pages of Hawthorne (2006).

applies the term to the deontological property P^d. According to the reference-magnetic account, these communities should semantically disagree because one of P^c or P^d has a much shorter canonical definition than the other, and hence is much more eligible for reference. But this is implausible: not only is it not obvious *which* property has the shorter canonical definition, *neither* property has a canonical definition that is *much* shorter than the other, and hence neither property will be *much* more eligible for reference than the other, which is what the eligibility response requires. In short, tying degrees of fundamentality to length of canonical definition won't yield an account of reference magnetism that can feature in a motivated eligibility response to the Moral Twin Earth for the synthetic naturalist.

Fortunately for the synthetic naturalist, the canonical definition approach to relative fundamentality is implausible when tied to REFERENCE MAGNETISM for quite general reasons. The failure of the Lewisian account of relative fundamentality to yield an account on which moral properties are reference magnets doesn't derive from any specific features of moral properties. It derives, instead, from general problems in the Lewisian canonical definition approach to relative fundamentality.

As Hawthorne (2007) notes, *many* properties that should, intuitively, count as highly eligible for reference will not be distinguished from their less eligible counterparts by length of canonical definition. Cats, for instance, won't be very special from the canonical definition perspective when compared with dogs-minus-toenails or other intuitively gerrymandered referents—the microphysical realization of cats is itself highly disjunctive and complex, which means that (however the canonical definition for cats is related to their microphysical realization) they will have highly gerrymandered canonical definitions. Thus while one would not wish to claim that moral properties are highly eligible for reference because they have sufficiently short canonical definitions, one should also not wish to claim that cats and other familiar objects are highly eligible, for the same reasons.[65]

After noting that the Lewisian conception of relative fundamentality will in general fail to underwrite a plausible version of REFERENCE MAGNETISM, Hawthorne (2007, 434) recommends jettisoning the Lewisian conception of relative fundamentality:

> It appears, then, that the 'more natural than' relation cannot serve Lewis's semantic purposes if it is tied to definitional length in the canonical lan-

[65] Williams (2007) can be seen as providing a further argument against wedding the Lewisian conception of relative fundamentality with reference magnetism.

> guage. We should thus be willing to give relative naturalness a life of its own, one that allows properties that are of equal definitional distance from the microphysical ground floor to be of radically unequal naturalness.[66]

The eligibility response to the Moral Twin Earth argument should follow Hawthorne in claiming that relative fundamentality is not simply a function of length of canonical definition. Once we accept that proper conception of degrees of fundamentality will let cats count as much more fundamental than other properties that have canonical definitions of the same length, the eligibility response can likewise claim that moral properties (either P^c or P^d or some other property) are much more fundamental than other properties that have canonical definitions of the same length.

Reference Magnetism, then, is a promising route for an eligibility response to the Moral Twin Earth argument, once a sufficiently nuanced conception of fundamentality is in play. Since the proper formulation of the thesis will make reference to degrees of fundamentality (rather than *absolute* fundamentality), we can sketch an account on which moral properties and natural kinds are highly eligible for reference for the same reason: namely, because they are both highly fundamental. Once we divorce our understanding of degrees of fundamentality from Lewisian notion of length of canonical

[66] This theme is echoed in Hawthorne (2006, 206):

> According to the austere physicalist, the perfectly natural properties will only be found at the microphysical groundfloor, relative naturalness being a matter of definitional distance from the perfectly natural properties: to calibrate the naturalness of a property, see how complicated the definition of that property would be in a 'canonical' language in which each predicate corresponded to a perfectly natural property. From such a perspective, the property of, say, being a chair will likely turn out hopelessly unnatural, far less natural than, say, the disjunctive property of being either a hydrogen atom or being fifteen feet from a quark. (Indeed, it wouldn't be surprising if the canonical definition of a chair was infinitary.) The 'emergentist' by contrast, believes that naturalness is not a matter of mere definitional distance from the microphysical groundfloor. (This kind of emergentist can of course allow that everything supervenes on the microphysical.) Perhaps being a cat is far more natural than certain properties far more easily definable in Lewis' canonical language. On the emergentist conception of things, there is no algorithm available for calibrating naturalness in terms of a perfect microphysical language ...
>
> How can the eligibility requirement provide some reasonable measure of determinacy for 'gavagai' if the property of being a rabbit turns out to be hopelessly gruesome? Far better, it seems to me, to opt for an emergentist physicalism, in which semantical joints remain a live option.

definition, we are not required to face compelling arguments that moral properties cannot be reference magnets because they are unprivileged from the canonical definition perspective. This is a promising route, but developing it requires some substantial further commitments. I outline some of them in §5.

2.5 Relative fundamentality and the autonomy of ethics

We can fill out the reference-magnetic picture by considering an instructive objection. It goes as follows: if REFERENCE MAGNETISM is correct as an account of how 'wrong' gets its reference, then a community's use of 'wrong' refers to a property partly in virtue of its relative fundamentality—a metaphysical property. And this, in turn, has first-order normative consequences: if 'wrong' in our mouths refers to a particular property, then it straightforwardly follows that we ought to avoid doing things that have that property. (After all, we ought to avoid doing what is wrong, and if 'wrong' in our mouths refers to a particular property, we ought to avoid doing things that have that property.) The problem with this consequence is that it presents a picture of ethical theorizing that is decidedly at odds with our ordinary practices, as it makes the *metaphysical* consideration about relative fundamentality a central consideration of relevance to theorizing about what we ought to do. But intuitively this isn't right: ethical theorizing should *not* defer to considerations proper to metaphysics. It should instead be autonomous, in the sense that it should give detailed consideration to ethically relevant features like whether an action causes harm, whether its bad consequences are intended, and whether it could be justified from an impartial perspective. It should not worry about what the metaphysician has to say about the relative fundamentality of the properties at issue.[67]

The autonomy objection claims that appeal to REFERENCE MAGNETISM in response to the Moral Twin Earth argument makes moral theorizing objectionably hostage to metaphysical considerations about fundamentality. This is clearly so on some ways of implementing REFERENCE MAGNETISM. Suppose, for instance, that degree of fundamentality is determined in the Lewisian way by length of canonical definition. (We have

[67] Thanks to Pekka Väyrynen for raising the concern that metaphysical fundamentality is the wrong sort of thing to bear on ethical questions like what it is wrong to do. As Peter Railton has pointed out in conversation, some ethically relevant properties might also be of interest to metaphysicians: the results of our theorizing about the nature of free will, causation, and intentionality could in principle be of relevance to the question of what we ought ethically to do. What I am claiming by the "autonomy" of ethics is, instead, simply that a purely metaphysical notion like relative fundamentality is not among these ethically relevant features. One might care about whether an action was freely done in assessing whether it is wrong, but one shouldn't defer to considerations about the degree of fundamentality of the properties the action instantiates in assessing whether it is wrong.

already rejected this approach on independent grounds, but it will useful for illustrating the autonomy objection.) According to REFERENCE MAGNETISM, then, 'wrong' refers to the property that best maximizes fit with use of 'wrong' and degree of fundamentality. But this seems implausible: supposing that both the consequentialist property P^c and deontological property P^d fit use of 'wrong' reasonably well, the way to settle which property 'wrong' refers to (and, by extension, to settle first-order normative debates between the relevant consequentialists and deontologists) is *not* to compare the length of the canonical definition for competing properties. Whether *every* way of implementing REFERENCE MAGNETISM has a similar consequence remains to be seen.

We can motivate one version that is immune to this objection by considering how REFERENCE MAGNETISM doesn't fall prey to analogues of the autonomy objection in other domains. It is plausible to say that fundamental physics, for instance, is closely related to what is absolutely fundamental in the metaphysical sense. This appears to be Lewis's position in the following passage:

> To a physicalist like myself, the most plausible inegalitarianism seems to be one that gives a special elite status to the 'fundamental physical properties': mass, charge, quark colour and flavour .... (It is up to physics to discover these properties, and name them; physicalists will think that present-day physics at least comes close to providing a correct and complete list.)[68]

Lewis's position is naturally elaborated in the following way: that certain properties (mass, charge, etc.) are metaphysically fundamental is a primitive metaphysical fact and cannot be explained in further terms. But we can give an account of how we come to *know* that these properties are metaphysically fundamental: in particular, we can look to fundamental physics to provide an epistemology of the fundamental. We can know that mass is fundamental, because it is countenanced by the best fundamental physical theory. This can be summarized as EMPIRICISM:

EMPIRICISM The fully fundamental properties are the properties singled out by complete physics.

It is important to note what the epistemology of the fundamental does *not* involve according to EMPIRICISM. It does not, in particular, involve settling on which properties are fundamental first, and then recognizing those properties as the ones countenanced

[68] Lewis (1984, 228)

by fundamental physics. Not only would it be unclear how such a procedure should go, it would also impugn the autonomy of fundamental physics, in a familiar sense. Just as ethics needn't defer to metaphysical considerations about fundamentality, so likewise should physics not need prior input concerning the fundamental.

The empiricist picture needn't be limited to the absolutely fundamental. Higher-order scientific theories plausibly implicate claims about *relative* fundamentality. For instance, chemistry tells us that acids are electron-pair acceptors. Electrons are, moreover, more fundamental than acids. According to general relativity, gravity is curvature in spacetime. Spacetime points are, moreover, more fundamental than gravity. And according to astronomy, a galaxy is a collection of stars (and other celestial objects) surrounded by an interstellar medium. Stars are, moreover, more fundamental than galaxies.[69]

These examples suggest a natural extension of Lewis's empiricist epistemology of fundamentality. EMPIRICISM gives fundamental physics the job of providing epistemic access to the facts about which properties are fully fundamental. The extension of this picture is one which gives higher-order sciences the job of providing epistemic access to the facts about which properties are more fundamental than others. We come to know the relative fundamentality fact that electrons are more fundamental than acids by learning of the chemical discovery that acids are electron-pair receptors. We come to know that spacetime points are more fundamental than gravity by learning from general relativity that gravity is curvature in spacetime. And we come to know that stars are more fundamental than galaxies by learning the astronomical discovery that galaxies are collections of stars (and other celestial objects) surrounded by an interstellar medium. This we can call GENERALIZED EMPIRICISM:

GENERALIZED EMPIRICISM The fully fundamental properties are the properties singled out by complete physics; the properties that are significantly (though not fully) fundamental are singled out by higher-order sciences.[70]

As before, we should be clear about what GENERALIZED EMPIRICISM picture does

[69] To use a metaphor that it sometimes helpful in articulating fundamentality claims: in making stars separated by an interstellar medium, God didn't have to do anything more to make galaxies—and similarly for our other examples. Of course, God presumably doesn't make stars and interstellar media "directly", but by making the microphysical grounds for these entities. The basic point remains, however: *once* the distribution of celestial bodies and gases are in place, nothing *further* has to happen for there to be galaxies. Similar metaphors could be used to motivate relative fundamentality claims about acids and gravity.

[70] Compare Schaffer (2004).

not involve. It does not, in particular, involve settling on which properties are highly fundamental first, and then recognizing those properties as the ones countenanced by chemistry, general relativity, or astronomy. Such a procedure would impugn the autonomy of these higher-order sciences, in a familiar sense. Just as ethics needn't defer to metaphysical considerations about fundamentality, so likewise should chemistry, general relativity, and astronomy not need prior input concerning the fundamental.

GENERALIZED EMPIRICISM is a natural implementation of Hawthorne's rejection of the canonical definitions approach to relative fundamentality. Recall Hawthorne's recommendation that we should be "willing to give relative naturalness a life of its own, one that allows properties that are of equal definitional distance from the microphysical ground floor to be of radically unequal naturalness". GENERALIZED EMPIRICISM tells us *when* properties that are of equal definitional distance from the groundfloor are of unequal degrees of fundamentality. Properties that are countenanced by higher-order sciences will have extremely long canonical definitions. But according to GENERALIZED EMPIRICISM the properties of being an electron-pair acceptor, for instance, will nonetheless be highly fundamental in spite of its long canonical definition.[71]

Which higher-order sciences provide epistemic windows into the facts about relative fundamentality is a large question I will not try to answer here. We have given plausible examples from general relativity, chemistry, and astronomy; we might also include psychology, economics, sociology, and other disciplines on the list. What is of relevance to the Moral Twin Earth argument is that it is at least a live option that we should be so liberal with the list of relevant disciplines so as to include ethical theorizing alongside the other higher-order sciences that provide epistemic access to the facts about relative fundamentality. That is, just as we learn from chemistry that electron-pair acceptors aren't fully fundamental but are nonetheless *highly* fundamental, the naturalist about ethics should take ethics as delivering similar results for wrongness. Suppose, for instance, that ethical investigation tells us that wrongness is failing to maximize net happiness. Then, one of the things we learn from ethics is that failure to maximize happiness is highly fundamental. This application of GENERALIZED EMPIRICISM is has several nice features for the purposes of an account that appeals to REFERENCE

[71] It is important to remind ourselves at this point that GENERALIZED EMPIRICISM isn't an account of what *makes* one property more fundamental than another: relative fundamentality, on the present approach, is an undefined metaphysical notion. The role of higher-order science is purely epistemic on the present approach, and so properties that play explanatory roles in distinct disciplines will retain a metaphysical similarity in virtue of being highly fundamental. This point will be crucial for avoiding the pitfalls of van Roojen's account.

MAGNETISM in giving an eligibility response to the Moral Twin Earth argument.

First, it explains why ethical properties should be highly eligible for reference *in exactly the same way* that natural kinds and other properties countenanced by the sciences are highly eligible for reference. They are both highly fundamental in spite of their long canonical definitions. Moreover, it does so without compromising the autonomy of ethical theorizing from metaphysical considerations about fundamentality. We emphasized how, according to GENERALIZED EMPIRICISM, in making metaphysical claims about relative fundamentality we defer to the deliverances of the higher-order sciences, rather than the other way around. Hence theorizing in chemistry, astronomy, etc. need not take prior account of which things are highly fundamental before delivering their theories of acids, galaxies, and the like. Instead, theorizing about acids and galaxies tells us which things are highly fundamental, *viz.* electron-pair acceptors and celestial bodies separated by an interstellar medium. If GENERALIZED EMPIRICISM is extended to allow in ethical theorizing, it will be autonomous in exactly the same way: the deliverances of ethical theory tell us which properties are highly fundamental, rather than deferring to considerations of fundamentality. Finally, this account explains why the assumption IE^+, which is at the center of the Moral Twin Argument, would be appealing, but not in general true. I investigate this claim in greater detail in closing in §6.

2.6 Conclusion: synthetic knowledge and metaphysical possibilities

The challenge for an eligibility response is not only to give a constructive account of what moral properties and natural kinds have in common, which makes both highly eligible for reference. The challenge is also to do this in a way which invalidates Horgan and Timmons' assumption IE^+. This requires that the synthetic naturalist not claim that moral properties are too similar to natural kinds; analogues of IE^+ are clearly *true* when applied to paradigmatic natural kind terms. That is, when one learns that one's term 'water' refers to H_2O, what one learns rules out the possibility where one's use of water bears the reference-determining relationship to a XYZ and not H_2O. The motivation for IE^+ is the assumption that the same thing goes for any terms which synthetically refer like natural kind terms. In learning what one's use of such terms refers to, one learns something that rules out a metaphysical possibility where one's use of the term bears the reference-determining relationship to a different property. Applying this general claim to moral terms, Horgan and Timmons' conclusion that there must be a second community to feature in an MTE pair follows.

The combination of GENERALIZED EMPIRICISM and REFERENCE MAGNETISM provides the outline of an account on which the general claim IE^+ is false, but is applicable in certain cases—in particular, is true of natural kinds. H_2O is highly eligible to be the referent for 'water', according to REFERENCE MAGNETISM, because H_2O is a highly fundamental kind—in particular, it is much more fundamental than the gerrymandered kinds that provide a better fit with some uses of 'water'. (Presumably, the same thing should be said about XYZ on Twin Earth.) According to GENERALIZED EMPIRICISM, we come to *know* that H_2O is highly fundamental through its role in chemical explanations. But chemistry is a contingent empirical science—the true explanations it offers in the actual world are *not* true with respect to others. In particular, chemical explanations for the solubility of salt (for example) will not be true with respect to the XYZ-world: there, it is facts about the structural features of XYZ that explains the solubility of salt-like substances.[72]

It is the contingency of chemistry (and other sciences that study what are traditionally called "natural kinds") that explains why there are other worlds where our natural kind terms, used in the same way, have different referents. That, in short, is why IE^+ is true of natural kind terms: chemistry and other empirical sciences are metaphysically contingent. But it is a mistake to assume that *every* synthetic discipline is contingent in the same way: it is plausible that some synthetic knowledge rules out no new metaphysical possibility, and the proponent of the reference-magnetic response to the Moral Twin Earth argument is free to appeal to this idea.

Implementing this claim in the framework of REFERENCE MAGNETISM and GENERALIZED EMPIRICISM is straightforward. While some of the theoretical disciplines that, according to GENERALIZED EMPIRICISM, provide epistemic access to relative fundamentality are empirical and contingent, it need not be that *every* such discipline is contingent. (Nor, for that matter, need they be empirical—hence even if the motivating case for GENERALIZED EMPIRICISM is the empirical sciences, extending it in the way I have been outlining points to one way in which 'empiricism' might be a misnomer.) In particular, the liberal approach that accords ethical theorizing a place alongside chemistry, astronomy, and the like is plausibly one on which some, but not all, of the disciplines that provide epistemic windows into the relative fundamentality facts state

[72] Nothing hangs on whether we deny the truth of H_2O-based explanations in the XYZ-world, or if we instead accept that the H_2O-based explanations are true but fail to apply to concrete substances in that world. Either way, H_2O-based explanations will fail to make predictions and capture generalizations about the distribution of matter across the XYZ-world. Whichever kind of contingency we prefer will be related to the fact that 'water' in that world refers not to H_2O, but instead to XYZ.

contingent claims. When one learns from ethical theorizing what 'wrong' refers to, on this view, one learns a synthetic fact, but not one that rules out a metaphysical possibility where 'wrong' refers to a different property. IE$^+$ is false, and for straightforward reasons.

The final point to make about the reference-magnetic solution on offer here is that it should not be confused with a solution to another, closely related problem. The Moral Twin Earth argument is a semantic objection to synthetic naturalism, one according to which the semantic commitments of the view have untoward consequences, failing to account for the full range of semantic disagreement. I have suggested, as part of the response to this argument, that the synthetic naturalist take ethical theorizing to single out properties that are highly fundamental, and hence highly eligible for reference. This might seem to fail to address the problem raised by the Moral Twin Earth argument altogether: after all, even if we think that ethical theorizing singles out a particular natural property as the property of wrongness, the whole point of the argument is that there is *another* community that uses 'wrong' differently and hence thinks that a different natural property constitutes wrongness. Nothing I have said here addresses how we know that we, and not the other community, has the correct view of wrongness.

It is true that REFERENCE MAGNETISM and GENERALIZED EMPIRICISM don't address this issue, but this is to be expected: the problem of knowing which community is correct is an epistemic problem for the synthetic naturalist, and not a semantic one. These problems are distinct: it could be that there is a single property that is the ethical property of wrongness, but that we fail to come to know what it is through mistakes in our ethical theorizing, or because our starting point is unfortunate in a way that makes subsequent theorizing bound to mislead us. All of this is consistent with the claim that moral properties are reference magnets, and so communities who diverge in their use of moral language nevertheless refer to the same property and hence semantically disagree. It just points to a way in which we might not be able to *know* which properties are the reference magnets, and hence know which community speaks truly in a disagreement. It is not obligatory for a response to the Moral Twin Earth argument to solve problems in epistemology of this kind. Thus to object to this aspect of REFERENCE MAGNETISM and GENERALIZED EMPIRICISM is to give up on the problems raised by the Moral Twin Earth argument, and to instead renew one's attack on a different front.

Chapter 3: Expressivism and Normative Metaphysics

The recent literature on Expressivism gives rise to the problem of differentiating *Expressivist* views about the normative from *realist* views. Dreier (2004) explains this problem as the "problem of creeping minimalism". But there are different (though related) kinds of problem that fall under this heading—one is a *semantic* thesis about what sentences Expressivists can accept; another is a *metaphysical* thesis about how Expressivists think the world is. I distinguish these problems in §2, and then show how, from within a metaphysical framework centered around the notion of fundamentality, we can articulate a solution to the metaphysical problem. According to this solution, realists are committed to metaphysical claims about the normative in giving an account of the *reference* of normative terms; Expressivists, meanwhile, incur no such commitments, owing to the fact that claims about the reference of normative terms are *plan-laden*.

3.1 First-order vs. theoretical differences

Expressivist views have been proposed for a variety of domains, including the normative, epistemic modality, and indicative conditionals.[73] Consider Expressivism about the normative (similar points might be made about epistemic modality and conditionality, though I will not pursue them here): the view is primarily a view about the *semantics* of words like 'ought'. The Expressivistic semantics presents a puzzle: it seems that, by adopting this treatment of the language for talking about the normative, one is thereby taking a distinctive stance on what the normative domain is *like*. It is easy enough state in rough terms why this seems to be so: according to Expressivist view, utterances of normative sentences merely express of a state of mind—and this state of mind is not most perspicuously described as a *belief* about what ought to be done. It therefore *seems* that the Expressivist, in virtue of her semantics for normative sentences, has a picture of the normative that differs substantially from an ordinary realist picture, which finds no motivation for a distinctive semantic take on the language required for speaking about the normative. The puzzle arises because stating this seeming in more precise terms turns out to be remarkably difficult.

[73] See Gibbard (1990, 2003), Yalcin (2011), and Bennett (2003).

The following passages from Finlay (2007, 822) constitute a concise and natural attempt to explain where, exactly, distinctiveness of Expressivism lies:

> The most modest face of moral realism is a semantic thesis. Its objects are moral *claims* (whether judgments, utterances, beliefs, or propositions), of which it holds that they or their contents have objective truth values. These truth values are 'objective' in that they are independent of the attitudes that anyone takes towards the moral claims. The strongest form of moral antirealism [...] is found in the *Expressivist* tradition.
>
> Expressivists [...] also reject the *ontological* face of moral realism (or 'descriptivism'). This form of realism takes as its objects the truth-makers of moral claims, holding that they include moral properties such as value (e.g., the *goodness* of charity) and moral entities such as practical reasons and obligations (e.g., *reasons* not to tell lies, *obligations* to keep promises).

In brief, Finlay thinks that Expressivism is committed to two things which differentiate it from realism. The first is that normative claims fail to have their truth-values independently of facts about the mental. And second, Expressivism entails that moral properties (such as goodness, wrongness, etc.) do *not* make English sentences containing 'good' true.

In what sense does Finaly think that, according to Expressivism, the normative is not independent of the mental? One natural interpretation is that Expressivism denies the following modal claim, where M is a mental fact (of an appropriate kind) and E is a normative one:

MODAL INDEPENDENCE (MI) $\exists w, w^*$: M and E are true at w, yet E is true at w^* and M is not true at w^*.

If Expressivists are committed to rejecting MI, they are committed to holding that there are certain ways for the mental facts to be, and these mental facts *determine* how things are normatively—no two worlds alike in the relevant mental respects differ in normative respects.

But Expressivists such as Blackburn (1984) and Gibbard (2003) famously *accept* the MI. Roughly, the Expressivist points out that MI makes a normative claim, and then applies the familiar Expressivist semantics for normative sentences to show why even theoretical claims of this kind are acceptable. To take an example, consider the claim that it's being the case that one ought not to murder is modally independent from the mental. This is true, according to MI, if there is a possible world w^* where the mental facts are in appropriate ways different from how they actually are, yet it is still true that one ought not murder in w^*. The question of whether one ought not to murder in w^* then a normative question about what to *do* in w^*. For Gibbard, what it is to accept that one ought not to murder in w^* is to plan, for such a situation, not to murder. Thus it is coherent for the Expressivist to think that one ought not murder in w^*, and hence to think that MI is true. [74]

The explanation for the acceptability of MI proceeds via an "indirect" mode of explanation that is the hallmark of Expressivist theorizing. The primary instance of this mode of explanation is found in the Expressivist explanation of basic normative sentences: their account of sentences like

T One ought not to tell lies

does not, in the first instance, proceed by telling us what *makes* **T** true. The Expressivist instead explains what it is to *accept* **T**.[75] Normative sentences like **T** are acceptable, then, because it is coherent to be in the mental state that constitutes accepting **T**. That is because the state that constitutes accepting **T** is the state of planning not to tell lies, and there is nothing incoherent about being in this state. By finding normative commitments in sentences like MI, Expressivists can extend this indirect method of explanation to show why it is acceptable.[76]

An indirect explanation of this kind shows that Finlay's second gloss on the distinctiveness of Expressivism fails as well. Finlay claims that another distinctive feature of Expressivism is that it cannot afford an ontology of properties such as goodness, and hence cannot accept the following claim:

PROPERTY-TRUTH CONNECTION (PTC) It is true that x is good iff x instantiates goodness.[77]

[74] Of course it might be that the difference in the mental facts in w^* include that one plans to murder in w^*. But this is just to say that *in those circumstances* one plans to murder; the question of whether MI is acceptable is a question of whether it is coherent to *actually* plan *now* not to murder in the case where one is someone who plans to murder.

[75] See Schroeder (2008a, Ch. 2) for more discussion.

[76] This is carried out in greater detail in Gibbard (2003, Ch. 5).

[77] I will not focus on the topic of whether Expressivists can say in addition that goodness is part of what *makes* it true that x is good. Although Finlay does use the language of truth-making in stating the issue, his official reason for holding that Expressivism has distinctive commitments in the area is that it lacks the ontological resources—i.e., is committed to the non-existence of certain properties or entities—to accept the biconditional.

Begin with the left-hand side of PTC: to accept it is is to accept a basic normative claim, that *x* is good. Its acceptability can then proceed along the lines of the explanation of **T**. Turn next to the right-hand side of PTC: Expressivists have claimed that to accept that something instantiates goodness requires nothing more than accepting that it is good.[78] So one must be in *exactly* the same state to accept either side of the biconditional; it is therefore coherent (and plausibly required) for the Expressivist to accept PTC.

3.2 Metaphysical and semantic versions of creeping minimalism

MI and PTC were claims that appeared *prima facie* to be acceptable only to realists—the quoted passage from Finlay gives voice to two very natural reactions to the Expressivist view by claiming that it is inconsistent with MI and PTC. But we have seen how the Expressivist's indirect method of explanation can, in fact, accommodate these claims. Dreier (2004) anticipates the possibility that an Expressivist-friendly indirect explanation for *every* such claim will be available, and he says:

> Suppose you'd never done a shred of meta-ethical thinking before, and you started out with Simon Blackburn's "How To Be a Moral Anti-Realist" (in his 1993), in which a story is told (sketched) about how creatures like us might start to talk and think the way we do when we are "moralizing". The story convinces you. Now you meet a self-styled moral realist, who tells you that there is something missing from the story you've come to accept. How is this realist going to say what that something is?
>
> In fact, it seems to me, the problem is not a problem *for* realists or *for* irrealists, but more a problem in meta-meta-ethics. It's not as if one side had better be able to come up with something clever to say about how to distinguish realism from irrealism or else the other side wins. It's rather that those of us who feel confident that there is some difference between the two meta-ethical camps should be concerned that we don't know how to say what that difference is.[79]

If the indirect method of explanation can be extended to every sentence that is intuitively consistent with realism, Dreier's question would appear to be a pressing one. It is puzzling if there seems to be a difference between the two views, but we "don't know how to say what the difference is". As a matter of fact, there are multiple puzzles here. This is because there are multiple ways to "say what the difference is".

[78] Dreier (2004, 26)

[79] Dreier (2004, 31)

Here are two different ways to do this. (I do not pretend that these are the only two.) The first way involves pointing to a limitation to the indirect method of explanation. There is no guarantee that, as Dreier conjectures, the indirect method can be used to explain the acceptability of *all* of the realist's sentences in the way it could explain MI and PTC. Let us call a response which does this an answer to the *semantic* question posed by Dreier's problem. A successful answer to this question points to a sentence, or a set of sentences, which are *prima facie* consistent with realism, yet cannot be shown to be consistent with Expressivism via the indirect method.

A second approach to the puzzle treats it as raising a *metaphysical* question. This question asks how the Expressivist incurs different metaphysical commitments from those of the realist by explaining the acceptability of realist-sounding claims via the indirect method. The metaphysical question requires an answer to the further question of what metaphysical commitments *are;* I will turn to this question in §3. But first, it will be useful to note that the semantic and metaphysical questions are distinct from each other.

This can be most easily seen by noting that the semantic question doesn't guarantee an answer to the metaphysical question. Suppose that the two meta-ethical theories are not consistent with exactly the same sets of sentences—Expressivism must reject some sentences that realists can accept, as a solution to the semantic question would have it. It is consistent with this that the sentences over which they differ do not carry any significant metaphysical implications. An answer to the semantic question would only constitute an answer to the metaphysical question if the sentences over which Expressivists and realists can be shown to disagree are sentences which themselves encode substantial metaphysical commitments.[80]

In the other direction, an answer to the metaphysical question need not constitute an answer to the semantic question. This can happen if the Expressivist, by virtue of her distinctive method of indirect explanation, incurs different metaphysical commitments from those of the realist, even if she ends up accepting all of the same sentences as the realist. This is, in short, a certain kind of difference in explanatory priority, and it need not presuppose an answer to the semantic question. In what follows, I will argue for

[80] In Dunaway (2010) I argued that a Blackburn-style Expressivism cannot accept all of the realist's sentences. This difference with realism is definitely a semantic difference; whether it is also a metaphysical difference is less clear. In what follows, I outline a framework for understanding metaphysical commitments generally, and then address the metaphysical version of the problem of creeping minimalism explicitly in this framework.

precisely this kind of answer to the metaphysical problem: that is, without presupposing that Expressivists are committed to rejecting any of the sentences realists accept, I will argue that they nonetheless hold a view that is metaphysically at odds with realism. The answer to the metaphysical question will come from within a particular framework for expressing metaphysical commitments more generally. This is a framework which takes the notion of metaphysical fundamentality as central; metaphysical differences, according to this framework, are differences over the *degree of fundamentality* of certain things or properties.

After explaining in more detail in the next section what this metaphysical framework amounts to, I will argue in the remaining sections that normative properties are, according to Expressivism, *less fundamental* than they are on the realist view.

3.3 Metaphysical fundamentality

A fundamentality-centric metaphysical framework is not uncontroversial. I won't attempt to defend use of the framework here, and will instead treat the subsequent argument as an extension of the metaphysical framework developed (in different ways) in Lewis (1983), Fine (2001, 2009), Sider (2012) and Schaffer (2009). I will follow these theorists by taking the notion as a primitive, not subject to explanation in independent terms.

3.3.1 Fundamentality: its explanatory character

Even if metaphysical fundamentality is primitive, it can still be clarified through its to closely related concepts and its applicability to concrete examples. For instance, it is characteristic of fundamentality that it is a kind of explanation: that it explains, in a distinctive metaphysical way, facts about the non-fundamental. Fine (2001, 22) says:

> [T]he relationship of ground is a form of explanation; in providing the ground for a given proposition, one is explaining, in the most metaphysically satisfying manner, what it is that makes it true. Thus a system of grounds may be appraised, in much the same way as any other explanatory scheme, on the basis of such considerations as simplicity, breadth, coherence, or non-circularity. Perhaps the most important virtue in this regard is explanatory strength, the capacity to explain that which stands in need of explanation and would otherwise be left unexplained.

Thus our catalogue of what is *most* fundamental is settled by the criteria by which we evaluate explanations, applied to a distinctive kind of metaphysical explanation. I will assume that the canons of evaluation for claims about metaphysical fundamentality include *parsimony*—what we countenance among the fundamental should, other things being equal, be limited to as little as possible.

Fine (2009) describes a specific case where, according to the fundamentality-theorist, there is disagreement over the relevant kind of metaphysical explanation. What distinguishes realist and irrealists about numbers, intuitively, is their respective stances on the metaphysics of numbers. But both parties can *accept* the unadorned sentence

There are numbers.

After all, even the irrealist view is compatible with the claim that five is the number of my fingers on my left hand and the claim that the sum of two and three is five. The existence of the number five straightforwardly follows from both claims.

What is distinctive of the realist about numbers, Fine says, is that she not only thinks numbers exist; she thinks that they are *fundamental*, as evidenced by the realist's refusal to give a further analysis or explanation of the facts about numbers. This is to be contrasted with the irrealist's approach which on which facts about numbers *can* be explained in further, more basic terms. Thus she is committed to thinking that there is something *more fundamental* than numbers, namely our counting practices, mental states, etc. The metaphysical difference between realist and irrealist views lies not at the level of first-order existence-claims. Rather, it is located in different stances over the fundamentality of numbers. This, in turn, amounts to a difference over whether facts about numbers can be given a further "metaphysically satisfying" explanation.

The small amount we have said about fundamentality so far gives us a *prima facie* case for thinking that it can be used to spell out an answer to the metaphysical question about Expressivism. Subsequent sections are dedicated to spelling out the case in greater detail, but here is the cause for optimism: in our above example, both realists about numbers and their irrealist opponents can *agree* on all of the first-order claims of mathematics, yet disagree about the metaphysics of numbers by disagreeing over whether numbers are fundamental. The disagreement over the metaphysical question is, at bottom, a disagreement over how the first-order facts about numbers are best explained. Without presupposing an answer to the semantic version of Dreier's worry, the account of the distinctiveness of Expressivism must have the same form: since Expressivists might accept all of the same first-order normative claims (construed broadly so as to include MI and PTC) as the realist, the differences must lie in *how* Expressivists explain these first-order claims. The difference in method of explanation will, if all goes well, give rise to a difference over the fundamentality of the normative. The task of subsequent sections is to show, in greater detail, how the Expressivist's indirect method of explanation gives rise to a difference in fundamentality of the kind the irrealist in mathematics is committed to.

3.3.2 Fundamentality: the relative notion

Before moving on to the positive argument for this conclusion, one further bit of stage-setting is needed. There is an important difference between realists about mathematics and realists about the normative: in the case of mathematics, realists are plausibly just those who take the domain to be irreducible—i.e., not grounded in anything. Irrealists, then, are those that think mathematics is grounded in *some* other domain; anyone who thinks the mathematical is not fundamental is not a realist. Things cannot proceed in exactly the same way when we are distinguishing realists about the normative from irrealists. This is because some realists about the normative, such as Railton (1986) and Schroeder (2007), are also naturalists. Since these views are views on which the normative is not fully fundamental, realism about the normative cannot simply be the view that the normative is most fundamental.

The fundamentality-theorist should, at this point, move away from a binary distinction between what fundamental and what is not, and toward a *relative* notion. The relative notion allows us to say that some things are more fundamental than others without being *fully* fundamental. Realists such as Railton and Schroeder, then, are plausibly realists in that their views entail that the normative is fundamental *to a high degree*.[81]

The relative notion of fundamentality is, like the absolute notion, an explanatory notion: if A is more fundamental than B, then A offers the same kind of metaphysical explanation of B. Perhaps it isn't the *final* such explanation, if A is not fully fundamental—in this case, there are some further facts that explain "in the most metaphysically satisfying manner" what makes both A and B true (possibly these further facts will be fully fundamental). The relative notion of fundamentality thus not only does many of the jobs that the absolute notion by itself can do; it also features in explanations that the absolute notion by itself is ill-suited for.

One of these jobs is *reference magnetism*; I will assume for the rest of this chapter that part of what it is to be highly fundamental is to be a reference magnet in the following

[81] These issues are covered in more detail in Chapter 1.

sense. Many have thought that fundamentality is naturally suited to solve a class of puzzles about reference, one of which is found in the "model-theoretic argument" of Putnam (1981, Ch. 2).[82] What fundamentality-based solutions to these puzzles have in common is the following.[83] They claim that the reference of a term is determined not only by how the term is used and which candidate referents "fit" the use of the term, but also by *how fundamental* the candidate referents are. More precisely: given a pattern of use of a term t by a community and a set of candidate referents $e_1 \ldots e_n$ that all fit the use of t to some degree, t as used by the community refers to whichever of $e_1 \ldots e_n$ scores best on the twin considerations of fit with use and eligibility. This is, to a first approximation, the phenomenon of reference magnetism. The relative notion of fundamentality is crucial to a reference-magnetic solution to Putnam-esque puzzles about reference, as any solution which appeals only to the absolute notion will fail to be sufficiently general.[84]

In is important to be clear about what the commitments of an appeal to relative fundamentality are. Lewis (1983) suggests that the scale of relative fundamentality (or "relative naturalness" in his terms) is fixed by length of definition in fundamental terms. Thus the property *being a lepton* is more fundamental than *being a cat*, on Lewis's view, *because* the latter has an extremely long definition in fundamental terms. Call such a definition a *canonical definition*. Tying length of canonical definition to degree of fundamentality will not underwrite the relationship between relative fundamentality and reference magnetism. Hawthorne (2006) points out that it is unlikely to explain what makes highly complex macroscopic objects eligible for reference, since they are likely to have infinitely long canonical definitions (as will their less-eligible counterparts).

But the Lewisian length-of-canonical-definition approach is not forced on us. After noting that this conception of relative fundamentality fails to adequately underwrite a theory of reference, Hawthorne (2007, 434) says:

> It appears, then, that the 'more natural than' relation cannot serve Lewis's semantic purposes if it is tied to definitional length in the canonical language. We should thus be willing to give relative naturalness a life of its own,

[82] Other similar puzzles are found in Quine (1960) and Kripke (1982).

[83] This idea is developed in Lewis (1983, 1984) and Sider (2012, Ch. 3), and is discussed in the concluding pages of Hawthorne (2006).

[84] For instance, we will want an explanation of why 'cat' refers to cats, and not cats-minus-a-toenail or cats-plus-a-few-molecules-of-surrounding-air. Since none of these is fully fundamental, an account which holds that only things that are most fundamental are more eligible for reference will not have anything to say about cats.

> one that allows properties that are of equal definitional distance from the microphysical ground floor to be of radically unequal naturalness.[85]

I will follow Hawthorne in what follows: relative fundamentality is not simply a function of length of definition in fundamental terms. Given this "emergentist" conception, the metaphysical consequences of Expressivism now come into view.

3.4 Realism and reference magnetism

Given a realist view of the normative, the referents of normative terms are fundamental to a significant degree and hence serve as reference magnets. One way to spell out the consequences of this package of theses is to consider them in light of the "Moral Twin Earth" argument of Horgan and Timmons (1992), though we can for present purposes remain neutral over how much reference magnetism can do for the realist in solving the alleged Moral Twin Earth problem. All that matters is that there are some possible communities who use moral language differently, yet owing to reference magnetism still refer to the same property. What Horgan and Timmons do is present us with a convenient (and familiar) example of such a case when they present an argument against synthetic naturalist versions of realism about the normative. It is based on a kind of thought experiment, involving what we can call a *Moral Twin Earth scenario*: we imagine a community of speakers who use the same normative language we use ('ought',

[85] This theme is echoed in Hawthorne (2006, 206):

> According to the austere physicalist, the perfectly natural properties will only be found at the microphysical groundfloor, relative naturalness being a matter of definitional distance from the perfectly natural properties: to calibrate the naturalness of a property, see how complicated the definition of that property would be in a 'canonical' language in which each predicate corresponded to a perfectly natural property. From such a perspective, the property of, say, being a chair will likely turn out hopelessly unnatural, far less natural than, say, the disjunctive property of being either a hydrogen atom or being fifteen feet from a quark. (Indeed, it wouldn't be surprising if the canonical definition of a chair was infinitary.) The 'emergentist' by contrast, believes that naturalness is not a matter of mere definitional distance from the microphysical groundfloor. (This kind of emergentist can of course allow that everything supervenes on the microphysical.) Perhaps being a cat is far more natural than certain properties far more easily definable in Lewis' canonical language. On the emergentist conception of things, there is no algorithm available for calibrating naturalness in terms of a perfect microphysical language …
>
> How can the eligibility requirement provide some reasonable measure of determinacy for 'gavagai' if the property of being a rabbit turns out to be hopelessly gruesome? Far better, it seems to me, to opt for an emergentist physicalism, in which semantical joints remain a live option.

See also Williams (2007).

'good', 'wrong', etc.), but use it very differently. The crucial claim Horgan and Timmons make about such scenarios is that the speakers in them *mean* the same thing by their normative terms, and hence *semantically disagree* with us when they use such terms differently.

In the original example from Horgan and Timmons (1992), we are asked to consider a scenario where one community applies normative terms in accordance with a consequentialist theory T_C, while the other applies the terms in accordance with a deontological theory, T_D. Horgan and Timmons then claim that the T_C-community semantically disagrees with the T_D-community over sentences like

A In circumstance c, one ought to ϕ,

where ϕ is an action which, in the maximally specific situation c, maximizes total happiness but fails to respect the humanity of agents in c.[86]

There are other Moral Twin Earth scenarios that fit this general form: in these scenarios, other possible communities that use their normative language differently will be like the T_C and T_D-communities, disagreeing with each other in their use of sentences like **A**.[87] Horgan and Timmons take the data from Moral Twin Earth scenarios to constitute an argument against synthetic naturalist theories. I won't go into their reasons for thinking this here, but they assume that any theory of reference available to naturalistic versions of realism will fail to account for the full range of disagreement in Moral Twin Earth scenarios.

Moral Twin Earth scenarios illustrate one nice feature of the normative realist's commitment to the fundamentality of the normative, as outlined above. Within such a fundamentality-centric framework, realists are committed to normative properties being highly eligible for reference—that is, they are reference magnets. This goes at least some of the way toward explaining, for the realist, why communities in some Moral Twin Earth scenarios refer to the same properties and disagree, even while using their normative terms differently. At least some such scenarios will in involve communities which apply their terms to properties that fail to be fundamental to a significant degree, and hence refer to much more fundamental nearby properties. The point for present

[86] Speakers disagree *semantically* over a sentence like **A** when what one community says by uttering **A** is inconsistent, by virtue of what the sentence means in the mouth of that community, with what the other community says by uttering the negation of **A**.

[87] It is controversial *how far* the disagreement plausibly extends. I am only assuming here that there are *some* cases that fit this general description, but make no assumptions about whether the disagreement data is as robust as Horgan and Timmons claim.

purposes isn't to claim that this can explain the full range of semantic disagreement in Moral Twin Earth scenarios—that claim is stronger than what we need here. Rather, it is only that, once the fundamentality-plus-reference-magnetism framework is in place, it will do the work of explaining at least *some* of these disagreement facts for the realist.

Moreover, interest in this explanation of the some of the disagreement facts is not limited to the naturalistic realists who hold the kind of view Horgan and Timmons wish to directly challenge with the Moral Twin Earth argument. It also applies to the *non*-naturalist account of reference in such scenarios. Even though Moral Twin Earth scenarios have not been presented as a direct challenge to the non-naturalist view, they still need some account of why communities in the relevant scenarios disagree. Given that non-naturalists hold a view on which the normative is *most* fundamental, reference magnetism seems to offer a promising account of the data for adherents to this view as well.[88]

Thus Moral Twin Earth scenarios highlight, from within the present metaphysical framework, the realist's commitment to a high degree of fundamentality in the normative. Once we are inside a framework where metaphysical fundamentality serves in part to distinguish realist from irrealist views, the normative realist—whether she is a naturalist or non-naturalist—also has a built-in explanation of the disagreement in a range of Moral Twin Earth scenarios. As I will show below, Expressivists can explain the same data. But their explanation does not avail itself of the same metaphysical facts about the fundamentality of normative properties via an appeal to reference magnetism. Since fundamentality is an explanatory notion (and is thereby subject to considerations of parsimony), this gives rise to a difference in metaphysical commitments for the Expressivist view.

3.5 Expressivism and reference

The essentials of the Expressivist account of meta-semantic phenomena involving reference can be found in Gibbard (2013), which I will loosely follow here. The crucial move is to expand on the familiar indirect explanation of first-order normative sentences like **T** found in standard Expressivist theories. Since claims about the reference of normative terms like 'ought', suitably qualified, conceptually entail first-order normative claims, they too should be treated withing the Expressivist's indirect explanatory apparatus. Crucially, claims about reference are *not* treated withing the

[88] See Chapter 4 for more on the non-naturalist view.

realist's reference-magnetic framework. This ensures a difference between Expressivists and realists in commitment to the fundamentality of normative properties.

3.5.1 Planning-states

When describing the Expressivist's semantics for normative terms, we must distinguish sharply between the *meaning* of a term and its *reference*. What a term means, in an Expressivist semantics, is explained by the mental state(s) expressed by a sentence containing the term. Gibbard offers the dictum "to explain the meaning of a term, explain what states of mind the term can be used to express."[89] But by saying what state of mind a normative sentence expresses, Expressivists do *not* explain the meaning of a sentence via a specification of the objects, properties and relations its constituent terms *refer* to. Thus the primary mode of explanation of meaning, in the Expressivist framework, doesn't directly explain the meaning of normative terms by specifying their referents.[90]

Expressivists can, from within this explanatory framework, take on different views concerning *which* mental states are expressed by normative sentences. There are some constraints on what kinds of states can play this role in a genuinely Expressivistic theory (it will not do for the Expressivist, for instance, to say that $\ulcorner\phi$-ing is wrong$\urcorner$ expresses the belief that ϕ-ing is wrong, with no further explanation of what the belief that ϕ-ing is wrong is), but I will not try to enumerate them here. I will instead follow a paradigm of Expressivist explanation by taking first-order normative sentences to express *planning-states* of the kind described in Gibbard (2003). This will give us the needed precision to discuss disagreement and reference in the Expressivist framework. But it is important to note that the upshot of much of what follows is independent of this choice. (See Appendix A for more on this.)

Planning-states are, to a rough approximation, commitments for how to act in various hypothetical scenarios, or scenarios that can be coherently imagined. Some of these hypothetical scenarios are also *actual*: I am committed to continue writing a paper on philosophy for the near future; you (I assume) plan to sleep sometime in the next 20 hours, etc. But we also have commitments to do things in circumstances that are not actual. For instance, for the circumstance of my being on a sunny beach in a tropical island right now, I do *not* plan on continuing to write a philosophy paper—I plan to sit

[89] Gibbard (2003, 7)

[90] This is not to deny that Expressivists can claim that normative terms *have* referents—as I will show below, then can. The present point is just that specifying a referent for 'wrong' does not, on the Expressivist view, constitute an explanation of what that term means.

and enjoy the sun. And for the circumstance of being in a rush to meet a deadline, you (perhaps) do not plan on sleeping in the next 20 hours.[91]

Planning-states are not themselves true or false, but sets of planning-states nevertheless are capable of standing in a kind of inconsistency with each other. In particular, they are capable of bearing a kind of *disagreement-relation* with each other. One such relation is PLAN-DISAGREEMENT:

PLAN-DISAGREEMENT Planning-states P_1 and P_2 *disagree* with each other just in case in is incoherent for a single person to be in both P_1 and P_2 at the same time.

For example: I plan, for the circumstance of being on a sunny beach, plan to sit and relax. But Sally, who values productivity over anything else, plans for the same circumstance to continue to work. Our plans disagree, in the above sense, as anyone would be guilty of a kind of incoherence by adopting both plans at once.

All of this concerns the meaning of normative terms for the Expressivist: that they express planning-states, and that these states stand in a relation of disagreement to each other. Given the notion of a planning state and PLAN-DISAGREEMENT, the Expressivist program aims to explain a range of related semantic phenomena including inconsistency, entailment, etc. Whether these resources are sufficient to the task is a large question I will not attempt to answer here.[92] What is important for our purposes here is simply the fact that *if* this style of explanation is successful, the Expressivist explains the semantic phenomena of inconsistency, disagreement, etc. without direct appeal to the notion of *reference*.

I will discuss below how the Expressivist can incorporate reference into this picture. But first, I will show how the resources of planning states and PLAN-DISAGREEMENT can

[91] Lurking in the background here is a distinction between kinds of plan for hypothetical scenarios: in the examples I gave above, there is a match between what I (or you) plan *now* to do in the hypothetical scenario, and what I (or you) plan *in the hypothetical scenario* to do. We can distinguish between these two aspects of my hypothetical plan; following Gibbard (2003, 50) and Hare (1981), we can call my current plans for a hypothetical scenario an *actual-for-hypothetical* plan, since it is a plan I now have for a situation that is potentially not mine. Meanwhile, the plan I have *in* a hypothetical scenario for that *same* scenario is a *hypothetical-for-hypothetical* plan. These notions can come apart, which is illustrated by the case in Gibbard (2003, 51-2) of the binge alcoholic who wants now not to drink, but knows that on Saturday night, he will want to drink. His actual-for-hypothetical plans for Saturday night diverge from the hypothetical-for-hypothetical plans for the same evening. I will, in what follows, be speaking exclusively of actual-for-hypothetical plans.

[92] The literature surrounding the Frege-Geach question, beginning with Geach (1960), addresses issues surrounding whether the Expressivist resources are adequate to explain inconsistency- and entailment-relations between sentences containing normative language. See Schroeder (2008b) for an especially insightful discussion.

explain the data provided by Moral Twin Earth scenarios. Importantly, reference need not be in the picture at this point; it emerges at a later stage.

3.5.2 Disagreement and Moral Twin Earth scenarios

Moral Twin Earth scenarios, I claimed in the previous section, bring out the importance of the realist's commitment to a high degree of fundamentality in the normative. The T_C- and T_D-communities from Horgan and Timmons' original Moral Twin Earth scenario, for instance, might be said to refer to the a single property which is highly fundamental and hence highly eligible for reference. An appeal to reference magnetism thus explains why communities in this kind of scenario semantically disagree with each other.[93] The realist explanation therefore has a particular structure, explaining (some of) the facts about disagreement in terms of antecedently explained facts about reference:

$$\begin{tikzcd}
\textbf{Reference} \arrow[r] & \textbf{Semantic Disagreement} \\
\textbf{Eligibility} \arrow[u] &
\end{tikzcd}$$

Fig. 3.1: The explanatory structure of realism

For the Expressivist, the explanation for the semantic disagreement between communities in Moral Twin Earth scenarios proceeds in terms of the notion of Plan-Disagreement. Recall the T_C- and T_D communities from earlier (though this is just for concreteness; the Expressivist should give an explanation on this model for whichever Moral Twin Earth scenarios contain genuine semantic disagreement). Since these communities use their normative terms in accordance with different normative theories, T_C-community will accept the sentence **A**, while the T_D-community rejects it:

A In circumstance c, one ought to ϕ

(where, as before ϕ-ing in circumstance c maximizes total happiness but fails in c to respect the humanity in some agents in the circumstance). The communities disagree over **A**, and this disagreement, according to the Expressivist, is constituted by a difference in *plans* for circumstance c.

[93] To reiterate, it is not at this point crucial to the argument that reference magnetism ground a disagreement between the T_C- and T_D-communities specifically. All that we need, for present purposes, is that *some* semantic disagreements between communities that diverge in use are explained on this model.

In more detail: speakers in the T_C-community plan to ϕ in c, while speakers in the T_D-community plan to refrain from ϕ-ing in c. This is what it is for speakers in the former community to accept **A** and for speakers in the latter community to reject **A**. These plans are inconsistent in the sense of PLAN-DISAGREEMENT. One cannot coherently both plan to do ϕ in c *and* plan to refrain from doing ϕ in c. And this is, according to the Expressivist, *all there is* to semantic disagreement. Hence the Expressivist explanation of the semantic disagreement in Moral Twin Earth scenarios bypasses reference, proceeding entirely in terms of the notion of disagreement in plan:

$$\textbf{Plan-Disagreement} \longrightarrow \textbf{Semantic Disagreement}$$

Fig. 3.2: Explanatory structure of Expressivism: disagreement

3.5.3 Plans and reference

Thus the realist and Expressivist explain the Moral Twin Earth data differently—the realist via a feature of normative properties (that they are highly eligible for reference), and the Expressivist via a feature of normative thought (that thinking that ϕ ought not to be done in c *just is* disagreeing in plan with communities who think otherwise). It is tempting, at this point, to say that we have arrived at a metaphysical difference between the views, and it is a simple difference over simple *existence* (no talk of relative fundamentality needed!). The reason is simply that the Expressivist explanation of Moral Twin Earth scenarios does not need reference to normative properties, so unlike the realist, they deny that normative terms refer to normative properties.[94]

But to say this would be to fail to appreciate the lesson from §1. Not only *can* Expressivists use minimalist techniques to explain why claims about reference to normative properties are acceptable, they need no further resources to do so: the tools used to explain the semantic disagreement in Moral Twin Earth scenarios are sufficient to do the job. Claims about reference can be explained for the Expressivist by appeal to planning-states and disagreement in plan. Here is how, in essentials, the account should go.[95]

[94] Thus the existence based difference, on this proposal is that realists accept

$\exists x$: x is a normative property and normative terms refer to x,

while Expressivists deny it.

[95] Much of what follows appears in Gibbard (2013), especially Ch. 2. Although the details in presentation are my own, the insights are due to Gibbard in his attempt to work out an Expressivistic account of *meaning* and closely related notions, which include reference.

The ultimate question we want an answer to is the question of what property 'ought' and other normative terms refer to in the mouths of speakers in the T_C-community, and what they refer to in the mouths of speakers in the T_D-community. Note that for the Expressivist we are not forced to say these communities refer to the same property *simply because* they disagree with each other. For the Expressivist, the communities disagree with each other because they *mean* the same thing with their normative terms. And speakers might mean the same thing with a term and yet differ in what they refer to with that term—you and I mean the same thing with 'I' yet refer to different people with the term.[96] Here, I will explain the Expressivist position on reference by addressing, in Expressivist terms, the simple question of what property is referred to by 'ought' in *our* mouths. Complications in extending claims about reference to other communities who use their terms differently are taken up in Appendix B to this chapter.

We can take guidance from our §1 discussion to see how the Expressivist can go about explaining that normative terms refer to certain properties. The trick is, as before, to find a *normative commitment* in the claims she is aiming to explain. She can then apply her characteristic indirect method of explanation to explain why the normative commitment in question is acceptable. As a way of approaching this strategy for explaining claims about what we refer to, begin with the following: for some natural property *P*, the following biconditional is acceptable:

B Necessarily, all and only things that have *P* are things one ought to do.

Instances of **B** carry normative commitments. Hedonistic consequentialists, for example, will accept an instance of **B** where *P* is replaced with 'the property of maximizing happiness'. Let us then suppose, for the sake of exposition, that this claim (which we can encode as $\mathbf{B}_C$) is true:

$\mathbf{B}_C$ Necessarily, all and only things that maximize total happiness are things one ought to do.

[96] Although this contextualist analogy for normative terms seems to be obviously inconsistent with the basic disagreement- and inconsistency-related data for normative terms, things are actually not so simple with an expressivistic semantics in place. For if two communities diverge over what they refer to with 'ought', it doesn't directly follow for the Expressivist that they don't disagree or make inconsistent claims. For the latter question is a question of whether they disagree in plan, and this is *prima facie* distinct from a question about reference. As I will argue below, however, the Expressivist is committed to normative terms not being contextually sensitive in this crude way, though it is a highly indirect relationship between this claim and the facts about disagreement and inconsistency.

$\mathbf{B}_C$ is acceptable to the Expressivist for a familiar reason: to accept $\mathbf{B}_C$ is just to accept its normative commitments. And it is coherent to accept these commitments, since such acceptance is constituted by planning, for a wide variety of hypothetical scenarios, to do the thing which maximizes total happiness.

$\mathbf{B}_C$ is simply a material mode claim about what ought to be done; nothing about the normative word 'ought' (and what it refers to) has been said yet. But it is straightforward to show that *if* $\mathbf{B}_C$ is acceptable, then so is a claim about the reference of 'ought'. Start first with a characterization of the conditions under which a word means what our word 'ought' means. This should be relatively straightforward within an expressivistic framework: 'ought' is a vehicle for expressing plans of a certain kind, namely plans for how to act and how to feel in various hypothetical scenarios. Thus any community whose use of 'ought' has these features is using that word with the same meaning. Say that these communities use 'ought' with its *normative meaning.*[97]

Given this stipulation, the expression 'satisfies 'ought' on its normative meaning' then applies to all and only the things our use of 'ought' applies to, across all possible worlds. In other words, **S** is true:

S Necessarily, all and only things that satisfy 'ought' on its normative meaning are things one ought to do.

By $\mathbf{B}_C$ and **S**, it follows that, 'ought' on its normative meaning applies to things that maximize total happiness. That is:

$\mathbf{S}_C$ Necessarily, all and only things that satisfy 'ought' on its normative meaning are things that maximize total happiness.

At this juncture, two points need to be made about $\mathbf{S}_C$. The first is that, like **A** and $\mathbf{B}_C$, it encodes a normative commitment. For the Expressivist, to accept it is to accept some constraints on how one plans. It is, to follow Gibbard's terminology, *plan-laden* in the following sense:

PLAN-LADEN A sentence *S* is *plan-laden* just in case there is some planning-state *p* such that being in *p* constitutes accepting *S*.

[97] 'ought' *could* be used to mean lots of other things—some possible communities use the word 'ought' to mean what we mean by 'green'. But these communities aren't expressing the right kind of plans tied to action and emotion. Hence they do not use the term with its normative meaning. For similar reasons, communities who use lexicographically distinct words with the same meaning will count as using these terms with the normative meaning of 'ought'.

Thus, **A** is plan-laden in this sense because to accept **A** is to plan to do ϕ in c. And $\mathbf{B}_C$ is plan-laden in the same way because to accept it is to be committed to doing all and only things which maximize total happiness across all possible scenarios.

$\mathbf{S}_C$ is likewise plan-laden, and this is because it *entails* $\mathbf{B}_C$. The entailment holds because something can't satisfy 'ought' on its normative meaning without being something one ought to do. This makes $\mathbf{S}_C$ plan-laden for the following reason. We have already noted that $\mathbf{B}_C$ is plan-laden, and so $\mathbf{S}_C$ entails a plan-laden claim. I will leave as an open question exactly *how* the Expressivist accounts for entailments like this; the only assumption I will make about the explanation is that it bars *naturalistic-to-normative* entailments. That is, a sentence which isn't plan-laden in the way normative sentences are cannot entail a plan-laden, normative sentence. This is roughly because naturalistic sentences serve to describe the world in a way which places no constraints on how to plan. $\mathbf{S}_C$, then, must be plan-laden as well. It it wasn't plan-laden, it couldn't entail $\mathbf{B}_C$, because of the ban on naturalistic-to-normative entailments.

The second point to make about $\mathbf{S}_C$ is that it is, or is at least trivially equivalent to, a claim about the reference of 'ought' on its normative meaning. After all, it gives an informative characterization about what things satisfy 'ought' on its normative meaning across every possible world, claiming that the things in the extension of the term at a world are those that maximize total happiness in that world. Thus if $\mathbf{S}_C$ is plan-laden, $\mathbf{R}_C$ must be as well:

$\mathbf{R}_C$ 'ought' on its normative meaning refers to the property of maximizing total happiness.

We have shown two things: that claims about reference like $\mathbf{R}_C$ are acceptable to the Expressivist, and that they are so because they are plan-laden. They are, in other words, acceptable for exactly the same reason that sophisticated claims like MI and PTC were acceptable to Expressivists: they embody a normative commitment, and the Expressivist can then explain why the commitment is acceptable by use of the by-now familiar indirect method of explanation.

Obviously, exactly *which* claim about the reference of 'ought' we accept will depend on our substantive normative views. But the upshot is relatively simple: first-order normative claims like **B** are entailed by a claim about the reference of 'ought'. And since first-order normative claims like **B** are plan-laden, analogous claims about the reference of 'ought' must be plan-laden as well.

3.5.4 Expressivist reference on Moral Twin Earth: a first pass

This is an account of what ‘ought’ in *our* mouths refers to. It does not go all the way toward explaining each of the claims about reference that the realist appeals to in explaining the disagreement in Moral Twin Earth scenarios. The realist explains the disagreement in these cases by claiming that the reference of ‘ought’ as used by *other* communities is the same as the reference of ‘ought’ in our mouths. So in order to show that all of the realist’s claim about reference are acceptable to the Expressivist, we need to show more than just that claims like $\mathbf{R}_C$ are acceptable; we also need to show that analogous claims about the reference of ‘ought’ in the mouths of speakers in *other* communities are acceptable too.

I show in Appendix B that the Expressivist commitments outlined here entail that she is committed to accepting the same claims about reference as the realist—accepting, that is, that both communities in a Moral Twin Earth scenario refer to the *same* property. The details proceed, in essentials, with the same resources of planning-states and Plan-Disagreement that we needed to explain the simple reference claim $\mathbf{R}_C$. Here, I want to focus on the consequences of explaining facts about reference in this way. In particular, I want to focus on the fact that these resources explain, for the Expressivist, the very claims about reference that the realist explains with reference magnetism. The Expressivist then has a picture with the following explanatory structure:

$$\begin{tikzcd}
\textbf{Plan-Disagreement} \arrow[r] \arrow[d] & \textbf{Semantic Disagreement} \\
\textbf{Reference} &
\end{tikzcd}$$

Fig. 3.3: The explanatory structure of Expressivism: reference

In a sentence: what the realist explains with reference magnetism, the Expressivist explains, via the indirect method, in terms of planning-states. Given a framework in which metaphysical commitments are commitments for claims about relative fundamentality, this gives rise to a substantial difference in the metaphysical commitments of the Expressivist. I will conclude by showing why this difference in explanatory structure gives rise to a difference in metaphysical commitments.

3.6 Conclusion: Expressivism and relative fundamentality

We are now in a position to answer the metaphysical version Dreier's problem, as set out in §2. The problem, we will recall, was set out in terms of the notion of relative fundamentality, which allows for the phenomenon of reference magnetism. Within this framework, it is very natural to suppose that the realist is committed to normative properties having a high degree of fundamentality, and to use this commitment in explaining the disagreement in Moral Twin Earth scenarios. §5 walks through the steps the Gibbard-style Expressivist can go through to accommodate the same claims about reference and semantic disagreement. And the important take-away lesson from that section was that they can be explained, but the explanatory directions differ. The question is now why this matters.[98]

The answer is straightforward within a fundamentality-centric framework. The metaphysical commitments about the normative, for the Expressivist, arise from the conjunction of the following two facts: (i) nowhere does the Expressivist explanation of what different communities refer to with normative terms require that candidate referents be highly fundamental, and (ii) relative fundamentality is a kind of explanation, and so considerations of parsimony require that we not posit a degree of fundamentality for a property beyond what is needed. Given (i) and (ii), Expressivism is committed to denying that, from the perspective of the reference of normative terms, normative properties are highly fundamental.[99]

[98] The difference in explanatory directions is by itself not entirely new. Gibbard hints at it when he says, "Once we distinguish properties from concepts, we have no need for non-natural properties to help us explain the special features of normative concepts." (Gibbard (2003, 181)) And Dreier says:

> [I]f the difference between normative (or evaluative, or "planning") concepts and descriptive (naturalistic) ones can also be stated as a difference in content, then at least it must be a comprehensible, substantive question whether the difference in concept is explained by (or if you prefer amounts to no more than) a difference in content, on the one hand, or rather it is explained by (amounts to) something else entirely, which in turn explains the difference in content. The divide between realism and irrealism, at least in meta-ethics, rests on the substance of questions about metaphysical explanation. (Dreier (2004, 42))

What both Gibbard and Dreier have to say seems right as far as it goes, but merely pointing out *some* difference in direction of explanation, as they do, does not go far enough to show anything like a *metaphysical* difference between the views. There are lots of differences of this kind that have no consequences for metaphysics.

[99] Here it is worthwhile to point out the import of rejecting a Lewis-style account of relative fundamentality in terms of length of canonical definition. For the realist and Expressivist will presumably *agree* on how long the microphysical definition of a property like maximizing total happiness is. (Moreover, the definition will in all likelihood be impossibly long, which is not a promising start for a view which

The realist, of course, is committed to normative properties being highly fundamental—this is required by the need to explain co-reference and semantic disagreement by appealing directly to the fundamentality of normative properties. In particular, the relevant explanation comes in the form of an appeal to reference magnetism which, in the framework of metaphysical fundamentality we are assuming, explains how communities which diverge in their use of terms manage to co-refer. Thus within the fundamentality-centric framework, we have found a difference in the metaphysical commitments of realism and Expressivism. The realist takes the normative to be highly fundamental; while the Expressivist is not free to adopt the same claim.

It remains a possibility, given what we have said here, that the Expressivist will need highly fundamental normative properties to explain different phenomena, and that this will eliminate any difference over the relative fundamentality of normative properties between the Expressivist and the realist. Although this is indeed a live possibility—and one which I have no direct argument against—it should be noted that such a possibility is highly unlikely. The Expressivist program is to explain the features of normativity that are of central concern to metaethicists in terms of their theory of normative terms/concepts. And this explanatory strategy, which eliminated the need for relative fundamentality in Expressivist explanations of the Moral Twin Earth phenomena, will plausibly eliminate the need in any other area of meta-ethical concern as well.

Thus we can tentatively conclude that we have found a metaphysical difference between realism and Expressivism, in the form of a difference over the relative fundamentality of the normative. What we have said so far, moreover, points toward a promising general strategy for resolving a version of Dreier's problem within a fundamentality-centric framework. This is because might be other aspects of the realist's view which, like her claims about reference, are symptoms of her commitment to a high degree of fundamentality in the normative. We can then follow a similar strategy for showing that the metaphysical commitments of Expressivism differ from the realist's by showing the normative properties are not fundamental in the same way on the Expressivist view. If the Expressivist can explain them at all, she will do so by indirect appeal to planning states and disagreement in plan.

A final point concerns the underlying framework of the argument. We granted at the

wants the properties referred to by normative terms to retain a significant degree of fundamentality.) But once we follow Hawthorne (2006, 2007) and allow the relative fundamentality facts to float free of length of canonical definition, it becomes an open possibility that the explanatory needs of the realist and Expressivist give rise to a difference in relative fundamentality.

outset that the fundamentality-centric framework where highly fundamental properties are reference magnets—a crucial assumption on which the entire argument depends—is controversial. Not everyone will be willing to accept it. One general point to make on this issue is that it is unlikely that any progress can be made on this issue absent some fairly specific metaphysical framework—one thing quasi-realist Expressivism shows is that benign garden-variety metaphysical notions like truth, fact, property, etc. are inadequate for articulating what is metaphysically distinctive about Expressivism. Once we go beyond these benign tools, any specific framework we adopt is bound to be controversial. But this isn't by itself objectionable; some controversial stance in the area is needed to gain leverage on the issue. This leads to a positive point in favor of the fundamentality-centric framework. Insofar as it is a controversial framework that also gives offers a clean solution to a difficult issue surrounding Expressivism, it thereby enjoys a *prima facie* point in its favor. Of course, it could be that other foundational approaches to metaphysics yield the same results; I have not tried to show that they do not. If they do, then the *prima facie* case in favor of the fundamentality-based framework disappears. But it is by no means obvious that other approaches will in fact yield the same results. This then constitutes a tentative point in favor of the legitimacy of fundamentality in metaphysical theorizing, and gives those who are skeptical of the framework some reason to rethink their opposition.[100]

[100] Additional thanks for discussion of this chapter go to Elizabeth Barnes, Ross Cameron, Gerald Lang, Sarah Moss, David Plunkett, Pekka Väyrynen, and others from audiences at the University of Leeds and the University of Michigan.

Appendix A: Generalized plan-disagreement and reference

This appendix aims to show that the account of the Expressivist's metaphysical explanation of the facts about reference does not rely on any idiosyncratic features of Gibbard's Expressivist view, which I use for exposition. I proceed by trying to identify the general versions of the central claims of that argument. Although I don't explicitly argue this here, I think that views which satisfy these general assumptions will intuitively differ from the realist view in the way Gibbard's does. They can also be shown to exhibit the same difference with the realist over the fundamentality of the normative.

The argument of §5—that Expressivists can explain the facts about semantic disagreement and reference that are at issue in Moral Twin Earth scenarios, albeit in a way that differs significantly from the realist's explanation of the same claims—proceeds with the Expressivist resources developed in Gibbard (2003). The details of the Gibbard's development of the view in terms of planning states allowed us to work out a fairly concrete explanation of claims about reference and semantic disagreement on Expressivist-friendly terms. But these details were in large part unnecessary; all that is needed are some quite general assumptions about how the Expressivist view works, and in this appendix I will re-trace the argument of §5 at a more abstract level.

For semantic disagreement between the T_C- and T_D-communities over a sentence like **A**:

A In circumstance c, one ought to ϕ,

the Expressivist needs only to have the following resources. The first is an account, in entirely descriptive terms,[101] of the non-primarily-descriptive semantic function served by normative sentences like **A**. That is, for normative sentences S,

SEMANTIC FUNCTION The primary semantic function of S—i.e., the most perspicuous explanation of the meaning of S—is not that of describing the world as being a certain way, but rather that of expressing a state of mind with broadly world-to-mind direction-of-fit.[102]

One caveat about SEMANTIC FUNCTION is immediately in order. The caveat is that SEMANTIC FUNCTION is *not* inconsistent with saying that normative sentences describe the world in some way—to deny this would be to fail to appreciate the lessons from §1.

[101] For more on the significance (and restrictions on) accounts that proceed in "descriptive" terms only, see Chapter 4. I will ignore the complications introduced there for the sake of exposition.

[102] For the canonical introduction of direction of fit, see Anscombe (1957, §32).

All it makes a claim about is the primary, or most basic, way in which the meaning of normative sentences is explained.

The second feature needed for an Expressivist explanation of the disagreement data surrounding **A** is an account, again in purely descriptive terms, of a disagreement-like relation the entities in the *explanandum* of SEMANTIC FUNCTION can bear to each other. This is the role played by Gibbard's notion of PLAN-DISAGREEMENT, which we can generalize as follows:

GENERALIZED DISAGREEMENT There is some relation R such that, for (a wide range of) normative sentences S_m, S_n that are intuitively inconsistent with each other, the primary semantic values S_m and S_n bear R to each other.

R must, moreover, be a relation which has some intuitive connection with inconsistency. Simply listing the ordered pairs of states that intuitively disagree with each other is *not* an acceptable way of specifying the disagreement relation GENERALIZED DISAGREEMENT specifies.

Many views in the non-cognitivist or Expressivist tradition satisfy—or at least aim to satisfy—SEMANTIC FUNCTION and GENERALIZED DISAGREEMENT. In other Expressivist-like views in the literature, the entities with the appropriate semantic function take the form of attitudes of approval and toleration in Blackburn (1988c), states of *being for* in Schroeder (2008a), and imperatival force in Hare (1949, 1952). The disagreement-relations in these views in these views are as varied as the basic semantic machinery; they are a primitive incompatibility relation between approval and toleration (in Blackburn), logical inconsistency between contents of *being for* attitudes (in Schroeder), logical inconsistency between the contents to which imperatival force is applied (Hare (1949)), or impossibility of coherent simultaneous acceptance of two imperatives (in Hare (1952)).

Views which satisfy SEMANTIC FUNCTION and GENERALIZED DISAGREEMENT will, with a few minimal assumptions in place, be able to explain the semantic disagreement between the T_C- and T_D-communities over a sentence like **A**. An account of the semantic function of utterances of **A** and its negation in the mouths of speakers in the T_C- and T_D-communities will associate different functions with each utterance. After all, the communities are inclined to *act* in very different ways, and given SEMANTIC FUNCTION, this should give rise to a difference in the semantic values of **A** in the mouth of the T_C-community and its negation in the mouths of the T_D-community. Call these values s_c and s_d, respectively. It is a basic aim of Expressivist-style views, regardless of the details, to be able to find a relation satisfying (for the most part) GENERALIZED DISAGREEMENT. Hence, *if* these views meet a basic adequacy constraint, they will specify a relation satisfying GENERALIZED DISAGREEMENT that holds between s_c and s_d. This completes the explanation of the semantic disagreement between the two communities.

There is a further question of whether Expressivists can accept claims about reference, such as $\mathbf{R}_C$, and whether they need highly fundamental reference-magnetic normative properties to explain why these claims are acceptable:

$\mathbf{R}_C$ 'ought' on its normative meaning refers to the property of maximizing total happiness;

The Gibbard-style explanation from §5 explained the acceptability of these claims without appeal to reference magnetism. And any of the views described above can do the same, given the following claims.

The first is that claims about the reference of 'ought' (if meaning is held fixed) have normative consequences:

NORMATIVE CONSEQUENCES The claim expressed by ⌜ 'ought' refers to Φ⌝ has normative consequences—namely, that one ought to do all and only things that satisfy Φ.

NORMATIVE CONSEQUENCES, when combined with a constraint on entailment-relations between sentences, has direct consequences for the semantics for claims about reference like $\mathbf{R}_C$. The constraint on entailment-relations is a generalization on the implementation of the ban on naturalistic-to-normative entailments in §5. In particular:

NORMATIVE ENTAILMENT For any sentence S that entails a normative claim N, S satisfies SEMANTIC FUNCTION and bears a relation of the kind described in GENERALIZED DISAGREEMENT to the negation of N.

That is: any sentence that entails a normative claim has a semantics of the kind described by the Expressivist; no non-normative statement could have these normative consequences.

Hence—given NORMATIVE CONSEQUENCES—claims about reference entail sentences that must be treated with the same semantic resources the Expressivist uses to explain ordinary normative claims. Given NORMATIVE ENTAILMENT, normative sentences must *themselves* be treated with these resources. Failing to do so is to allow that non-normative sentences can entail normative sentences.

Thus our explanation of the basic explanatory structure of reference and disagreement in Expressivism will apply to any Expressivist view that meets some very minimal conditions. What the details of the account of reference look like will be as variable as the Expressivist-style treatments of first-order normative claims.

Appendix B: Cross-community planning and co-reference

In §5, we argued that there are some planning-states that constitute acceptance of $\mathbf{R}_C$.

$\mathbf{R}_C$ 'ought' on its normative meaning refers to the property of maximizing total happiness.

This was restricted to claims which, like $\mathbf{R}_C$, are about the reference of 'ought' *in our own mouths*. In this Appendix, I will show how the account generalizes to an account for what 'ought' as used by different communities refers to. Moreover, this extension tells us *which* properties other communities refer to with 'ought'—they refer to the same properties we refer to. As with the discussion in the main text, this generalization owes much to the work in Gibbard (2013). But as the details below show, the results here are non-trivial: once we appreciate the Expressivist distinction between meaning and reference, it isn't straightforwardly *required* on the Expressivist picture that communities who disagree with each other refer to the same properties. Thus even though we will end with the conclusion that the disagreeing communities co-refer, the denial of such a conclusion is, in some sense, a live possibility.

This is of some interest given the different kind of projects we discussed at the beginning of this chapter. There, we noted that a solution to the "semantic" problem need not constitute a solution to the metaphysical questions raised by the problem of creeping minimalism. By showing that Expressivism can be committed to a difference in fundamentality in the normative without also being committed differing with the realist over claims concerning the reference of normative terms in the mouths of other communities, we can show that the present solution to the metaphysical problem does not carry with it the claim that Expressivists must reject some non-metaphysical sentences that realists accept.

Begin with the question of which plan $\mathbf{R}_C$ expresses. This is a question that goes beyond the main text—all we argued there was that, owing to the ban on naturalistic-to-normative entailments, $\mathbf{R}_C$ must express *some* planning-state. On a Gibbard-inspired proposal, $\mathbf{R}_C$ expresses a plan for what to do with the word 'ought' in various hypothetical circumstances. In accepting that 'ought' on its normative meaning refers to the property of maximizing happiness, one thereby plans to apply 'ought' to all and only things that have this property.[103]

This answer gives us a clue about how to make claims about the reference of *others* within the Expressivist framework. To see this, we can consider an instructive example from Gibbard (2013, Ch. 2). The context for Gibbard's example is one in which he assumes that 'means' is a normative term, and hence subject to Expressivistic treatment. He then asks which property it refers to, and his answer is instructive for our own case.

In a variant of the Kripkenstein example in Kripke (1982), the hypothetical subject Quursula has no dispositions for use of the sign '+' when presented with queries involving '+' and numbers over 50. (She simply answers randomly for such questions, or looks confused and doesn't answer at all.) For simpler questions involving '+', she reliably adds. Her surrounding community, however, not only adds for the simple questions, they *also* take a stance on the more complicated questions—they respond with the *quum*, answering '5' for all of them.

Now consider two different views about meaning—Tyler's, on which an individual's dispositions and the dispositions of her linguistic community determine what she means, and Jerry's, on which an individual's dispositions and the relative simplicity of candidate referents determine what she means. Tyler and Jerry hold two incompatible meta-semantic theories, and Gibbard describes their disagreement in the familiar Expressivist way, making reference to the *plans* that constitute acceptance of the various claims, and pointing out that these plans disagree in the sense of PLAN-DISAGREEMENT. Gibbard says:

> An Expressivist could further maintain that the issue is just this: which statements of Quursula's to accept if one is Quursula. At issue between Jerry and Tyler is at least this: whether, if one is Quursula, to accept or reject the sentence she writes '68 + 57 = 125' [...] It is a difference on what to accept if in a hypothetical circumstance, that of being Quursula.[104]

The disagreement in plan is a disagreement between plans of the same kind that constitute acceptance of sentences like $\mathbf{R}_C$. Acceptance of $\mathbf{R}_C$, we said, amounts to planning to use 'ought' in a certain way in certain hypothetical circumstances. Similarly, Tyler and

[103] Here I restrict attention to circumstances where 'ought' meets conditions for having its normative meaning—there are obviously hypothetical circumstances where 'ought' means what 'green' actually means. One need not plan to apply 'ought' to non-green utility-maximizing things in these circumstances in order to count as accepting $\mathbf{R}_C$.

[104] Gibbard (2013, 46)

Jerry's disagreement is, according to Gibbard, a disagreement in plans concerning use of the sentence '68 + 57 = 125' in hypothetical circumstances.

So far, this is just an extension of the Expressivistic explanatory strategy to meaning taken normative. But our main concern is about *reference*, and this falls out of the story in the way we suggested in §5. Jerry and Tyler's disagreement over how to use '+' in various hypothetical circumstances, Gibbard says, is *also* a disagreement over which property 'means PLUS' refers to:

> If Jerry is right, then in both Jerry's and Tyler's mouth, the phrase 'means PLUS' signifies the property P_J, intrinsic to the brain [i.e., the property of scoring the best on both metrics of fit with the individual's linguistic dispositions and simplicity]. If Tyler is right, then 'means PLUS' in both their mouths signifies the community property P_T [i.e., the property of being the best fit with the linguistic dispositions of both an individual and her linguistic community].[105]

Thus Gibbard is claiming that the disputants, Jerry and Tyler, refer to the *same* property (though we can't say which property is the common referent without taking a stance on the meta-semantic dispute between Jerry and Tyler). For instance: if Jerry is right, then they both refer with the predicate 'means PLUS', to the property P_J of scoring highest on the metrics of fit with use of '+' and simplicity. Similarly, if Tyler is right, the common referent is the community property P_T. Once we identify the reasoning behind this claim, we will be able to extend it to show that other communities who use 'ought' differently nonetheless use it to refer to the same property we do (and as before, saying which property is the common referent requires taking a first-order normative stance).

Why do Tyler and Jerry refer to the same property? Suppose we deny this, thereby allowing that Tyler refers to P_T with his expression 'means PLUS' (while Jerry continues to refer to P_J). Jerry is then committed to the pair of claims

$\mathbf{R}_J$ 'means PLUS' refers to the property P_J of scoring the highest on the metrics of fit with linguistic dispositions and simplicity;

$\mathbf{R}_T$ 'means PLUS' in the mouth of Tyler refers to the property P_T of fitting the best with community-wide linguistic dispositions.

Given that claims about meaning are plan-laden, the plans that $\mathbf{R}_J$ and $\mathbf{R}_T$ express disagree in plan. Take $\mathbf{R}_J$ first—Jerry accepts it, and thereby plans, for any hypothet-

[105] Gibbard (2013, 49), my notation.

ical scenario where ‘means Plus’ has its actual meaning, to apply it to speakers who instantiate P_J. Since Tyler’s circumstance is one where ‘means Plus’ retains its actual meaning, it is one for which Jerry, in virtue of accepting R_J, plans to apply ‘means Plus’ to speakers who instantiate P_J.

But to accept $\mathbf{R}_T$, Jerry would need to plan, for the circumstance of being Tyler, to apply ‘means Plus’ to speakers who instantiate P_T instead. These circumstances where one is Tyler are a *subset* of the hypothetical circumstances of scenarios for which Jerry plans to apply ‘means Plus’ to speakers who instantiate P_J. It is incoherent to plan, for some circumstance, to apply ‘means Plus’ only to speakers who instantiate P_J *and* only to speakers who instantiate P_T. So we learn, from the fact that Jerry and Tyler disagree, that their use of ‘means Plus’ refers to the same property. *Which* property this is is a question of which substantive view about meaning is correct.

Things go similarly for claims about the reference of ‘ought’ in Moral Twin Earth scenarios. That is, we can prove in exactly the same way as before, that the attitudes which $\mathbf{R}_C$ and $\mathbf{R}_D$ express disagree in plan:

$\mathbf{R}_C$ ‘ought’ on its normative meaning refers to the property P_C of maximizing total happiness;

$\mathbf{R}_D$ ‘ought’ in the mouths of speakers in the T_D-community refers to the property P_D of respecting the humanity in others.

$\mathbf{R}_C$ is a claim about what ‘ought’ in our mouths refers to; $\mathbf{R}_D$ is a claim about what the same term (with the same meaning) refers to in the mouths of the T_D-community. As with Jerry and Tyler, these sentences express plans that disagree. In accepting $\mathbf{R}_C$, one plans to apply ‘ought’ with for any hypothetical scenario where ‘ought’ has its normative meaning to apply it to all and only things that have P_C. One thereby plans (on pain of incoherence) to do the things which instantiate P_C. Yet in accepting $\mathbf{R}_D$, one plans for those hypothetical scenarios where one is in the T_D-community to apply ‘ought’ to the things that instantiate P_D, and hence to avoid doing some things have P_C. We need to reject $\mathbf{R}_D$ and similar instances, in just the same way that Jerry needs to reject $\mathbf{R}_T$. We need to reject these claims on which others refer to different properties in order to avoid incoherence in plans for what to do in hypothetical scenarios with the relevant terms.

As with Jerry and Tyler, *which* property different communities use ‘ought’ to refer to is a first-order normative question. Thus those who disagree with the first-order normative assumption built into $\mathbf{R}_C$ will need to revise the example, but the general point about co-reference remains.

Thus all of the claims about reference that the realist explains can also be explained by the Expressivist, via her indirect method of explanation. There need not be a semantic difference over the reference of normative terms between realists and Expressivists. The difference between realism and Expressivism lies solely in the *manner* of explanation of their claims about reference, not the claims about reference themselves. Metaphysical fundamentality, as outlined in §3, is a promising resource for capturing the metaphysical significance of this explanatory difference.

Chapter 4: Supervenience arguments and normative non-naturalism

4.1 Defining non-naturalism

Frank Jackson (1998) gives an argument against familiar non-naturalist views about the normative which has been endorsed, in essentials, elsewhere in the metaethics literature. (The most prominent endorsements include Brown (2011) and Streumer (2008)). The primary aim of Jackson's argument (and those modeled on it) is to establish the DESCRIPTIVISM thesis, which is characterized as follows:

DESCRIPTIVISM Every normative property is identical to a descriptive property.

Descriptive properties, for Jackson, are just properties which can be expressed with descriptive language—that is, with language that includes no normative vocabulary such as 'right', 'good', 'reason', etc.[106] DESCRIPTIVISM thus implies that every normative property can be expressed using descriptive vocabulary only. (We will return in §2 to the question of *why*, according to Jackson et al., DESCRIPTIVISM is supposed to be true.)

Most parties to the debate—both those friendly to Jackson's argument and those concerned to resist it—assume that *if* DESCRIPTIVISM is true, then the traditional non-naturalist views about the normative found in Moore (1903) and elsewhere are false.[107] Let us label this thesis IMPLICATION:

IMPLICATION If DESCRIPTIVISM is true, then non-naturalism about the normative is false.

[106] Jackson (1998, 113, 117). Gibbard (2003, 99) draws the same distinction using the term 'natural', stipulating that supernatural, mathematical and psychological properties count as "natural" in the relevant sense. Brown (2011) doesn't explicitly accept the same definition of 'descriptive property', presumably on the grounds that his version of the argument is supposed to avoid the "linguistic detour" present in Jackson's. He fails, however, to offer an alternative characterization of the notion. I will not try to settle this question for Brown; but it should be clear that the points I make against Jackson's argument should apply *mutatis mutandis* to Brown's version *if* he were to accept the same characterization of what descriptive properties *are*.

[107] In addition to Brown (2011) and Streumer (2008), the discussion of the argument by non-naturalists such as Shafer-Landau (2003, 94 ff.), Fitzpatrick (2008, 199) and Suikkanen (2010) do nothing to question this claim. Instead, they argue against a prior step concerning the metaphysics of properties—according to which any properties that share an intension are identical—to block the argument to DESCRIPTIVISM. Schmitt and Schroeder (2011, 146-7) label the same step as a weakness in Jackson's argument.

Jackson claims to find a commitment to the denial of DESCRIPTIVISM at the center of paradigmatic non-naturalist views. Speaking of Moore (1903), he says:

> What he really wants to insist on, I think, is an *inadequacy* claim: what is left of language after we cull the ethical terms is in principle inadequate to the task of ascribing the properties we ascribe using the ethical terms. He wants to object to exactly the claim I will be making.[108]

This, however, is not *obvious* given the descriptions non-naturalists provide for their own view. Moore, for instance, preferred (at one point) to explain his view in terms of the absence of a certain kind of *definition* of normative properties:

> When we say, as Webster says, 'The definition of horse is "A hoofed quadruped of the genus Equus," ' we may, in fact, mean three different things. (1) We may mean merely: 'When I say "horse," you are to understand that I am talking about a hoofed quadruped of the genus Equus.' [...] (2) We may mean, as Webster ought to mean: 'When most English people say "horse," they mean a hoofed quadruped of the genus Equus.' [...] But (3) we may, when we define horse, mean something much more important. We may mean that a certain object, which we all of us know, is comprised in a certain manner: that it has four legs, a head, a heart, a liver, etc., etc., all of them arranged in definite relations to one another. *It is in this sense that I deny good to be definable.* I say that it is not composed of any parts, which we can substitute for it in our minds when we are thinking of it.[109]

I do not wish to treat Moore's comments on the non-naturalist view as definitive.[110] What I do wish to point out is that there are plausible senses of 'definition', as Moore explains it, which are stronger than Jackson's inadequacy claim. That is: it makes sense to say that, even though descriptive language is adequate for expressing normative properties, it cannot provide a *definition* of those properties. And if we take Moore's talk of composition on board for a moment, it is easy to see why: even if descriptive language could pick out a normative property, it might not do so by delineating the *parts* of the

[108] Jackson (1998, 121)

[109] Moore (1903, 60), my italics.

[110] This is in part because Moore's own views on the topic were famously (and self-admittedly) confused. What he means by "definable" is not entirely clear (many will not find the language of composition helpful in a discussion of properties like goodness), and he retracted some his earlier claims about non-naturalism in his reply to C. D. Broad in Moore (1942). All I want to establish here that it is worthwhile and coherent to ask, in an investigation of Jackson's argument, whether IMPLICATION is true.

property. Thus, given Moore's conception of definition, normative properties might not be definable in descriptive terms, even though descriptive language is adequate for describing them.[111] The (limited) conclusion I wish to draw here is simply this: some of Moore's claims suggest that non-naturalism is strictly speaking *consistent* with DESCRPTIVISM, and that IMPLICATION is therefore false.[112] It is the aim of the present chapter to motivate and develop this Moorean line of response.

In §2, I set out Jackson's original argument for DESCRIPTIVISM. The argument, in its essentials, claims that DESCRIPTIVISM follows from the global supervenience of the normative on the descriptive plus some additional assumptions about the metaphysics of properties. I then argue in §§3-4 that these additional assumptions imply DESCRIPTIVISM *even under the supposition that the supervenience claim is false.* But, I argue, non-naturalism is clearly *true* under such a supposition. This constitutes a kind of *reductio* of Jackson's auxiliary assumptions including IMPLICATION, showing them to be too strong—even under a supposition where non-naturalism is clearly true, they imply its denial.

Then, in §5, I ask which of the auxiliary assumptions is at fault. My contention will be that the proof of §§3-4 makes it extremely natural to pin the fault on IMPLICATION. In particular, I will show how we can use the notion of *metaphysical fundamentality*, as it is conceived of in the (meta-)metaphysics literature, to formulate a conception of non-naturalism that is consistent with DESCRIPTIVISM. This is an approach which fits with the broad contours of the Moorean approach outlined above, and it has the following virtues: it explains why non-naturalism *must* be true if supervenience fails, and it explains why Jackson's argument fails even when the supervenience assumption is back in place. But it doesn't overgenerate: it doesn't allow us to prove in the same way

[111] Analogies that are friendly to Moore abound here. Given the Ideal Gas Law, the volume of an ideal gas can be described entirely in terms of the amount, temperature, and pressure of the gas. But no one would suggest that volume of an ideal gas has pressure, among other things, as a constituent part. (Note that analogous reasoning would lead to the conclusion that pressure has volume as a constituent part.)

[112] Allowing that descriptive language is adequate to *describe* normative properties, as I will do, is consistent with the central lesson of Moore's famous "Open Question Argument". Moore's arguments, as I understand them, show that there is no *conceptual* connection between normative and descriptive concepts—there is no descriptive concept D which is such that who thinks that something falls under the concept GOOD but not D is making a conceptual mistake (see Moore (1903, 11), also Gibbard (2003, Ch. 2)). Moore apparently thought that the absence of such a conceptual connection showed something important about the *property* of goodness—a highly suspect move. For our purposes here, however, it is sufficient to note that Moore's contention that there is no conceptual definition GOOD is consistent with DESCRIPTIVISM, which does not claim that the identity of normative and descriptive properties is a conceptual truth.

that non-naturalism *must* be true when the supervenience premise is back in place.

Before proceeding, one disclaimer is in order: other ways to reject Jackson's auxiliary assumptions have been discussed in the literature.[113] In particular, opponents of Jackson have focused on the assumptions that feature in the derivation of DESRIPTIVISM, which I describe in §2. (Remember, it is generally a point of agreement between Jackson's friends and opponents that IMPLICATION is true, and so establishing DESCRIPTIVISM is sufficient for arguing against non-naturalism.) I will not have much to say about these alternative routes to rejecting Jackson's auxiliary assumptions here. Rather, my aim will be to use the §§3-4 *reductio* to develop and motivate a rejection of IMPLICATION; the question of whether alternative responses enjoy similar merits is a question for another time.

4.2 Jackson's supervenience argument

4.2.1 Rightness-entailing predicates

As mentioned in §1, Jackson's argument proceeds from apparently minimal assumptions which, when taken together, imply DESCRIPTIVISM. Given IMPLICATION, the falsity of non-naturalism immediately follows.[114]

Jackson begins with the *global supervenience* of the normative on the descriptive:

GS $\forall w, w^*$: if w and w^* are exactly alike descriptively, then they are exactly alike normatively.[115]

He then argues that GS implies that normative predicates are equivalent to descriptive predicates. Here is his explanation of why this follows from a later paper:

> Consider any right action R_1. It must have some particular descriptive nature or other, as it is impossible to be right without having some descriptive nature or other. Let "x is D_1" is be the open sentence that ascribes that nature and also fully specifies descriptive nature elsewhere in R_1's world. It must then be the case that "x is D_1" entails "x is right" [...] Now consider any other right act R_2. With D_2 specified as for D_1 above but with "2" for

[113] See, for instance, the references given above.

[114] Gibbard (2003, Ch. 5) also develops an argument along these lines, though the argument is adapted to a setting where the semantics for normative expressions is expressivist. And Kim (1978) originally outlined the form of argument, abstracting away from Jackson's concern with normative properties in particular. I discuss the general argument at the end of this chapter.

[115] Jackson (1998, 119).

> "1", we get the result that "x is D_2" entails "x is right". From which it follows that "x is D_1 or D_2" entails "x is right". Repeating the process for every right act in logical space, we get "x is D_1 or D_2 or D_3 ..." entails "x is right". But, as we included every right act in logical space, the entailment must also run the other way. We have thus derived the logical equivalence of the infinite disjunctive open sentence "x is D_1 or D_2 or ..." with "x is right".[116]

That is: given a right action, there is a descriptive predicate D which describes the intrinsic features *and* worldly environment of that action. From GS, it follows that there is no other possible action which satisfies D and is not right. Call such a predicate *righness-entailing*. By disjoining each descriptive rightness-entailing predicate—one for each possible right action—we arrive at a big disjunctive descriptive predicate that is equivalent to 'right'.

The notions of entailment and equivalence, as I will use them here, are *modal*. For any predicates A and B, *A entails B* just in case for every possible world w, the set of objects B is true of at w includes the set A is true of at w; A and B are *equivalent* just in case these sets are the same at every possible world. Similar definitions of entailment and equivalence are available for properties: for any properties α and β, *α entails β* just in case for every possible world w, the set of objects that instantiate β at w includes the set that instantiate α at w; α and β are *equivalent* just in case these sets are the same at every possible world.

Extending the Jackson argument to show the equivalence of normative and descriptive properties, and not just predicates, requires a further assumption. For it might be that certain predicates fail to express a property (and perhaps the big disjunctive descriptive predicate is a candidate for such a predicate). So Jackson needs the following, which we can call *predicate-property correspondence*, or PPC:

PPC For every predicate P, P expresses a property α_P where for any world w, the set of objects P is true of at w is the same as the set of objects that instantiate α_P at w.[117]

4.2.2 Excursus: accidental expression

There are complications at this point, which Jackson and his followers recognize. Our §1 gloss on DESCRIPTIVISM has it that, if P is a descriptive predicate, then the

[116] Jackson (2001, 655)

[117] This principle needs to be restricted to avoid paradox; I will assume that an appropriately restricted principle will license all of the uses to which PPC is put to here and below.

property α_P that P expresses (by PPC) is a descriptive property. But it won't do to say that *any* predicate containing descriptive language only counts as expressing a descriptive property. For instance, consider the predicate

D is the property I am actually thinking about now.

D is a descriptive predicate, as it contains no normative language. And by PPC, it expresses a property which, given what we have said, is a descriptive property. But this is problematic for the purposes of Jackson's argument. For suppose the property I am actually thinking about now is rightness; we then have an implausibly easy proof that rightness is a descriptive property.[118] Something must be said to explain why this isn't sufficient to refute non-naturalism.

Jackson (1998, 119, fn. 10) and Streumer (2008, 538-9) restrict the notion of a descriptive property to those properties that are expressed by descriptive predicates that do not contain *property-denoting* terms of the form 'the property such that Φ'. That is, since **D** contains the expression 'the property I am actually thinking about now', the property it expresses fails to thereby count as descriptive. (Though it could, in principle, still be a descriptive property—so long as there is *another* descriptive predicate which expresses it without recourse to a property-denoting expression.)

This move is not only highly artificial; it is inadequate. For the *same* problems that arise for **D** also arise if we consider the predicate

D* is actually being thought about by me now.

D* contains only descriptive language, yet in the appropriate circumstances, it expresses rightness. Intuitively, moreover, the explanations for why **D** and **D*** fail to express descriptive properties of the kind needed for Jackson's argument are the same. But **D*** contains no property-denoting expression. Another route is needed.[119]

The natural thought to have here is that both **D** and **D*** do not count for Jackson's purposes because they contain indexical expressions like 'I' and 'actually'. We can then add a general requirement that descriptive properties be expressed *non-accidentally* by

[118] More generally, I could, in principle, be thinking of *any* property, and **D** would thereby pick out that property. But not every property is a descriptive property.

[119] The same points apply to the example in Streumer (2008, 538), 'is the property Fred is thinking about at *t*'. Even though this expression eliminates the indexical expressions 'I' and 'now', it still expresses different properties at different *worlds*. So we should not want the property it expresses at the actual world to thereby count as a descriptive property. But, as before, the property-denoting expression is dispensable: the same points apply to the predicate 'is being thought about by Fred at *t*'.

descriptive language. More precisely, let an expression e non-accidentally expresses the property P just in case, holding fixed the meaning of e, for every world considered as a context of utterance w_c, e in w_c expresses P. Thus on this approach, rightness may be expressed by $\mathbf{D}^*$ (or $\mathbf{D}$), but it isn't *non-accidentally* expressed, since there are worlds (considered as contexts of utterance) where the predicates, on their actual meanings, fail to express rightness.[120]

Even with this amendment in place, the property that Jackson's big disjunctive predicate expresses still counts as a descriptive property. Since the rightness-entailing predicates from which the predicate is constructed contain no indexical-like terms, which property the big disjunctive predicate expresses is not accidental in the relevant sense. So at this point the argument can proceed as before, with our amendment in place.

4.2.3 Identity and Descriptivism

Jackson's argument, however, is for Descriptivism, which requires more than that rightness is *equivalent* to a descriptive property—rightness, according to the desired conclusion, must *be* a descriptive property. The road from equivalence to identity is easy for Jackson, who accepts the following thesis about the individuation of properties, which we can call the *equivalence thesis*, or ET:

ET For any properties α and β, if α and β are equivalent—i.e., if they share an extension at every possible world—then $\alpha = \beta$.

With these premises in place, Jackson's argument can be summarized as follows: GS guarantees the existence of descriptive a predicate equivalent to 'right'; PPC tells us that this predicate expresses a (descriptive) property, and ET implies that this property is the *same* property as rightness. Since we can repeat the same style of argument for any

[120] It might be desirable to complicate the definition of non-accidental expression to require that P is accidentally expressed by e only if e on its actual meaning expresses different properties in different worlds *in virtue of the linguistic rules governing e*. The reason for this is that some theoretical terms e_t might express different properties in different worlds (while holding meaning fixed) on account of the fact that e_t expresses the property which best fits the theoretical role delineated by e_t. If different properties fit this role best in different worlds, e_t will fail, according to our first-pass definition, to express a property non-accidentally. But e_t could be a term from a well-confirmed empirical science, and would in this case be a paradigmatic case of a term that expresses a descriptive property. The proposed revision, on which it is a necessary condition on accidental expression that a predicate expresses different properties in different worlds in virtue of the linguistic rules governing the expression, remedies this difficulty. e_t intuitively expresses different properties in different contexts in virtue of a fact about the modal profiles of the various properties that e_t expresses in different worlds, and not in virtue of the linguistic rules governing e_t.

normative property, these premises imply DESCRIPTIVISM. Given IMPLICATION, non-naturalism is false as well.

4.3 DESCRIPTIVISM without supervenience

In this section and the next I will argue that there is something wrong with this argument against non-naturalism. The reason is that even if we assume that GS is *false*, Jackson's auxiliary premises establish the *same* conclusion—in particular, we can still construct descriptive rightness-entailing predicates under the supposition that GS is false. The rest of the argument then proceeds exactly as before, yielding the conclusion that non-naturalism is false. But non-naturalism is clearly *true* under such a supposition. Something has gone wrong in Jackson's argument, although nothing in these sections will single out any specific premise for rejection. PPC, ET, and IMPLICATION are all suspects.

4.3.1 Metaphysical consequences of failures of supervenience

Our first step is to argue that if GS is is false, then non-naturalism *must* be true.[121] If GS is false, then the following holds:

NO GS $\exists w, w^* : w$ and w^* are exactly alike descriptively, and there is some normative respect in which w differs from w^*.

NO GS requires, in other words, that the following obtain: there are two descriptively alike worlds which differ over whether some action is right—i.e., there is some world where an action is right, and a second world which is descriptively identical to the first, but where that same action is not right.

I am claiming that non-naturalism follows from NO GS. Or, to be precise, *if* a broadly realist theory of the normative is true, non-naturalism is the only viable candidate under the NO GS supposition.[122] This is because non-naturalism is supposed to be a view

[121] It is important, in what follows, to distinguish the attitude of supposition from other attitudes one might bear to the failure of GS. In particular, supposing GS fails does not require *believing* that it fails or even having some minimal *credence* that it does; one can adopt such a supposition while being absolutely certain that it does not fail. With this distinction in mind, we can separate the present issues, which concern whether DESCRIPTIVISM is consistent with non-naturalism, from other issues concerning non-naturalism and supervenience. In particular, Mackie (1977, Ch. 1) and Blackburn (1988b) have argued that the non-naturalistic realist will have trouble *explaining* the supervenience of the normative on the natural. I will set this (important) issue aside here; the present point is independent of them, as it asks us to suppose that supervenience fails then to see what follows.

[122] Some extreme versions of irrealism might also be consistent with NO GS. I do not wish to deny this point; rather, my present purpose is to evaluate Jackson's argument in an intramural dispute between

according to which the normative is, in some important sense, *independent* of the descriptive. We haven't explained what the appropriate sense of 'independent' is, but it is clear that whatever the appropriate sense is, the normative is independent of the descriptive in the relevant way if it fails to even *supervene* on the descriptive. One way to illustrate this is by drawing attention to similar debates in other domains. Take dualism about the mental, for instance: we might say that it is likewise an independence thesis, holding that certain aspects of the mental (perhaps *qualia*, the qualitative aspects of experience) are independent of the physical, biological, chemical etc. Chalmers (1996) defends this kind of thesis, in a context where broadly realist theories of the mental are assumed to be the only live options, by arguing for the falsity of the following global supervenience thesis:

GS-MENTAL $\forall w, w^*$: if w and w^* are exactly alike in all physical, biological, chemical, etc. respects, then they are exactly alike mentally.

Hence much of the debate in this area centers around the metaphysical possibility of so-called "philosophical zombies": *if* they are in fact metaphysically possible, then GS-MENTAL fails and dualism is vindicated. Quite plausibly, the same is true for GS and non-naturalism about the normative: *if* the supervenience thesis fails, non-naturalism is vindicated.[123]

4.3.2 Descriptive rightness-entailing predicates under NO GS

Our next step is to show that Jackson's auxiliary premises imply that non-naturalism is false *even under the assumption that* NO GS *is true*. But before explaining why this is so, a point of clarification is in order: Jackson is correct to say that his preferred method for constructing descriptive rightness-entailing predicates requires GS.[124] What this ne-

realists, and so will restrict attention to broadly realist theories of the normative, both naturalist and non-naturalist. It is important to be clear, however, that I am assuming that the supposition of NO GS doesn't *ipso facto* require a rejection of realism, broadly construed. This seems like a reasonable assumption: as I describe below, in other domains, analogues of non-naturalistic realism follow from the falsity of relevant supervenience theses, and the normative does not seem to be anomalous in this respect.

[123] The terminology is tricky here, but this shouldn't obscure the underlying issues: Chalmers claims that his view, which denies GS-MENTAL, is nevertheless a version of *naturalism* (Chalmers (1996, xiii)). Thus it would be misleading to say that his view is an instance of non-naturalism about the mental. But this is because Chalmers has a distinctive theory about what 'naturalism' is, which may not capture the sense of 'naturalism' at issue in metaethics. I am assuming that there is still a clear sense in which the Chalmersian about the mental and the Moorean about the normative maintain that these domains are independent of others, regardless of the terminology we use to mark the similarity.

[124] Suppose that NO GS is true, and take the worlds w_1 and w_2 that differ only in normative respects. There is then some action which is the *same* with respect to its intrinsic descriptive features and envi-

glects is that there are other ways to construct the rightness-entailing predicates needed by Jackson's argument, and these don't require require the supervenience assumption.

In particular, we can construct rightness-entailing predicates by using identity and reference worlds. Letting $i_1, i_2 \ldots$ designate all possible instances of rightness, and $w_1, w_2 \ldots$ designate possible worlds in which there is at least one instance of rightness, there are rightness-entailing predicates of the following form:

I $x = i_n$ and x is in w_j[125]

where i_n is an instance of rightness in w_j.

Each predicate of the same form as **I** is rightness-entailing, as no possible action satisfies the predicate yet fails to be right. This is so even if NO GS is true. If NO GS is true, then there is a pair of worlds—let them be w_1 and w_2—which differ only in whether some action i is right. Suppose that i is right in w_1; then, the predicate

$\mathbf{I}_1$ $x = i$ and x is in w_1

fails to be rightness-entailing only if there is some action that is not right, yet satisfies $\mathbf{I}_1$. But i in w_2 is *not* such an action—while i in w_2 is not right, it *also* fails to satisfy the predicate $\mathbf{I}_1$, as the non-right action is in w_2, which is distinct from w_1. Thus predicates in the form of **I** can be rightness-entailing even if there are pairs of worlds that differ only over whether a particular action is right.[126]

4.3.3 Excursus: refining supervenience

Let us suppose, for the moment, that rightness-entailing predicates in the form of **I** are also *descriptive* vocabulary. (I defend this assumption in §4.) There is an apparent tension in how we have described the situation: on the one hand, rightness-entailing predicates in the form of **I** naming individuals and worlds are descriptive predicates. But, on the other hand, these predicates are supposed be able to pick out right actions *even if* supervenience fails, which is to say *even if* the only difference between the

ronment, but which is right in w_1 but not w_2. So descriptive characterizations constructed according to Jackson's method will not be rightness-entailing if NO GS is true.

[125] Here I conflate predicates with open sentences. The open sentence can easily be converted into a predicate by the device of lambda-abstraction which is familiar from formal semantics. Where $\ulcorner F(x) \urcorner$ is an open sentence with the free (unbound) variable x, binding the free x with the lambda-operator, yielding $\ulcorner \lambda x.F(x) \urcorner$, denotes a function from objects to truth-values (namely a function which assigns 'true' to an object in case that object is F, and assigns 'false' otherwise). The lambda-abstracted expression therefore has the same semantic function as a predicate.

[126] I owe this point to discussion with Mark Schroeder and Campbell Brown.

right action and a not-right action is a normative difference. We can't have it both ways; if predicates in the form of **I** are descriptive, then there is never only a normative difference between two actions. This might make it tempting to view our construction of alternative rightness-entailing predicates not as problematic for Jackson's argument, but rather as an indictment of our attempt to suppose that GS is false. After all, our *reductio* of Jackson's argument, which draws out the consequences of his assumptions under the supposition of No GS, isn't very promising if No GS is an incoherent supposition.

What this really shows is simply that more precision is needed in specifying the descriptive supervenience base for the normative. An appropriate revision will leave Jackson's argument unaffected, and will leave our construction of alternative rightness-entailing predicates in tact. And it will have the added advantage of making the failure of supervenience minimally intelligible.

Here is some terminology: there are some non-normative disciplines which intuitively are sufficient for describing a global supervenience base for the normative: these include, at the very least, microphysics. Perhaps psychology is needed (if the mental doesn't supervene on the normative), and theology is needed if there might be non-physical beings.[127] We can call the vocabulary from these disciplines *MPT vocabulary*, and can use it to formulate a minimal global supervenience base for normative properties. This gives us a revised version of global supervenience, namely:

GS$^-$ $\forall w, w^*$: if w and w^* are exactly alike in all MPT respects, then they are exactly alike normatively.

Even if our alternative rightness-entailing predicates are descriptive, they are not among the MPT vocabulary, as they contain world-names that are not replaceable by long descriptions in microphysical, psychological, and theological terms. So it makes sense to suppose that they are descriptive rightness-entailing predicates, yet the following denial of GS$^-$ also obtains:

No GS$^-$ $\exists w, w^*$: w and w^* are exactly alike in all MPT-respects, and there is some normative respect in which w differs from w^*.

Moreover, it is still extremely intuitive for the same reasons as before that non-naturalism is true under the supposition that GS$^-$ fails—that is, if No GS$^-$ is true,

[127] If one thinks that psychology necessarily reduces to microphysics, or that the supernatural entities in theology are impossible, then one can omit these elements from our discussion in what follows, and focus on microphysics. If one thinks that higher-order scientific properties such as those studied by biology are not reducible to microphysics, then one will need to supplement the list. For doubts, see Sturgeon (2009).

then so is non-naturalism. But since, supposing No GS$^-$ is true, we can still use a disjunction of predicates in the form of **I** to express rightness, Descriptivism is true. Hence, even with a move to characterizing supervenience in terms of GS$^-$ rather than GS, our *reductio* is still in place.

This naturally raises the question of whether there should be a corresponding change in the Descriptivism as a characterization of non-naturalism. Implication as stated claims that *any* descriptive means for picking out normative properties secures the falsity of non-naturalism. But we might imagine a revision of Descriptivism which mimics our revision to GS, as follows:

Descriptivism$^-$ Every normative property is identical to a MPT property,

where MPT properties are those that can be expressed using MPT vocabulary. A corresponding revision to Implication then would hold that only Descriptivism$^-$ implies the falsity of non-naturalism. Our new rightness-entailing predicates in the form of **I** are, on this revision, a non-starter for a *reductio* of Jackson's method of argument, since they employ non-MPT vocabulary.

While a move to Descriptivism$^-$ would avoid the *reductio*, it is in a way unsatisfactory. I will return to this issue in more detail in §5, but here is a basic summary of what is at issue. If this move is to rescue Jackson's argument, it rests on the following assumption: that non-naturalism consistent with Descriptivism, but is not consistent with Descriptivism$^-$. This would be a very peculiar feature of non-naturalism, and is one that should be explained. For it would raise the question of how the difference between the MPT-vocabulary and descriptive vocabulary is relevant to non-naturalism, a metaphysical thesis about the normative. According to the Descriptivism$^-$ proposal, it is evidently a quite important difference: non-naturalists can allow that normative properties are picked out by descriptive vocabulary, but cannot allow that these properties are picked out by the subset of descriptive vocabulary that is MPT-vocabulary. Whether we can explain this difference in a way that is a consistent with Jackson's argument is very much an open question.

4.3.4 Summary

Here is where we are so far: under the supposition of No GS$^-$, non-naturalism is clearly true. But if the rightness-entailing predicates we have constructed under the supposition of No GS$^-$ are also descriptive predicates, we can then proceed as Jackson does before: a big disjunction of such predicates yields a descriptive predicate that is equivalent to 'right'. PPC implies that there is a descriptive *property* equivalent to rightness, and ET guarantees that these properties are identical. That is the thesis of DESCRIPTIVISM. When conjoined with IMPLICATION, the falsity of non-naturalism immediately follows. This is an implausible result, since we derived it even while assuming that NO GS$^-$ is true. Before proceeding, three items of clarification are in order.

First: one might think that the force of the above *reductio* lies in that it shows the conjunction of NO GS$^-$, PPC, ET and IMPLICATION to be be inconsistent: it implies that both non-naturalism and its denial are true. But this by itself would be a weak argument. It would indistinguishable from the argument that the conjunction of a logical contradiction with PPC, ET and IMPLICATION implies both non-naturalism and its denial. This latter argument doesn't show anything interesting about Jackson's auxiliary premises; both non-naturalism and its denial follow from the logical contradiction alone. And the former argument doesn't seem much better, since it relies on the patently false NO GS$^-$.

Second: the *reductio* I am offering should instead be understood as consisting in two separate claims. The first is that, on the supposition that NO GS$^-$ is true, non-naturalism is also true. This is a claim about NO GS$^-$ alone; not about its conjunction with other premises. The second is that the falsity of non-naturalism can be derived from Jackson's auxiliary assumptions by themselves. In other words, the assumption of GS$^-$ plays no essential role in the argument. One way to show this is by explicitly supposing its denial—i.e., NO GS$^-$—and then showing that the derivation of the falsity of non-naturalism from these assumptions *still* goes through. The auxiliary assumptions are therefore too strong in this sense: they, by themselves, would license a derivation of the falsity of non-naturalism, and would do so even under the assumption that GS$^-$ is false. Putting these two claims together yields a *reductio*: Jackson's auxiliary assumptions all by themselves license the conclusion that non-naturalism is false, and therefore license the conclusion that non-naturalism is false even under suppositions where it is clearly true. An argument with this structure prevents us from placing blame on the supposition of NO GS$^-$.[128]

Third: something more should be said about the role played by the attitude of *supposition*. The supposition that NO GS$^-$ is true plays a central role in the argument, and it is important to be clear about what this does and does not mean. In particular, it does not mean that we are claiming that NO GS$^-$ is, or even might, be true. Supposition is a flexible attitude; one can even suppose the denial of known logical truths in order

[128] Thanks to Brian Weatherson for helpful discussion on this point.

to carry out a proof by *reductio*. We are taking advantage of this flexibility in the present argument, as non-naturalists typically accept GS$^-$, and some even go so far so as to grant it the status of a conceptual truth. Everything we have said here is consistent with these claims. Supposing that it is false merely helps to bring out the interesting properties of Jackson's auxiliary assumptions.[129]

4.4 World-names and Ramsification

The big assumption in all of this is that the rightness-entailing predicates in the form of **I**, when constructed under the No GS$^-$ supposition, are genuinely descriptive. This assumption can be challenged: we might worry that, especially in the use of world-names under the assumption of No GS$^-$, we are covertly smuggling in normative vocabulary. My contention in this section that these terms can be shown to be (or to be replaceable by) terms that clearly do not contain normative vocabulary. This will complete the *reductio* argument that Jackson's auxiliary assumptions are too strong, as they even imply the falsity of non-naturalism under the supposition of No GS$^-$.[130]

We can make the worry that use of world-names constitutes use of normative terms explicit, as follows. Given No GS$^-$, there are worlds w_1 and w_2 which differ in normative but not MPT respects—for simplicity, we will suppose that these worlds differ only in that some act i is right in w_1 but not in w_2, and are otherwise exactly the same. The name for w_1 must then describe it as a world where i is right. Otherwise, it could not succeed in referring to w_1 and not w_2—a description of w_1 that omits the claim that

[129] Note that, given the previous paragraphs, the analogy with proofs by *reductio* isn't borne out in every respect. In particular, a *reductio* proof establishes the negation of its supposition. But the facts that (i) the auxiliary premises are themselves susceptible to questioning, and (ii) a contradiction doesn't follow directly from No GS$^-$ also, together suggest that blame should be placed on the auxiliary assumptions.

[130] Could the instance-names for right actions contain normative vocabulary? I find this suggestion less worrying, but here is one way to show that they do not.

Start by picking out actions via the region of spacetime they occupy. For instance, if regions of spacetime are identified by sets of 4-tuples such as '$\{\langle x_i, y_j, z_k, t_l\rangle, \langle x_m, y_n, z_o, t_p\rangle \ldots\}$' (which are surely descriptive), we could identify right actions by descriptions of the form

x is at $\{\langle x_i, y_j, z_k, t_l\rangle, \ldots\}$ and x is in w.

Complications arise for this simple proposal if there can be multiple objects or actions occupying the same region of spacetime. Here two components of descriptive vocabulary that, plausibly, will serve to distinguish such objects or actions. First, coincident actions or objects might satisfy different sortals: a right action, for instance, might fall under the description 'intentional human action' whereas an event that happens to coincide with the action does not. Second, there are descriptive *modal properties*: properties about what some thing might, or must be. (A coincident statue and lump of clay do not share the property of *possibly surviving being squashed*. One could reasonably hope that sortals and predicates expressing modal properties of this kind will individuate instances of rightness sufficiently.

i is right is a description that w_2 also satisfies. But then the name for w_1 is a piece of normative vocabulary, as it refers to w_1 and not w_2 by describing i as right. Our big disjunction of rightness-entailing predicates cannot be assumed to be descriptive, and so a crucial step in Jackson's argument is missing under the NO GS$^-$ supposition.

Where does this line of reasoning go wrong? Certainly *one* way of distinguishing w_1 is by referring to as the world where i is right. But this doesn't show that we *must* use normative vocabulary to pick out w_1, rather than w_2. Here is an alternative method.

Recall the complete list of microphysical, psychological and theological terms, or what we have been calling the *MPT vocabulary*.[131] Worlds as a whole can be described in MPT terms, and some of these descriptions are *complete*—i.e., every true claim about a world w with MPT vocabulary is made by a complete MPT description of w. Let D_w be such a description of w. We can then form a description which not only completely describes w in MPT terms, but also *says that* it is a complete MPT description of the world:

> x is D_w and x has no other MPT properties.

Call such a description MPT-*closed*. If MPT descriptions like D_w are to give a genuine supervenience base for the normative, the MPT vocabulary must include the MPT-closure.[132]

With NO GS$^-$ in place, worlds that differ normatively will satisfy the same complete MPT-closed descriptions. For instance, the worlds w_1 and w_2, which differ only in whether some action i is right or not, satisfy the same MPT-closed description. w_1 and w_2, then, cannot be distinguished in MPT terms.

But we can still distinguish between them using non-MPT descriptive vocabulary. Begin with a "functional" characterization of rightness in other normative terms—i.e., a characterization of the *intra-normative* connections between rightness, having a reason, etc.[133] I don't want to take on any specific claims about what exactly the proper intra-normative connections of rightness are, but it is highly plausible that the something like following will be correct (those who disagree can substitute their favored theory in what follows):

[131] Recall also that these are placeholders for whatever vocabulary is intuitively needed to formulate a minimal global supervenience base for normative properties.

[132] What I am calling an MPT-closure is analogous to the notion of a physical description of a world with a "stop clause" found in Jackson (1998, 13). While not all negative existentials involving MPT vocabulary themselves constitute MPT vocabulary, I will assume that the closure clause is a special case.

[133] See Ewing (1947, 148-9), Gibbard (1990, 51), and Scanlon (1998, 97).

R1 Necessarily, $\forall x$, if x is right, then there is some reason to do x;

R2 Necessarily, $\forall x$, if x is right, and one can do x, then one has overall reason to feel guilt if one doesn't do x;

R3 Necessarily, $\forall x$, if x is right, then one has overall reason to blame someone who can do x, but does not.

Let a *Ramsification* of the intra-normative connections of rightness be the result of, first, replacing all of the normative terms in R1-R3 with distinct variables.[134] This leaves us with the descriptive open sentences R1*-R3*:

R1* Necessarily, $\forall x$, if x has F, then doing x has G;

R2* Necessarily, $\forall x$, if x has F, and one can do x, then one bears H to feeling guilt if one doesn't do x;

R3* Necessarily, $\forall x$, if x has F, then one bears H to blaming someone who can do x, but does not.[135]

Since we are interested in forming a term that designates rightness, we can then conjoin the open sentences R1*-R3* and bind the variables that replaced normative terms besides 'right' in with existential quantifiers to obtain its Ramsification. This is the predicate **R**:

R $\exists G, \exists H$: necessarily, $\forall x$: if x has F, then doing x has G; if x has F, and one can do x, then one bears H to feeling guilt if one doesn't do x; $\forall x$, if x has F, then one bears H to blaming someone who can do x, but does not.

We know that rightness is one of the properties that satisfies **R**, the bound Ramsified intra-normative connections of rightness. Moreover, **R** uses no normative vocabulary.[136]

But readers familiar with using Ramsified theories to define theoretical terms will immediately wonder whether **R** is sufficient for our purposes: Ramsifications can suffer from the problem of being satisfied by *too many* properties. And if **R** is satisfied not only by rightness, but by other properties besides, it will not do as a descriptive specification of rightness.

[134] See Lewis (1970).

[135] I have sightly modified the syntax of R1-R3 to accommodate F, G and H as first-order variables which take properties as assignments. This is for ease of expression; nothing essential hangs on the difference as we could easily have used second-order variables to accomplish the same task.

[136] Jackson agrees: see Jackson (1998, 141).

The solution to this worry is to not use **R** to directly specify rightness—rather, it is used as a *component* in a larger descriptive predicate of the kind sketched in §3. Such a predicate requires wholly descriptive names for worlds; these cannot simply be MPT-closed descriptions. (Under the assumption of No GS^{-}, purely MPT-closed descriptions fail to uniquely characterize worlds.) **R** merely supplements MPT-closed descriptions of worlds: those worlds that are identical in all MPT respects can be such that they differ in whether certain things have a further non-MPT property that satisfies **R**. To take an example: even though w_1 and w_2 satisfy the same MPT-closed description—$D_{1\&2}$, say—they are different in the following respect. In w_1, i has a *further* property, not mentioned in the MPT-closed description $D_{1\&2}$, which satisfies **R**. In w_2, i *lacks* such a further property. That is:

w_1 is such that

- it is $D_{1\&2}$ and there are no other MPT properties that it instantiates;
- there is a further property Z such that Z satisfies **R** and i has Z in w_1.

w_2 is such that

- it is $D_{1\&2}$ and there are no other MPT properties that it instantiates;
- there is no further property Z such that Z satisfies **R** and i has Z in w_2.

By replacing world-names with descriptions of this kind, we then have a rightness-entailing descriptive predicate in $\mathbf{I}_N$:

$\mathbf{I}_N$ $x = i$ and x is in a world w such that (i) w is $D_{1\&2}$ and there are no other MPT properties that it instantiates and (ii) there is a further property Z such that Z satisfies **R** and x has Z in w.

How does this address the worry that there are too many properties that satisfy **R**? Notice that the existence of MPT properties that satisfy **R** is irrelevant here—even if such properties exist, they don't satisfy clause (ii) of $\mathbf{I}_N$, which requires a further non-MPT property that satisfies **R**. So the success of **R** does not depend on its being uniquely satisfied.

This still leaves open the possibility that another non-MPT property (that is, another normative property) besides rightness also satisfies **R**. I don't have a further argument that rules out this possibility. Instead, I will simply point out that a proper characterization of the intra-normative role for rightness will no doubt require more detail than what is found in R1-R3, and once further spelled out, its Ramsification will plausibly not be satisfied by other normative properties like *being a reason*, goodness, etc. This is a substantive claim that would ideally receive further defense. But it requires settling the first-order normative question about the nature of rightness, which is a project I will not attempt here.[137]

To summarize: the big disjunctive predicate of §3 will quite plausibly require no normative vocabulary. Such a predicate expresses rightness non-accidentally, so the property it expresses genuinely counts as the kind of property IMPLICATION says is inconsistent with non-naturalism. (It includes property-denoting expressions, but as we argued in §2, we have not been given a good motivation for claiming these expressions do not necessarily express properties that are 'descriptive' in the relevant sense.) We can then conclude that, even under the supposition of NO GS$^-$, there is a descriptive predicate equivalent to 'right'. The rest of Jackson's argument can proceed as before, concluding with the falsity of non-naturalism. This amounts to a *reductio* of Jackson's original premises, since non-naturalism is clearly true under the supposition of NO GS$^-$.

4.5 Non-naturalism as a fundamentality thesis

It is natural at this point to ask which of the auxiliary premises is at fault. Criticisms have been raised in the literature in response to Jackson's original argument; these almost exclusively focus on the premise ET that features in the derivation of DESCRIPTIVISM.[138] Instead of debating the merits on such responses, I will explain and develop a different option for rejecting Jackson's auxiliary assumptions. This option relies on a rejection of IMPLICATION, the premise that embodies Jackson's conception of non-

[137] We should note in this context that Smith (1994, 48-56) offers a positive reason for thinking that Ramsified descriptions of normative properties will be satisfied by multiple normative properties. The Ramsified descriptions of the color property redness will be satisfied by other color properties, and Smith suggests that this is a general problem which affects all Ramsified descriptions, including those of normative properties. I don't think Smith is correct to draw any general pessimistic conclusion from his observation about color properties—there is an important disanalogy between the Ramsified description for rightness, R1-R3, and Ramsified descriptions of (say) redness. As Smith points out, the characterization of redness and yellowness are *symmetric*, in that they are exactly the same, save for substitution of appropriate color terms. Hence, when these characterizations are replaced with bound variables, the resulting Ramsified descriptions of redness and yellowness are exactly the same. But rightness and reasons aren't the like this: their intra-normative connections aren't symmetric, and the property *being a reason* doesn't satisfy the Ramsified description of rightness. One might reasonably conjecture that the intra-normative connections for rightness can be filled out in sufficient detail so as to generate the same asymmetry with every other normative property.

[138] ET has been subjected to criticism by Shafer-Landau (2003) and Suikkanen (2010), and Schmitt and Schroeder (2011) identify this assumption as a weakness of the argument. PPC could also be a culprit, as one could easily imagine applying a "sparse" theory of properties in the style of Armstrong (1978) which renders PPC false.

naturalism as a thesis about the "inadequacy" of descriptive language. As we saw in §§3-4, descriptive language turns out to be very powerful—too powerful, in fact, for non-naturalism to be plausibly construed as an inadequacy thesis. Moore himself suggested a stronger conception of non-naturalism, one tied to the impossibility of descriptive *definition* of normative properties. The response I develop fits in this Moorean mold; it requires a rejection of IMPLICATION in favor of a stronger conception of non-naturalism.

The essentials of the approach are sketched in §5.1, where I show how the resource of *metaphysical fundamentality*, together with our earlier *reductio*, motivates a rejection of IMPLICATION. In particular, it allows us to explain why non-naturalism is true under the supposition of NO GS^-, even if we explicitly endorse the other premises in Jackson's argument. Rejecting IMPLICTATION in this way also predicts a failure of Jackson's argument when GS^- is back in place. The upshot, then, is that the problem with Jackson's argument is very naturally diagnosed as deriving from IMPLICATION only. The assumptions of PPC and ET are, in this context, just a diagnostic tool; we assume them to illustrate that the IMPLICATION-rejecting response does not require rejecting other assumptions in Jackson's argument.

§5.2 illustrates how the fundamentality-based conception of non-naturalism doesn't require acceptance of ET. One can, in fact, motivate a rejection of ET in this fundamentality-centric setting. But this doesn't show that rejecting IMPLICATION is optional. The opposite is true: even if the fundamentality-based response is taken to motivated a rejection of ET, IMPLICATION cannot be left untouched. Thus the lesson of §§5.1-5.2 will be that non-naturalism construed as a fundamentality thesis requires a rejection of IMPLICATION, regardless of its relation to ET.

Finally, §5.3 returns some issues left unresolved by earlier discussion. Metaphysical fundamentality explains why certain descriptive expressions, including the indexical predicates of §2, are such that it is consistent with non-naturalism that they express normative properties. The upshot is that we can explain a point of agreement that has so far gone unexplained. Jackson was right to restrict his attention to certain kinds of descriptive language, but appreciating the reason for this restriction has unwelcome consequences for his own argument.

4.5.1 Ramsification and fundamentality

Begin with the notion of metaphysical fundamentality as it is explained under the heading of "Reality" in Fine (2001, 2009), "grounding" in Schaffer (2009), and "structure" in Sider (2012). Collections of microphysical particles are (for instance) more fundamental than tables, chairs, rocks or trees. This is to say that there is a kind of explanatory dependence between the microphysical and the tables, chairs, etc.: the microphysical gives the most metaphysically satisfying explanation of the facts about these macrophysical objects. Take some microphysical predicate M which, if it is true of some things, entails that there is a tree. The first of the following fundamentality-claims is then true while the second is false:

Particles that are M are more fundamental than trees.

Trees are more fundamental than particles that are M.

There is a corresponding distinction in fundamentality in language: predicates like M can be said to be fully fundamental terms expressions, as they are constructed out of terms like (we may assume) 'quark', 'spin', etc. There are other non-fundamental expressions for the same property: these include macrophysical predicates like 'table', 'chair', 'rock', and 'tree'. With these resources in hand, we can sketch an account of why the IMPLICATION assumption is responsible for the *reductio* in §§3-4.

That *reductio* required us to construct a descriptive predicate equivalent to 'right' by using, in part, a Ramsification of the intra-normative connections of rightness. A complete descriptive predicate equivalent to 'right' would include (among other things) the following predicate:

R $\exists G, \exists H$: necessarily, $\forall x$: if x has F, then doing x has G; if x has F, and one can do x, then one bears H to feeling guilt if one doesn't do x; $\forall x$, if x has F, then one bears H to blaming someone who can do x, but does not.

Rightness satisfies **R**. But the success of **R** in this respect *depends* on the normative in a way that has consequences for the fundamentality-facts in the area.

Let us call an expression *existential* just in case the expression uses an existential quantifier like 'some' or 'there is'. Using **R** in the descriptive specification of rightness involves using an existential expression: worlds with an action i is right are worlds where *there is* a further non-MPT property that satisfies **R** and is instantiated by i. It is quite plausible that existential expressions bear the following relationship to fundamentality: truths that can be stated using an existential expression can always be stated in more fundamental terms by using only non-existential expressions. Thus, there is someone in the next room in virtue of, or because of, Sally's being in the next room. Similarly, something has a property which satisfies the Ramsified functional role for rightness in virtue of it's being right. That is:

> That *i* is right is more fundamental than *i*'s having a non-MPT-property that satisfies **R**.

At the level of language, 'right' is a more fundamental expression than a predicate containing **R**.

IMPLICATION makes non-naturalism blind to this distinction in fundamentality. Even though the big disjunctive predicate of §§3-4 utilizes **R** to express rightness in purely descriptive terms is an existential, and therefore non-fundamental predicate that expresses rightness, IMPLICATION does not care. So long as there is *some* way to express the property using descriptive vocabulary, non-naturalism is false according to IMPLICATION. The fact that we must use a provably non-fundamental expression in order to express rightness doesn't enter into the picture, given Jackson's assumptions.

This suggests a diagnosis of the *reductio* which faults IMPLICATION. We should replace IMPLICATION, according to this approach, with a characterization of non-naturalism that is sensitive to the distinction between fundamental and non-fundamental expressions. A simple form of such a characterization holds that non-naturalism is the view that FUNDAMENTAL DESCRIPTIVISM is false:

> FUNDAMENTAL DESCRIPTIVISM Every normative property is identical to a property that is **most fundamentally** a descriptive property only,

where a property is most fundamentally descriptive only just in case the only way of expressing it in fundamental terms uses purely descriptive vocabulary. FUNDAMENTAL DESCRIPTIVISM is then one way of pursuing a diagnosis of Jackson's argument as ignoring distinctions in fundamentality. And it requires a rejection of IMPLICATION *only*, so long as we make some additional assumptions about the metaphysics of fundamentality. I describe one such set of assumptions below.

If the non-naturalist rejects FUNDAMENTAL DESCRIPTIVISM, she accepts that the normative is highly fundamental, and hence accepts a claim in the form of **F**, where *D* is a descriptive predicate necessarily co-extensive with 'right':

F That *i* is right is more fundamental than that *i* is *D*.

But what she can't do, without also rejecting ET, is accept that **F** is true because of some fact about the fundamentality-relations between distinct properties of rightness and *D*-ness. That is, she cannot accept that, if **F** is true, then this is because **F-prop** is also true:

F-prop The property rightness is more fundamental than the property *D*-ness.

An approach to non-naturalism as the denial of FUNDAMENTAL DESCRIPTIVISM then needs another way of explaining fundamentality-claims like **F**, if it is not to be committed to a rejection of ET, in addition to IMPLICATION.

Sider (2012) provides a framework for fundamentality-claims on which **F-prop** might be false, even if **F** is true. There is, then, at least approach that allows rejecting IMPLICATION but not ET. (There may well be others; I will stick with Sider's framework for the purposes of exposition.) Sider accepts a thesis he calls PURITY, a about which truths count as fundamental. In particular:

PURITY Fundamental truths involve only fundamental notions.

Given PURITY, **F** cannot be a fundamental truth—it contains *D* which is, by hypothesis, a descriptive expression containing non-fundamental terms. Notice that the *same* is true for **F-prop**, which also contains non-fundamental notions, namely *D*. Hence it cannot be a fundamental truth, for the same reasons. So there is no temptation, in Sider's framework, to accept **F-prop** as an explanation of why **F** is true. It would itself require a further explanation, but no further fact about the fundamentality-relations between properties is forthcoming.

Sider opts for a different explanatory strategy, one that is at home with PURITY. We begin with the notion of a "metaphysical semantics": a metaphysical semantics gives, for any sentence containing non-fundamental terms *S*, the truth-conditions for *S* using only fully fundamental vocabulary. Thus, if M_T specifies the necessary and sufficient microphysical grounds of trees' having leaves, the metaphysical semantics for 'trees have leaves' claims the following:

$\mathbf{MS}_T$ 'trees have leaves' is true iff M_T.

$\mathbf{MS}_T$ explains, for Sider, what cannot be a fundamental fact given PURITY, namely **M**:

M M_T is more fundamental than that trees have leaves.[139]

The metaphysical semantics for 'trees have leaves' thus stands proxy for the fundamentality-fact **M**. But it does so without positing two sets of properties, namely the property of being a tree, and the microphysical properties. All it requires is the fundamental fact M_T, and two ways of talking about it: one using 'trees', and one using the fully fundamental vocabulary of microphysics.

So Sider provides us with a framework for talking about the relationship between the non-fundamental and the fundamental, without positing a domain of non-fundmental properties. There are only fundamental properties to talk about on this picture. It then provides an outline of a metaphysically realist approach to **F** that does not require **F-prop**, and hence does not imply a proliferation of properties that is inconsistent with ET.

Here is why: the less-than-fully fundamental vocabulary in D means that, by PURITY, **F** is (like **M**) not a fundamental fact. It needs to be explained, and the Siderian approach is to give a metaphysical semantics for the component containing non-fundamental terms, 'i is D'. This will be a claim in the form

$\mathbf{MS}_{...}$ 'i is D' is true iff ....

Given that 'right' is, on the non-naturalist view, a fully fundamental piece of vocabulary, then it is eligible to appear in the truth-conditions given by a metaphysical semantics. That is, the metaphysical semantics for 'i is D' can be filled in to read as follows:

$\mathbf{MS}_R$ 'i is D' is true iff i is right.

Thus the non-naturalist's claim **F** is explained in the Siderian framework by appeal to a metaphysical semantics that includes $\mathbf{MS}_R$. As before, ET can stay: $\mathbf{MS}_R$ tells us only that there are two ways of talking about the fundamental fact that i is right: one in less metaphysically perspicuous descriptive terms, and one in fully fundamental normative terms.[140]

[139] Of course the claim of the metaphysical semantics $\mathbf{MS}_T$ won't be fully fundamental itself, given PURITY: the linguistic expression 'trees have leaves' isn't fully fundamental, and hence the biconditional itself stands in need of further explanation. The advantage of a metaphysical semantics is that a further explanation is readily forthcoming: we can also give a metaphysical semantics for the sentence $\mathbf{MS}_T$: this gives the truth-conditions, in fully fundamental terms, for ' 'trees have leaves' is true iff M_T'. Obviously, this process repeats itself; the salient point here is that we never run out of explanatory material.

[140] The non-naturalist view does require some amendments to Sider's notion of a metaphysical semantics, but these are cosmetic changes that affect nothing of substance. First, readers will notice that with

There is then a coherent version of non-naturalism that denies FUNDAMENTAL DESCRIPTIVISM, but maintains ET. It denies FUNDAMENTAL DESCRIPTIVISM by allowing normative terms on the right hand side of a metaphysical semantics. Nothing said here implies that the basic idea presented here *requires* an assumption of ET. I will return to these issues in the next subsection, giving special attention to how things go under alternative assumptions which require the rejection of ET. But first, it will be important to notice in detail how this approach represents a resolution of our earlier *reductio*, and how it doesn't overgenerate in undesirable ways.

By characterizing non-naturalism as the denial of FUNDAMENTAL DESCRIPTIVISM, we thereby explain why non-naturalism is straightforwardly true under the supposition of NO GS^-. When we introduced the notion of fundamentality at the beginning of this section, we noted that "existential" means of expressing properties are often not the most fundamental means of expressing those properties. And, given that NO GS^- holds, the only descriptive means of expressing rightness will involve the existential expressions outlined in §§3-4. The non-existential means of expressing the property must then involve normative vocabulary. Hence, supposing NO GS^- is true, it follows that rightness is *most fundamentally* a normative property, and that FUNDAMENTAL DESCRIPTIVISM is false.

GS^- back in place, there is also a highly complex descriptive predicate, using only fully fundamental terms, that is eligible to provide a metaphysical semantics for D—call this $F(D)$. Thus, $\mathbf{MS}_R$ isn't the *only* admissible metaphysical semantics for 'i is D', given our current characterization of the notion, since there is a metaphysical semantics that also proceeds in terms of $F(D)$. The Siderian explanation for **F**, then, should be that $\mathbf{MS}_R$ is *among* the admissible metaphysical semantics for 'i is D'; not that it is the only one.

The existence of two admissible metaphysical semantics—one normative, and one descriptive—leads to a second point. In characterizing the notion of a metaphysical semantics, Sider says:

> The metaphysical semanticist seeks to explain *some* of the same phenomena as does the linguistic semanticist. For example, just like the linguistic semanticist, she wants to help explain why English speakers will point to the salient horse, rather than the salient car, when they hear the sounds "Point to the horse!" (Sider (2012, 113))

This is a kind of explanatory constraint on a metaphysical semantics: the metaphysical condition on the right side explains, in part, speakers' linguistic behavior with the sentence that occurs on the left side. But a non-naturalist might accept that the normative is fundamental, without thereby accepting that it has a causal profile of the kind that allows it to play a role in the relevant linguistic explanations. This is, more plausibly, a feature of a descriptive metaphysical semantics, which appeals to $F(D)$. There are many nuances to be explored in this setting, but the general lesson here is that we should not insist on explanation of linguistic behavior as a general feature of a metaphysical semantics.

This resolution can be shown to avoid overgeneration when we assume GS$^-$; with this assumption back in place, non-naturalism does *not* straightforwardly follow. This is as it should be: even if non-naturalism is true, there is no easy proof of it—or, at least, it can't be shown to be true with a straightforward proof like the one we gave above under the supposition of No GS$^-$. Suppose, then, that GS$^-$ holds. Unlike the above argument, we cannot rely on the *existential* character of the descriptive predicate expressing rightness to show that there must be a more fundamental characterization of the property—indeed, there is no reason at all to think that the descriptive means of expressing rightness will be existential in this way. Hence, we can't argue that there must be a more fundamental means of expressing the property, and that it must use normative vocabulary. So the falsity of FUNDAMENTAL DESCRIPTIVISM doesn't follow in the same way as before.

Of course, we can't show from Jackson's premises alone that FUNDAMENTAL DESCRIPTIVISM is *true*, either—merely from the fact that there is *some* descriptive characterization of rightness, nothing follows about whether the only fundamental characterization of the property is descriptive. Thus, Jackson's premises neither imply that FUNDAMENTAL DESCRIPTIVISM is true, nor do they imply that it is false. If non-naturalism is the view that FUNDAMENTAL DESCRIPTIVISM is false, the status of non-naturalism is left entirely up in the air by the other premises in Jackson's argument.[141]

The failure of Jackson's premises to establish FUNDAMENTAL DESCRIPTIVISM contains a larger lesson. What would settle the question of whether rightness is most fundamentally a descriptive property? The answer is not appeals to highly general linguistic and metaphysical assumptions like Jackson's ET and PPC. These don't give us any information about the fundamentality of rightness, and so they tell us nothing about whether FUNDAMENTAL DESCRIPTIVISM is true. What we need instead is to do substantive metaethical theorizing; we need, in other words, to ask what roles normative properties play, and to ask whether there are reasonably explanatory descriptive properties—properties from biology, psychology, and the like—that can play (most of) these roles. This kind of theory-building constitutes the strongest case for the kind of reductive thesis Jackson intends to establish; if successful, it shows that we can accommodate the data the non-naturalist accounts for, but without her metaphysical

[141] The difference can be put in Siderian terms: supposing No GS$^-$, normative vocabulary must be used in giving a metaphysical semantics; with GS$^-$ back in place, nothing directly follows concerning whether a metaphysical semantics requires normative vocabulary.

extravagance.[142] Absent this kind of detailed theorizing, we will not gain traction on the question of the fundamentality of the normative and, by extension, the truth of non-naturalism. This is where Jackson's argument goes wrong in the most general terms: it uses resources that are not precise enough to give us leverage on the question of non-naturalism.[143]

4.5.2 Giving up ET

The previous subsection sketches how non-naturalism, characterized as the denial of FUNDAMENTAL DESCRIPTIVISM, is consistent with ET. But this does not mean that we *must* accept ET in order to give the fundamentality-based, IMPLICATION-rejecting response to Jackson's argument developed there. The point of explicitly assuming ET was simply to show the fundamentality-based response most directly implicates IMPLICATION as the faulty premise in Jackson's argument, and is therefore distinct from the ET-rejecting responses in the literature. In this subsection, I will give a brief sketch how the fundamentality-based approach outlined above might be wed to an explicit rejection of ET. The upshot of this discussion, however, will be much the same as before: in faulting Jackson's argument for not making needed distinctions in fundamentality, we need to reject the IMPLICATION premise.

The crucial claim of §5.1 is that the fundamentality claim **F** is consistent with ET, but it required further assumptions—in particular, that **F** is not explained by the truth of a claim like **F-prop**, and is instead to be explained in terms of something like Sider's metaphysical semantics. As I have been emphasizing, this doesn't imply that one *must* take such a stance; it only indicates that doing so is not incoherent. One could instead accept **F-prop** as well, and thereby reject ET.

The main point to make here is that, while this is a viable approach, it isn't by itself plausible. Rather, a position which rejects ET for the reason described above *overgener-*

[142] For examples of this kind of theorizing, see Railton (1986) and Schroeder (2007). I confine the discussion here to the intramural debate between realists—there are obviously other alternatives to non-naturalism, including error theory (see Mackie (1977)) and expressivism (see Gibbard (2003)).

[143] This can be elaborated in the context of Moore's claims about "definition" which we mentioned in §1. While Moore's talk of parthood is potentially confusing, parts generally stand in an explanatory relationship with the things they compose—things have their (intrinsic) properties *in virtue of* the (intrinsic) properties of their parts. (This is not to say that the interaction between part and whole might not be highly complex—*cf.* Moore's discussion of "organic unities" in Moore (1903, 82 ff.).) By denying that Good has parts, Moore could be denying that the *kind* of explanatory relationship that part/whole explanations provide is unavailable for Good—there is, in other words, nothing more to be said to explain why Good has the features it does. The conception of reduction as a "constitutive account" in Wedgwood (2007, 145) might also be put to use in a similar way.

ates if it retains IMPLICATION. That is, it predicts that non-naturalism is obviously true, even under the assumption of GS^-, and it predicts that this is so for exactly the same reason that non-naturalism is true under the supposition of NO GS^-. IMPLICATION needs to be rejected as well.

Suppose NO GS^-, and that **F-prop** follows from **F**. Non-naturalism follows: normative properties are distinct from descriptive properties because, according to **F-prop**, they are more fundamental than descriptive properties.[144]

The problem is, essentially the same line of reasoning can be repeated to reach the conclusion that non-naturalism is false when we assume GS^-. There must be *some* difference in fundamentality between the normative and the descriptive if GS^- is in place; let's make the assumption that is most hostile to the non-naturalist in the present setting by supposing that the descriptive is more fundamental. That is, for some descriptive predicate *D*:

F* That *i* is *D* is more fundamental than that *i* is right.

Given **F***, there is on our present assumptions a corresponding truth about the fundamentality-relations between properties. That is **F-prop***:

F-prop* The property of *D*-ness is more fundamental than the property of rightness.

F-prop* is similar to **F-prop** in one respect: it says that there is a difference in fundamentality between normative properties and descriptive properties, and so implies that normative properties are distinct from the descriptive properties they are necessarily co-extensive with. DESCRIPTIVISM is then false, and non-naturalism follows as before.

But here we were assuming GS^-, and even made other assumptions that are *hostile* to non-naturalism; nonetheless, the same same straightforward proof on non-naturalism was available. Something has gone wrong. The present set of premises overgenerates, as they imply that non-naturalism is straightforwardly true given *both* GS^- and NO GS^-. But clearly there is an asymmetry between these assumptions; only one licenses a straightforward inference to the truth of non-naturalism.

[144] Strictly speaking, IMPLICATION itself doesn't yield this conclusion; it is a conditional which only states that DESCRIPTIVISM is a sufficient condition for the *falsity* of non-naturalism. It doesn't say anything about what happens when DESCRIPTIVISM is false. But it is natural to take Jackson's inadequacy thesis as an assertion of not only IMPLICATION, but also its converse (assuming bivalence). This amounts to the assumption that, if DESCRIPTIVISM is false, then non-naturalism is true. Given this, **F-prop** implies the falsity of DESCRIPTIVISM, which then implies that non-naturalism is true.

IMPLICATION needs to be rejected alongside ET in order to accommodate this asymmetry. In §5.1, we noted that characterizing non-naturalism as inconsistent with FUNDAMENTAL DESCRIPTIVISM avoids this kind of overgeneration. But once we reject ET, FUNDAMENTAL DESCRIPTIVISM won't be of any help, since if ET is false, normative properties aren't *fundamentally* descriptive properties—they aren't descriptive properties *at all.* What we need is a version of FUNDAMENTAL DESCRIPTIVISM which is adapted to a setting where necessarily co-extensive properties are distinct when they differ in degree of fundamentality. Non-naturalism should then be characterized as a view which is inconsistent with REVISED FUNDAMENTAL DESCRIPTIVISM:

REVISED FUNDAMENTAL DESCRIPTIVISM Every normative property necessarily co-extensive with a descriptive property that is **more fundamental** than it.

Such an understanding of non-naturalism allows for a view which accounts for why non-naturalism follows from NO GS^- without overgenerating. Under the supposition of NO GS^-, non-naturalism still follows: so long as our earlier argument for **F** is sound, **F-prop** follows. Since there is no *other* descriptive way of picking out a property that is necessarily co-extensive with rightness, rightness is more fundamental than other descriptive properties that are necessarily co-extensive with it. REVISED FUNDAMENTAL DESCRIPTIVISM is then false given NO GS^- and the falsity of ET.

With GS^- back in place, the line of reasoning doesn't go through: Jackson's assumptions tell us nothing about whether **F** or **F*** is true. Nothing follows, that is, about whether normative properties are more fundamental than their necessarily co-extensive but distinct descriptive counterparts. Thus nothing follows concerning REVISED FUNDAMENTAL DESCRIPTIVISM, and the argument avoids overgeneration. We can then conclude: rejecting ET for fundamentality-based reasons is a viable option within a fundamentality-centric framework, but it does not a provide a diagnosis of Jackson's assumptions which leaves IMPLICATION untouched.

4.5.3 Revisiting DESCRIPTIVISM$^-$ *and non-accidental expression*

Let us return briefly to some questions raised in §§2-3. In §2, we noted that non-naturalism is consistent with the possibility that normative properties are accidentally expressed by descriptive predicates like **D** or **D***:

D is the property I am actually thinking about now.

D* is actually being thought about by me now.

But we said nothing there about *why* descriptive language of this kind is consistent with non-naturalism.[145] From the fundamentality-centric perspective introduced in this section, there is a very natural explanation. Although descriptive indexical expressions can be used to capture truths about the world, reality is not fundamentally indexical—or, more precisely, to capture how things are fundamentally, one does not need to use indexical language. Hence it does not follow that FUNDAMENTAL DESCRIPTIVISM is true if one can use a descriptive indexical to express a normative property: there must be some more fundamental, non-indexical predicate that expresses this same property. And this very well could be a normative predicate. Indexical expressions are not the right kind of tool for attacking non-naturalism.[146]

A similar point sheds light on Jackson's claims about "property-denoting" expressions, and our discussion of "existential" expressions above. The descriptive predicate we constructed in §§3-4 uses an existential predicate to express rightness, but this (we claimed) doesn't show that non-naturalism is false. The reason is that, just as reality isn't fundamentally indexical, it also isn't fundamentally existential—to capture how things are fundamentally, one does not need to use existentially quantified expressions (or, at least, one does not need to use the particular existential expressions under discussion here). Jackson's "property-denoting" expressions are just an instance of existential expressions. Hence, even if they express rightness, they aren't the most fundamental means for doing so. As with indexical expressions, existential expressions aren't the right kind of tools for attacking non-naturalism.[147]

Finally, attending to distinctions in fundamentality gives us traction on a fall-back option for the proponent of Jackson's argument. In §2, we entertained a version of the argument which aims to establish not DESCRIPTIVISM, but rather DESCRIPTIVISM^{-}:

[145] We could legislate that the term 'descriptive' is reserved for non-indexical expressions, but this does nothing to explain the metaphysical question of why non-naturalism is consistent with this possibility.

[146] This way of making the point assumes that ET is still in place, but we could easily make the point with ET assumed to be false. If reality isn't fundamentally indexical, then for any descriptive property expressed by an indexical predicate, there must be a more fundamental property that is expressed by a non-indexical predicate. Since this might be a normative predicate, REVISED FUNDAMENTAL DESCRIPTIVISM does not follow.

[147] In a sense, we are agreeing with Jackson that existential and property-denoting expressions can't be used to show non-naturalism is false. Two points arise here, though: first, Jackson's motivation for this claim, which we glossed in §2, is lacking (this was the occasion for introducing the notion on non-accidental expression). Second, we are here giving an explanation for something Jackson simply stipulates. Moreover, the explanation isn't one Jackson can help himself to: it appeals to a characterization of non-naturalism as the denial of FUNDAMENTAL DESCRIPTIVISM—a characterization which, we have shown, has disastrous consequences for his original argument. So our agreement fails to offer much in the way of consolation to Jackson.

Descriptivism$^-$ Every normative property is identical to a MPT property.

If non-naturalism were characterized as the denial of Descriptivism$^-$, we would have a version of Jackson's argument that escapes *reductio*.

It won't do, however, just to find *some* restriction on the language that can be used to specify normative properties. Consider Descriptivism$_{52}$: the view that normative properties can be specified in a canonical descriptive language using less than fifty-two connectives. It is not, in an intuitive sense, an interesting metaphysical question whether Descriptivism$_{52}$ is true or false; very little of metaphysical significance hangs on the precise number of connectives needed to express rightness in a fundamental language. Whether Descriptivism$^-$ represents a more interesting thesis is a question of whether expressibility in MPT vocabulary marks a metaphysically significant property, unlike expressibility with fifty-two or fewer connectives.

It is quite natural to think that it does, on the grounds that MPT properties are *fundamental* properties, and so expressibility in MPT vocabulary is a metaphysically significant fact about the relationship of a property to the fundamental. The fact that a property can only be specified using fifty-three, rather than fifty-two, connectives has no analogous consequences. This is quite natural as an explanation of the difference between Descriptivism$^-$ and Descriptivism$_{52}$, but it doesn't support the proposed characterization of non-naturalism in the present setting. This is because the non-naturalist view we are trying to evaluate is precisely one on which normative vocabulary stands alongside MPT vocabulary as fully fundamental means of expressing a property. Descriptivism$^-$ cannot, absent some question-begging assumptions, provide a metaphysically interesting characterization of non-naturalism in this setting. If we antecedently assume that normative vocabulary is not among the fully fundamental vocabulary, then MPT-expressible properties represent a metaphysically interesting kind—for the MPT terms would exhaust the fully fundamental terms. But, of course, such an assumption is unwarranted here, since it is a straightforward denial of non-naturalism. If we avoid this assumption—that is, if we allow that fully fundamental vocabulary includes normative terms— Descriptivism$^-$ then risks being, like Descriptivism$_{52}$, a metaphysically uninteresting thesis about the normative. Once the MPT vocabulary is acknowledged to not correspond to the entire range of fully fundamental terms, it loses its metaphysical interest.[148]

[148] I do not wish to police use of the word 'non-naturalism', so one could in principle continue to use the word to talk about the denial of Descriptivism$^-$. Jackson's argument would then succeed in

4.6 Supervenience and reduction in general: a cautionary tale

I have sketched and motivated a response to Jackson's argument according to which non-naturalism is consistent with DESCRIPTIVISM, the claim that normative properties are identical to descriptive properties. The problem with Jackson's argument, I have suggested, is that it assumes that DESCRIPTIVISM is inconsistent with non-naturalism. Remedying this defect by rejecting IMPLICATION has ramifications for metaphysical theorizing about domains other than the normative. In closing, I will offer the briefest sketch of why this might be so.

Some arguments in metaphysics aim to establish that properties from different domains are identical. But if non-naturalism is, given some Jackson-like assumptions, consistent with the identity of normative and descriptive properties, the bare identity fails to have substantial metaphysical import for the metaphysics of the normative—it fails to even rule out the metaphysically extravagant non-naturalist view. This suggests that caution is needed when drawing metaphysical conclusions from claims about property-identity.[149]

For instance, one natural thought is that property-identities are metaphysically significant because a *reduction*, in a suitable metaphysical sense, immediately follows. Kim (2008) claims that this is not only true; it is obvious:

> There is no question about the reductive import of identity. If pain = N_1 [where N_1 is a specific neurophysiological state], there is no pain over and above N_1 [...] This is an open-and-shut affair if anything in philosophy ever is: Identities do reduce [...]
>
> [A]s far as reduction goes, nothing beats identities. That appropriate identities achieve reduction is intuitively obvious and beyond any philosophical second thoughts.[150]

showing that "non-naturalism", so construed, is false. But this would leave a metaphysically interesting thesis concerning the metaphysics of the normative untouched—namely, a thesis according to which the normative is fully fundamental. Whether we call this interesting thesis "non-naturalism" is not the main concern here.

[149] Here are some examples: Schmitt and Schroeder (2011, 145-6) observe that any supervenience thesis is committed to the existence of necessary connections between properties in the supervening set and subvening base. The best explanation for such necessary connections, they claim, holds that each supervening property is identical with a property in the subvening base. (See also Block and Stalnaker (1999, 24) and Kim (2008, 101).) And Kim (1989, 45) argues that once we derive the equivalence of properties from a supervenience claim, we cannot hold that properties in the supervening class are causally efficacious unless they are identical to subvening properties. The present point is simply that, even if these arguments are sound, more needs to be said if they are to yield metaphysically interesting conclusions.

[150] Kim (2008, 100; 113). Fodor agrees:

But it is overwhelmingly intuitive that non-naturalism is a metaphysically interesting view in part because it *denies* that the normative is reducible. Insofar as non-naturalists can hold that normative properties are identical with descriptive properties yet not reducible, we need to be careful to check whether the background assumptions in place really do license this conclusion. Should they be analogous to those in Jackson's argument for DESCRIPTIVISM, the present chapter casts significant doubt on the Kim's claim that a reduction follows.

These closing suggestions are quite obviously not decisive—after all, Kim might have additional background assumptions in mind that license his inference from identity to reduction. One obvious respect in which this is so is that in the case of non-naturalism, I have claimed that the identity of normative and descriptive properties follows only under the assumption of Jackson's ET—an assumption which we have acknowledged is not mandatory.[151] What these considerations show is not that inferences from identity to reduction are invalid; only that there are lots of moving parts in play when the relationships between supervenience, identity, and reduction are at issue. The possibility of rejecting Jackson's IMPLICATION premise as outlined above shows that there are some classes of assumptions under which metaphysically interesting reductive theses do *not* follow from bare identity-claims. Whether similar issues arise in other areas of metaphysics is a question that deserves a second look.[152]

> Functionalists are required to deny that pain is *identical to* the disjunction of its realizers [...] And the reason they have to say *that* is that *otherwise multiple realization wouldn't be an argument against reduction.* (Fodor (1997, 155), Fodor's emphasis.)

[151] Alternatively, Kim and others might have an assumption about fundamentality in the background; perhaps they are presupposing that *if* mental states are identical to neurophysiological states, then these states are *most fundamentally* neurophysiological states.

[152] Additional thanks go to Campbell Brown, Jamie Dreier, Ishani Maitra, Sarah Moss, David Plunkett, Bart Struemer, and Brian Weatherson for helpful discussion of the various issues covered in this chapter.

Bibliography

Robert Merrihew Adams. Theories of Actuality. *Nous*, 8(3):211–231, 1974.

G.E.M. Anscombe. *Intention.* Basil Blackwell, 1957.

D.M. Armstrong. *A Theory of Universals, vol. 1.* Cambridge Univsersity Press, 1978.

Jonathan Bennett. *A Philosophical Guide to Conditionals.* Oxford University Press, 2003.

Marie François Xavier Bichat. *Anatomie générale appliquée à la physiologie et à la médicine.* Brossom, Gabon et Cie., 1801.

Simon Blackburn. *Spreading the Word.* Oxford University Press, 1984.

Simon Blackburn. How to be an Ethical Antirealist. *Midwest Studies in Philosophy*, XII: 361–375, 1988a.

Simon Blackburn. Supervenience Revisted. In Geoffrey Sayre-McCord, editor, *Essays on Moral Realism*, pages 59–75. Cornell University Press, 1988b.

Simon Blackburn. Attitudes and Contents. *Ethics*, 98(3):501–517, 1988c.

Ned Block and Robert Stalnaker. Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review*, 108:1–46, 1999.

Richard N. Boyd. How to be a Moral Realist. In Geoffrey Sayre-McCord, editor, *Essays on Moral Realism.* Cornell University Press, 1989.

David O. Brink. Moral Realism and the Skeptical Arguments from Disagreement and Queerness. *Australasian Journal of Philosophy*, 62(2):111–125, 1984.

Campbell Brown. A New and Improved Supervenience Argument for Ethical Descriptivism. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 6*, pages 205–218. Oxford University Press, 2011.

David Chalmers. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press, 1996.

Noam Chomsky. *Knowledge of Language: Its Nature, Origins and Use.* Praeger Publishers, 1986.

Paul M. Churchland. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy*, 78(2):67–90, 1981.

David Copp. Milk, Honey, and the Good Life on Moral Twin Earth. *Synthese*, 124:113–137, 2000.

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

Michael Dummett. Realism. *Synthese*, 52:55–112, 1982.

Billy Dunaway. Minimalist Semantics in Meta-ethical Expressivism. *Philosophical Studies*, 151(3):351–371, 2010.

David Enoch. The Epistemological Challenge to Metanormative Realism. *Philosophical Studies*, 148(3):413–438, 2010.

A.C. Ewing. *The Definition of Good*. Routledge and Kegan Paul, 1947.

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

Kit Fine. The Question of Ontology. In David Chalmers, David Manley, and Ryan Wasserman, editors, *Metametaphysics*, pages 157–177. Oxford University Press, 2009.

Stephen Finlay. Four Faces of Moral Realism. *Philosophy Compass*, 2(6):820–849, 2007.

William Fitzpatrick. Robust Ethical Realism, Non-naturalism, and Normativity. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 3*, pages 159–206. Oxford University Press, 2008.

Jerry Fodor. Special Sciences: Still Autonomous after All These Years. *Philosophical Perspectives*, 11:149–163, 1997.

Peter Geach. Ascriptivism. *The Philosophical Review*, 69(2):221–225, 1960.

Allan Gibbard. *Wise Choices, Apt Feelings*. Harvard University Press, 1990.

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

Allan Gibbard. *Meaning and Normativity*. Oxford University Press, 2013.

R. M. Hare. Imperative Sentences. *Mind*, 58(229):21–39, 1949.

R. M. Hare. *The Language of Morals*. Oxford University Press, 1952. All citations to the 1964 edition.

R. M. Hare. *Moral Thinking*. Oxford University Press, 1981.

John Hawthorne. Deeply Contingent A Priori Knowledge. *Philosophy and Phenomenological Research*, 265:247–269, 2002.

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

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

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

Terrence Horgan and Mark Timmons. Copping Out on Moral Twin Earth. *Synthese*, 124: 139–152, 2000.

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

Frank Jackson. Responses. *Philosophy and Phenomenological Research*, 62(3):653–664, 2001.

C.S. Jenkins. Realism and Independence. *American Philosophical Quarterly*, 42(3):199–211, 2005.

Christopher Kennedy. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. PhD thesis, University of California, Santa Cruz, 1997.

Jaegwon Kim. Supervenience and Nomological Incommensurables. *American Philosophical Quarterly*, 15(2):149–156, 1978.

Jaegwon Kim. The Myth of Nonreductive Materialism. *Proceedings and Addresses of the American Philosophical Association*, 63(3):31–47, 1989.

Jaegwon Kim. Reduction and Reductive Explanation. In Jakob Hohwy and Jesper Kallestrup, editors, *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press, 2008.

Saul Kripke. *Wittgenstein on Rules and Private Language*. Harvard University Press, 1982.

David Lewis. How to Define Theoretical Terms. *Journal of Philosophy*, 67(13):427–446, 1970.

David Lewis. New Work for a Theory of Universals. *Australasian Journal of Philosophy*, 61(4):343–377, 1983.

David Lewis. Putnam's Paradox. *Australasian Journal of Philosophy*, 62(3):221–236, 1984.

David Lewis. *On the Plurality of Worlds*. Basil Blackwell, 1986.

J.L. Mackie. *Ethics: Inventing Right and Wrong*. Penguin, 1977.

David Manley. Introduction: A Guided Tour of Metametaphysics. In David Chalmers, David Manley, and Ryan Wasserman, editors, *Metametaphysics: New Essays in the Foundations of Ontology*, pages 1–37. Oxford University Press, 2009.

Alexander Miller. Realism. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/sum2010/entries/realism/, 2010.

G.E. Moore. *Principia Ethica*. Cambridge Univsersity Press, 2nd edition, 1903. All page numbers to the revised second edition, published in 1993.

G.E. Moore. A Reply to My Critics. In P. A. Schlipp, editor, *The Philosophy of G. E. Moore*, pages 535–677. Open Court, 1942. All citations to the 3rd edition, published 1968.

Thomas Nagel. *The View from Nowhere*. Oxford University Press, 1989.

Alvin Plantinga. *The Nature of Necessity*. Clarendon Press, 1978.

Alvin Plantinga. Two Concepts of Modality: Modal Realism and Modal Reductionism. *Philosophical Perspectives*, 1:189–231, 1987.

Hilary Putnam. The Nature of Mental States. In Ned Block, editor, *Readings in the Philosophy of Psychology*. Harvard University Press, 1980.

Hilary Putnam. *Reason, Truth and History*. Cambridge Univsersity Press, 1981.

W. V. O. Quine. *Word and Object*. M.I.T. Press, 1960.

Peter Railton. Moral Realism. *The Philosophical Review*, 95(2):163–207, 1986.

Peter Railton. Naturalism and Prescriptivity. *Social Philosophy and Policy*, 7:151–174, 1989.

Steven Reynolds. Realism and the Meaning of 'Real'. *Nous*, 40:468–494, 2006.

Gideon Rosen. Modal Fictionalism. *Mind*, 99(395):327–354, 1990.

Gideon Rosen. Objectivity and Modern Idealism: What is the Question? In Michaelis Michael and John O'Leary-Hawthorne, editors, *Philosophy in Mind: The Place of Philosophy in the Study of Mind*, pages 277–319. Kluwer Academic, 1994.

Bertrand Russell. *The Problems of Philosophy*. Oxford University Press, 1912.

Geoffrey Sayre-McCord. 'Good' on Twin Earth. *Philosophical Issues*, 8:267–292, 1997.

T.M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.

Jonathan Schaffer. Two Conceptions of Sparse Properties. *Pacific Philosophical Quarterly*, 85:92–102, 2004.

Jonathan Schaffer. On What Grounds What. In David Chalmers, David Manley, and Ryan Wasserman, editors, *Metametaphysics*. Oxford University Press, 2009.

Johannes Schmitt and Mark Schroeder. Supervenience Arguments under Relaxed Assumptions. *Philosophical Studies*, 155:133–160, 2011.

Mark Schroeder. Realism and Reduction: the Quest for Robustness. *Philosophers' Imprint*, 5(1):1–18, 2005.

Mark Schroeder. *Slaves of the Passions*. Oxford University Press, 2007.

Mark Schroeder. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford University Press, 2008a.

Mark Schroeder. How Expressivists Can and Should Solve Their Problem with Negation. *Nous*, 42(4):573–599, 2008b.

Russ Shafer-Landau. *Moral Realism: A Defense*. Oxford University Press, 2003.

Theodore Sider. *Writing the Book of the World*. Oxford University Press, 2012.

Michael Smith. *The Moral Problem*. Wiley-Blackwell, 1994.

Robert Stalnaker. Possible worlds. *Nous*, 10:65–75, 1976.

C.L. Stevenson. The Emotive Meaning of Ethical Terms. *Mind*, 46:14–31, 1937.

Sharon Street. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, 127(1):109–166, 2006.

Bart Streumer. Are there irreducibly normative properties? *Australasian Journal of Philosophy*, 86(4):537–561, 2008.

Nicholas Sturgeon. Doubts about the Supervenience of the Evaluative. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 4*, pages 53–90. Oxford University Press, 2009.

Jussi Suikkanen. Non-naturalism: the Jackson Challenge. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 5*, pages 87–110. Oxford University Press, 2010.

Bas van Fraassen. *The Scientific Image*. Oxford University Press, 1980.

Mark van Roojen. Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 1*. Oxford University Press, 2006.

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

Alfred North Whitehead and Bertrand Russell. *Principia Mathematica*. Cambridge University Press, 1910.

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

Seth Yalcin. Nonfactualism about Epistemic Modality. In Andy Egan and Brian Weatherson, editors, *Epistemic Modality*, pages 295–332. Oxford University Press, 2011.